


The SAGE Handbook of  
Research Methods in  
Political Science and  
International Relations



2 Volume set

Edited by  
Luigi Curini and  
Robert Franzese



The SAGE Handbook of  
**Research Methods in  
Political Science and  
International Relations**



**SAGE** was founded in 1965 by Sara Miller McCune to support the dissemination of usable knowledge by publishing innovative and high-quality research and teaching content. Today, we publish over 900 journals, including those of more than 400 learned societies, more than 800 new books per year, and a growing range of library products including archives, data, case studies, reports, and video. SAGE remains majority-owned by our founder, and after Sara's lifetime will become owned by a charitable trust that secures our continued independence.

Los Angeles | London | New Delhi | Singapore | Washington DC | Melbourne

# The SAGE Handbook of Research Methods in Political Science and International Relations



Volume 1

Edited by  
Luigi Curini and  
Robert Franzese

 SAGE reference

Los Angeles | London | New Delhi | Singapore | Washington DC | Melbourne





Los Angeles | London | New Delhi  
Singapore | Washington DC | Melbourne

SAGE Publications Ltd  
1 Oliver's Yard  
55 City Road  
London EC1Y 1SP

SAGE Publications Inc.  
2455 Teller Road  
Thousand Oaks, California 91320

SAGE Publications India Pvt Ltd  
B 1/1 1 Mohan Cooperative Industrial Area  
Mathura Road  
New Delhi 110 044

SAGE Publications Asia-Pacific Pte Ltd  
3 Church Street  
#10-04 Samsung Hub  
Singapore 049483

---

Editor: Natalie Aguilera  
Editorial Assistant: Umeeka Raichura  
Production Editor: Jessica Masih  
Copyeditor: Sunrise Setting  
Proofreader: Sunrise Setting  
Indexer: Marketing Manager: Susheel  
Gokarakonda  
Cover Design: Naomi Robinson  
Typeset by Cenveo Publisher Services  
Printed in the UK

At SAGE we take sustainability seriously. Most of our products are printed in the UK using responsibly sourced papers and boards. When we print overseas we ensure sustainable papers are used as measured by the PREPS grading system. We undertake an annual audit to monitor our sustainability.

Apart from any fair dealing for the purposes of research or private study, or criticism or review, as permitted under the Copyright, Designs and Patents Act, 1988, this publication may be reproduced, stored or transmitted in any form, or by any means, only with the prior permission in writing of the publishers, or in the case of reprographic reproduction, in accordance with the terms of licences issued by the Copyright Licensing Agency. Enquiries concerning reproduction outside those terms should be sent to the publishers.

**Library of Congress Control Number:**  
2020931911

**British Library Cataloguing in Publication data**

A catalogue record for this book is available from the British Library

978-1-5264-5993-0

Editorial arrangement, Foreword & Introduction © Luigi Curini and Robert Franzese, 2020  
Preface © Gary King, 2020

Chapter 1 © William Roberts Clark, 2020  
Chapter 2 © Adam McCauley and Andrea Ruggeri, 2020  
Chapter 3 © Branislav L. Slantchev, 2020  
Chapter 4 © Ravi Bhavnani, Karsten Donnay and Mirko Reul, 2020  
Chapter 5 © Mitchell Goist and Burt L. Monroe, 2020  
Chapter 6 © Ezequiel González-Ocantos, 2020  
Chapter 7 © Thomas Bräuninger and Tilko Swalve, 2020  
Chapter 8 © John Aldrich and Jim Granato, 2020  
Chapter 9 © Rose McDermott, 2020  
Chapter 10 © Maxfield J. Peterson and B. Guy Peters, 2020  
Chapter 11 © Adam Meirowitz and Kris Ramsay, 2020  
Chapter 12 © James Adams, Samuel Merrill III and Roi Zur, 2020  
Chapter 13 © Charles Cameron and Nathan Gibson, 2020  
Chapter 14 © Lanny W. Martin and Georg Vanberg, 2020  
Chapter 15 © James D. Morrow and Jessica S. Sun, 2020  
Chapter 16 © Deborah Beim, 2020  
Chapter 17 © Scott de Marchi and Brandon M. Stewart, 2020  
Chapter 18 © Scott J. LaCombe and Frederick J. Boehmke, 2020  
Chapter 19 © Gerardo L. Munck, Jørgen Møller and Svend-Erik Skaaning, 2020  
Chapter 20 © Christopher J. Fariss, Michael R. Kenwick and Kevin Reuning, 2020  
Chapter 21 © Lucas Leemann and Fabio Wasserfallen, 2020  
Chapter 22 © Dominic Nyhuis, 2020  
Chapter 23 © Pablo Barberá and Zachary C. Steinert-Threlkeld, 2020  
Chapter 24 © David Darmofal and Christopher Eddy, 2020  
Chapter 25 © Richard Traumnüller, 2020  
Chapter 26 © Ken Benoit, 2020  
Chapter 27 © Benjamin C.K. Egerod and Robert Klemmensen, 2020  
Chapter 28 © Sarah B. Bouchat, 2020  
Chapter 29 © Luigi Curini and Robert A. Fahey, 2020  
Chapter 30 © Ernesto Calvo, Joan C. Timoneda and Tiago Ventura, 2020  
Chapter 31 © Robert Franzese, 2020  
Chapter 32 © Suzanna Linn and Clayton Webb, 2020

Chapter 33 © Vera Troeger, 2020  
Chapter 34 © Mark Pickup, 2020  
Chapter 35 © Kentaro Fukumoto, 2020  
Chapter 36 © Marco Steenbergen, 2020  
Chapter 37 © Tobias Böhmelt and Gabriele Spilker, 2020  
Chapter 38 © Eric Neumayer and Thomas Plümper, 2020  
Chapter 39 © Scott J. Cook, Jude C. Hays and Robert Franzese, 2020  
Chapter 40 © Christopher L. Carter and Thad Dunning, 2020  
Chapter 41 © Jake Bowers and Thomas Leavitt, 2020  
Chapter 42 © Richard A. Nielsen, 2020  
Chapter 43 © Luke Keele, 2020  
Chapter 44 © Matias D. Cattaneo, Rocio Tituniuk and Gonzalo Vazquez-Bare, 2020  
Chapter 45 © Jennifer N. Victor and Elsa T. Khwaja, 2020  
Chapter 46 © John P. Schoeneman and Bruce A. Desmarais, 2020  
Chapter 47 © Jong Hee Park and Soohahn Shin, 2020  
Chapter 48 © Shawn Treier, 2020  
Chapter 49 © Florian M. Hollenbach and Jacob M. Montgomery, 2020  
Chapter 50 © Jeff Gill and Simon Heuberger, 2020  
Chapter 51 © Rebecca Morton and Mateo Vásquez-Cortés, 2020  
Chapter 52 © Betsy Sinclair, 2020  
Chapter 53 © Anna M. Wilke and Macartan Humphreys, 2020  
Chapter 54 © Gustavo Díaz, Christopher Grady and James H. Kuklinski, 2020  
Chapter 55 © Kakkia Chatsiou and Slava Jankin Mikhaylov, 2020  
Chapter 56 © Santiago Olivella and Kelsey Shoub, 2020  
Chapter 57 © Adrian Duşa, 2020  
Chapter 58 © Imke Harbers and Matthew C. Ingram, 2020  
Chapter 59 © Chiara Ruffa, 2020  
Chapter 60 © Klaus Brummer, 2020  
Chapter 61 © Claire Greenstein and Layna Mosley, 2020  
Chapter 62 © Virginie Van Ingelgom, 2020  
Chapter 63 © Xymena Kurowska and Berit Bliesemann de Guevara, 2020  
Editors' Afterword © Luigi Curini and Rob Franzese, 2020

# Contents

<i>List of Figures</i>	xi
<i>List of Tables</i>	xvii
<i>Notes on the Editors and Contributors</i>	xx
<i>An Introduction</i>	xxxix
<i>Foreword</i>	xliii

## VOLUME 1

Preface: So You're A Grad Student Now? Maybe You Should Do This <i>Gary King</i>	1
<b>PART I FORMULATING GOOD RESEARCH QUESTIONS AND DESIGNING GOOD RESEARCH PROJECTS</b>	<b>5</b>
1. Asking Interesting Questions <i>William Roberts Clark</i>	7
2. From Questions and Puzzles to Research Project <i>Adam McCauley and Andrea Ruggeri</i>	26
3. The Simple, the Trivial and the Insightful: Field Dispatches from a Formal Theorist <i>Branislav L. Slantchev</i>	44
4. Evidence-Driven Computational Modeling <i>Ravi Bhavnani, Karsten Donnay and Mirko Reul</i>	60
5. Taking Data Seriously in the Design of Data Science Projects <i>Mitchell Goist and Burt L. Monroe</i>	79
6. Designing Qualitative Research Projects: Notes on Theory Building, Case Selection and Field Research <i>Ezequiel González-Ocantos</i>	104
7. Theory Building for Causal Inference: EITM Research Projects <i>Thomas Bräuninger and Tilko Swalve</i>	121
8. EITM: Applications in Political Science and International Relations <i>John Aldrich and Jim Granato</i>	142

<b>PART II</b>	<b>METHODS OF THEORETICAL ARGUMENTATION</b>	<b>159</b>
9.	Political Psychology, Social Psychology and Behavioral Economics <i>Rose McDermott</i>	161
10.	Institutional Theory and Method <i>Maxfield J. Peterson and B. Guy Peters</i>	173
11.	Applied Game Theory: An Overview and First Thoughts on the Use of Game Theoretic Tools <i>Adam Meirowitz and Kristopher W. Ramsay</i>	192
12.	The Spatial Voting Model <i>James Adams, Samuel Merrill III and Roi Zur</i>	205
13.	New Directions in Veto Bargaining: Message Legislation, Virtue Signaling, and Electoral Accountability <i>Charles Cameron and Nathan Gibson</i>	224
14.	Models of Coalition Politics: Recent Developments and New Directions <i>Lanny W. Martin and Georg Vanberg</i>	244
15.	Models of Interstate Conflict <i>James D. Morrow and Jessica S. Sun</i>	261
16.	Models of the Judiciary <i>Deborah Beim</i>	277
17.	Wrestling with Complexity in Computational Social Science: Theory, Estimation and Representation <i>Scott de Marchi and Brandon M. Stewart</i>	289
18.	Learning and Diffusion Models <i>Scott J. LaCombe and Frederick J. Boehmke</i>	311
<b>PART III</b>	<b>CONCEPTUALIZATION AND MEASUREMENT</b>	<b>329</b>
19.	Conceptualization and Measurement: Basic Distinctions and Guidelines <i>Gerardo L. Munck, Jørgen Møller and Svend-Erik Skaaning</i>	331
20.	Measurement Models <i>Christopher J. Fariss, Michael R. Kenwick and Kevin Reuning</i>	353
21.	Measuring Attitudes – Multilevel Modeling with Post-Stratification (MrP) <i>Lucas Leemann and Fabio Wasserfallen</i>	371

<b>PART IV</b>	<b>LARGE-SCALE DATA COLLECTION AND REPRESENTATION METHODS</b>	<b>385</b>
22.	Web Data Collection: Potentials and Challenges <i>Dominic Nyhuis</i>	387
23.	How to Use Social Media Data for Political Science Research <i>Pablo Barberá and Zachary C. Steinert-Threlkeld</i>	404
24.	Spatial Data <i>David Darmofal and Christopher Eddy</i>	424
25.	Visualizing Data in Political Science <i>Richard Traummüller</i>	436
26.	Text as Data: An Overview <i>Ken Benoit</i>	461
27.	Scaling Political Positions from Text: Assumptions, Methods and Pitfalls <i>Benjamin C.K. Egerod and Robert Klemmensen</i>	498
28.	Classification and Clustering <i>Sarah B. Bouchat</i>	522
29.	Sentiment Analysis and Social Media <i>Luigi Curini and Robert A. Fahey</i>	534
30.	Big Relational Data: Network-Analytic Measurement <i>Ernesto Calvo, Joan C. Timoneda and Tiago Ventura</i>	552

## VOLUME 2

<b>PART V</b>	<b>QUANTITATIVE-EMPIRICAL METHODS</b>	<b>575</b>
31.	Econometric Modeling: From Measurement, Prediction, and Causal Inference to Causal-Response Estimation <i>Robert Franzese</i>	577
32.	A Principled Approach to Time Series Analysis <i>Suzanna Linn and Clayton Webb</i>	599
33.	Time-Series-Cross-Section Analysis <i>Vera Troeger</i>	616
34.	Dynamic Systems of Equations <i>Mark Pickup</i>	632

35.	Duration Analysis <i>Kentaro Fukumoto</i>	659
36.	Multilevel Analysis <i>Marco Steenbergen</i>	679
37.	Selection Bias in Political Science and International Relations Applications <i>Tobias Böhmelt and Gabriele Spilker</i>	701
38.	Dyadic Data Analysis <i>Eric Neumayer and Thomas Plümper</i>	717
39.	Model Specification and Spatial Interdependence <i>Scott J. Cook, Jude C. Hays, and Robert Franzese</i>	730
40.	Instrumental Variables: From Structural Equation Models to Design-Based Causal Inference <i>Christopher L. Carter and Thad Dunning</i>	748
41.	Causality and Design-Based Inference <i>Jake Bowers and Thomas Leavitt</i>	769
42.	Statistical Matching with Time-Series Cross-Sectional Data: Magic, Malfeasance, or Something in between? <i>Richard A. Nielsen</i>	805
43.	Differences-in-Differences: Neither Natural nor an Experiment <i>Luke Keele</i>	822
44.	The Regression Discontinuity Design <i>Matias D. Cattaneo, Rocío Titiunik, and Gonzalo Vazquez-Bare</i>	835
45.	Network Analysis: Theory and Testing <i>Jennifer N. Victor and Elsa T. Khwaja</i>	858
46.	Network Modeling: Estimation, Inference, Comparison, and Selection <i>John P. Schoeneman and Bruce A. Desmarais</i>	876
47.	Bayesian Methods in Political Science <i>Jong Hee Park and Sooahn Shin</i>	895
48.	Bayesian Ideal Point Estimation <i>Shawn Treier</i>	910
49.	Bayesian Model Selection, Model Comparison, and Model Averaging <i>Florian M. Hollenbach and Jacob M. Montgomery</i>	937
50.	Bayesian Modeling and Inference: A Postmodern Perspective <i>Jeff Gill and Simon Heuberger</i>	961

51.	Laboratory Experimental Methods <i>Rebecca Morton and Mateo Vásquez-Cortés</i>	985
52.	Field Experiments on the Frontier: Designing Better <i>Betsy Sinclair</i>	999
53.	Field Experiments, Theory, and External Validity <i>Anna M. Wilke and Macartan Humphreys</i>	1007
54.	Survey Experiments and the Quest for Valid Interpretation <i>Gustavo Diaz, Christopher Grady, and James H. Kuklinski</i>	1036
55.	Deep Learning for Political Science <i>Kakia Chatsiou and Slava Jankin Mikhaylov</i>	1053
56.	Machine Learning in Political Science: Supervised Learning Models <i>Santiago Olivella and Kelsey Shoub</i>	1079
<b>PART VI QUALITATIVE AND ‘MIXED’ METHODS</b>		1095
57.	Set Theoretic Methods <i>Adrian Duşa</i>	1097
58.	Mixed-Methods Designs <i>Imke Harbers and Matthew C. Ingram</i>	1117
59.	Case Study Methods: Case Selection and Case Analysis <i>Chiara Ruffa</i>	1133
60.	Comparative Analyses of Foreign Policy <i>Klaus Brummer</i>	1148
61.	When Talk Isn’t Cheap: Opportunities and Challenges in Interview Research <i>Claire Greenstein and Layna Mosley</i>	1167
62.	Focus Groups: From Qualitative Data Generation to Analysis <i>Virginie Van Ingelgom</i>	1190
63.	Interpretive Approaches in Political Science and International Relations <i>Xymena Kurowska and Berit Bliesemann de Guevara</i>	1211
	Editors’ Afterword <i>Luigi Curini and Robert Franzese</i>	1231
	<i>Index</i>	1232

*This page intentionally left blank*

# List of Figures

1.1	Relationship between the height of fathers and sons	12
1.2	Partisanship and beliefs about ‘room to maneuver’: the conditional effects of knowledge and age	19
2.1	From specific to abstract (and vice versa)	29
2.2	Where good research questions come from	32
2.3	On research question types	34
2.4	Stylized democratic peace research agenda	36
2.5	Stylized civil war research agenda	36
4.1	Model development process	65
4.2	Flow diagram for MERIAM EDM	66
4.3	Data construction overview	68
4.4	EDM modeling cycle	70
4.5	Mock-up of SAMT (main window)	73
4.6	Mock-up of SAMT (intervention configuration)	74
5.1	Searches for ‘basketball’ have the expected seasonality and a plausible pattern of growth	88
5.2	The prominent dynamic in searches for ‘islam’ is the same as that in searches for ‘essays’	89
5.3	Comparing to a reference term captures more relevant shifts in attention	90
5.4	Profiling of open-ended responses in the Mood of the Nation Poll (Wave 4), run by Penn State’s McCourtney Institute for Democracy	92
5.5	Downsampling aggregates pixels (observations) into new pixels, reducing the resolution of the image	94
5.6	Effect of dimension reduction through singular value decomposition (SVD)	96
5.7	A kernel density estimation of the Lincoln images pixel intensities convolves a Gaussian curve (normal distribution) over the observed distribution of values	97
5.8	Effect of convolutional image kernels	97
7.1	Models, theoretical implications, and features of the real world	125
7.2	Entry game	129
8.1	Simulated and actual volatility	150
8.2	Deregulation effects	150
11.1	The bargaining problem with unilateral outside options	195
12.1	Illustrative placements of a voter $v$ and parties $A, B$ , on left–right ideology	206
12.2	Illustrative placement of a voter $v$ and parties $L$ and $R$ in a two-dimensional space	207
12.3	How two-party, positional competition motivates party convergence	209
12.4	The dynamics of three-party positional competition	210
12.5	Centripetal incentives in a four-party election: the peripheral parties converge toward their ideological ‘neighbors’	211
12.6	Example of four-party Nash equilibrium configuration in competition over a bimodal voter distribution	211



12.7	Distribution of American survey respondents' Left–Right self-placements and their mean party placements, 2016 National Election Study	212
12.8	Distributions of citizens' Left–Right self-placements and their mean party placements in the UK, Germany, Canada and Finland	213
12.9	How valence affects voter choice in a model with one positional dimension	214
12.10	Party strategies in elections with one positional and one valence dimension	215
12.11	Examples of probability distributions over the median voter's position	217
12.12	How party positioning affects parties' election prospects when there is uncertainty over the median voter position	218
13.1	Passage margin of cloture votes, 2009–2018	234
13.2	Futile cloture votes, 1975–2018	234
14.1	Laver–Shepsle ministerial autonomy model	251
19.1	The parts of a concept: term, sense and reference	333
19.2	The structure of a concept: levels and relationships	334
19.3	Conceptual systems I: typologies	335
19.4	Conceptual systems II: two hierarchical structures	335
19.5	The concept–fact interface	337
19.6	The production of data on indices: two basic situations	345
19.7	Combining data on properties: reflective and formative measurement models	346
20.1	Latent variables and item parameters	357
20.2	Example of additive scale function	358
20.3	Identification issues in latent variables	360
21.1	Cantonal estimates and true approval rate	375
21.2	Cantonal estimates and uncertainty	376
22.1	Response from the Wikipedia API (Mannheim page views between November 3 and November 9, 2018)	390
22.2	Response from the Wikipedia API (Mannheim coordinates)	391
22.3	Interpreted version of the Mannheim Wikipedia page (excerpt), November 13, 2018	394
22.4	Source code of the Mannheim Wikipedia page (excerpt), November 13, 2018	395
22.5	Response from the <i>Google Translate</i> server to the query	400
25.1	Four scatter plots of four data sets that show wildly different patterns – although summary statistics are identical	437
25.2	Left panel: average number of figures in all articles published in the <i>AJPS</i> , 2003–18. Right panel: relative frequency of graphical formats	439
25.3	The relationship between graph use and citation counts of <i>AJPS</i> articles, 2003–17. Left panel: simple scatter plot with jitter along the x-axis. Right panel: adjusted variable plot relating residuals of figure count to residuals of citations eliminating the effect of publication date	441
25.4	Table lens plots of Swiss census data 2010	443
25.5	Parallel coordinate plot of the policy preferences of over a thousand political candidates	447
25.6	Parallel coordinate plot of the policy preferences of over a thousand political candidates. Axes have been re-sorted	447
25.7	Parallel coordinate plot of the policy preferences of over a thousand political candidates. Lines are colored according to party affiliation	447
25.8	Parallel coordinate plot of the policy preferences of political candidates of the FDP and the Left	448

25.9	Parallel coordinate plot of the policy preferences of political candidates of the CDU/CSU and SPD	448
25.10	Parallel coordinate plot of the policy preferences of political candidates of the AfD and the Greens	448
25.11	Small multiples of support for school vouchers by geography and income (Gelman, 2009)	449
25.12	Small multiples of support for school vouchers by geography, income and ethno-religious group (Gelman, 2009)	450
25.13	Small multiples of support for school vouchers by geography, income and ethno-religious group. Rows have been re-sorted by support	451
25.14	Plot ensemble for exploratory model analysis of the determinants of civil war onset	452
25.15	Line-up visual inference with 20 histograms	455
25.16	Line-up for the relation between the individual education effect on political participation (y-axis) and state-level education (x-axis)	456
25.17	Line-up for the relation between survival vs. self-expression values (x-axis) and traditional vs. secular-rational values (x-axis) clustered by cultural zone	457
26.1	From text to data to data analysis	464
26.2	From text to tokens to matrix	482
26.3	Word cloud of influential hashtags from a sample of Tweets about Brexit	491
27.1	On a scale from Medvedev to Putin	506
27.2	Perception and governor speeches	507
27.3	Validating Wordfish estimates against a human benchmark	510
27.4	Which words define the policy space?	511
27.5	How the performance of scaling algorithms vary with the comparability of texts	514
27.6	Document length and performance of scaling models	515
27.7	Error in scales and bias in econometric models	519
30.1	API interface and pre-processing of large Twitter network	555
30.2	Sample code to create a dataframe with selected variables	558
30.3	Twitter ‘Trump’ network (1)	559
30.4	Twitter ‘Trump’ network (2)	564
30.5	A complex network with three local topologies: (a) a ring subnetwork, (b) a fully connect subnetwork and (c) a follow the leader subnetwork	565
30.6	Activation of eight different hashtags during the Kavanaugh Confirmation Hearings, September 19, 2018	569
30.7	Primary connected network of #Kavanaugh	570
30.8	Sampling distribution of the bandwidth selection procedure (250 samples with 100 observations each)	571
30.9	PWR results in the network format	572
31.1	The logically necessarily sigmoidal relation $p(y) = f(x)$	585
31.2	‘Multiple Hands on the Wheel’ model of complex context-conditionality in monetary policymaking	587
31.3	Substantive dynamic-effect estimates of real interest-rate net of growth impacts on public debt	588
31.4	Maps depicting the initial (left panel) and LRSS (right panel) spatial ALM-spending responses to +1 shock in Germany	590

31.5	BSVAR Estimated responses from a system of Israel↔Palestinian, US→Israel, US→Palestinian actions	592
32.1	IRF and CIRF for the effect of consumer sentiment on presidential approval	608
32.2	IRFs for three variable VAR model of economic performance consumer sentiment and presidential approval	610
34.1	Impulse response and dynamic multiplier functions	639
34.2	Structural impulse response function	640
34.3	Two-dimensional space	648
34.4	Smooth states for vote intention	656
36.1	Three canonical multilevel-data structures	680
36.2	Three types of random-coefficient models	684
36.3	Conceptualizing ordinal outcomes	695
39.1	Manifestations of spatial association	733
41.1	Distribution of Difference-in-Means estimator as $h \rightarrow \infty$	777
41.2	Distribution of Difference-in-Means estimates under (a) unblocked and (b) blocked assignment	779
41.3	Distribution of Difference-in-Means test statistic under test of strong null of no effect when $\mathbf{z}_g$ is the realized assignment	784
41.4	Distribution of observed test statistic	785
41.5	Distribution of Fisherian $p$ -values for test of strong null of no effect under all realizations of data as the size of the experimental pool grows from one copy of the study, $h = 1$ ( $n = 6$ ); to two copies of the study, $h = 2$ ( $n = 12$ ); to three copies of the study, $h = 3$ ( $n = 18$ )	788
41.6	Distribution of Neymanian $p$ -values under all realizations of data	790
41.7	Distribution of type I error probabilities for different $\alpha$ levels	791
41.8	Distributions of Difference-in-Means test statistic under strong null of no effect when $\mathbf{z}_g$ is the realized assignment under (a) a model where an unobserved covariate has no effect on odds of treatment, $\Gamma = 1$ , and (b) a model where an observed covariate doubles the odds of treatment, $\Gamma = 2$	797
42.1	The number of articles per year using statistical matching in the 12 leading IR journals	806
43.1	Schematic representation of response in treated and control groups, before and after treatment, with and without transformation to log scale	825
44.1	Estimated density of running variable	850
44.2	RD effects on predetermined covariates	851
44.3	RD effects on predetermined covariates	852
44.4	Effect of victory at $t$ on vote margin at $t + 1$ . Incumbent party, Brazilian mayoral elections, 1996–2012	853
44.5	Window selector based on covariates. Incumbent party, Brazilian mayoral elections, 1996–2012	853
46.1	Density plots for degree distribution and betweenness centrality for all three networks	884
47.1	The number of journal articles containing Bayesian as a keyword excluding game theory papers	900
50.1	How Markov Chain Monte Carlo works	968
50.2	Categories illustration	971
50.3	Posterior predictive probabilities by ordered outcomes	971
50.4	Posterior predictive probabilities by outcome categories	972

50.5	Log-likelihood functions for configurations of component likelihoods	978
53.1	A directed acyclic graph (DAG) for a non-parametric bargaining model	1009
53.2	The hazards of trying to stitch experimental results together to form a theory	1015
53.3	Three selection graphs	1021
55.1	Machine learning and related tasks	1055
55.2	Building a hierarchy of complex concepts out of simpler concepts	1058
55.3	Visualization of features	1059
55.4	Examples of Digital Globe (left) and Planet (right, Michoacán) imagery	1059
55.5	Convolution of a black and white image of a heart – the essential elements	1060
55.6	Convolution of a black and white image of a heart – step one	1060
55.7	Convolution of a black and white image of a heart – step two	1061
55.8	Convolution of a black and white image of a heart – step three	1061
55.9	Max pooling operation – using a 2×2 input window and stride 2	1062
55.10	The task in image classification is to predict a single label (or a distribution over labels as shown here – percentages indicating level of confidence) for a given input image	1062
55.11	Illustration of a single-layer convolutional neural network architecture for sentence classification	1064
55.12	Illustration of a convolutional neural network for sentence classification	1065
55.13	A simple RNN network with a feedback loop. A simple RNN, $A$ , looks at some input, $x_t$ , and outputs a value, $y_t$	1066
55.14	A sequence of simple RNNs	1067
55.15	The repeating module in a standard RNN with a single layer (A1) and an LSTM with four interacting layers (A2)	1067
55.16	The BreakingNews dataset	1069
56.1	Left panel: feature space comprised of ages and education levels, partitioned into regions that are relatively homogeneous with respect to probability of turnout (indicated with lighter shades of gray for every combination of feature values). Right panel: binary-tree representation of feature-space partition depicted on the left panel	1082
56.2	Left panel: post-stratified predictions of turnout during the 2008 presidential election in the United States at low levels of demographic aggregation. Predictions produced by a multilevel model along the y-axis, and predictions along the x-axis produced by a Bayesian additive regression tree model. Right panel: post-stratified predictions of vote intention for McCain during the 2008 presidential election in the United States at low levels of demographic aggregation	1086
56.3	Left panel: classifying by a Maximum-Margin Hyperplane: a linearly separable problem, with the corresponding separating line (solid line) and classification margins (dashed lines). Right panel: non-linearly separable classification problem of Class A (outer light gray triangles) and Class B (inner dark gray circles) instances	1087
56.4	Illustration of kernel trick: tilted projection of instances depicted in right panel of Figure 56.3 onto space with an additional dimension $z = -(X_1^2 + X_2^2)$ (original two-dimensional representation can be seen at the bottom)	1089
57.1	Calibration in the set of developed countries	1104
57.2	Other possible calibration functions	1104
57.3	Correlation (left) and subset sufficiency (right)	1108

57.4	Fuzzy correlation (left) and fuzzy subset sufficiency (right)	1109
57.5	Incomplete inclusion/large coverage (left) and complete inclusion/ low coverage (right)	1109
57.6	Unique coverage of A (hashed area)	1110
57.7	X as a trivial necessary condition for Y	1111
57.8	Bidimensional vector space (left) and the corresponding truth table (right)	1113
58.1		1120
58.2		1122
61.1	Interviews in political science journals, 2000–17	1173
61.2	Longer-term patterns in political science articles	1174
61.3	Interviews in political science books, % of responses	1175
61.4	Uses of interviews in political science books, % of responses	1175

# List of Tables

1.1	Military conflicts, 1958–65	9
6.1	Specifying the field	109
8.1	The list of variables for the model	148
8.2	Parameter values	149
13.1	Summary statistics on vetoes, 1975–2018	231
13.2	Hopeless over-ride attempts, 1975–2018	231
13.3	Massive rolls of presidential vetoes, 1975–2018	232
20.1	Example of additive scale function	357
22.1	Common components of URLs	389
22.2	Parameters of the Wikipedia query	390
23.1	Descriptive statistics from Sifter data	414
23.2	Main results	415
23.3	Robustness checks – tweets from phones	416
26.1	A map of approaches to the analysis of political text	468
26.2	Stages in analyzing text as data	478
27.1	A stylized model of political text generation and model assumptions in each step	501
29.1	Results for Liu & Hu sentiment lexicon	546
29.2	Results for VADER sentiment analysis	546
29.3	Initial results for classification algorithms	547
29.4	Results for final classification algorithm	548
30.1	Summary of layouts for network visualization	561
30.2	Summary of the algorithms for community detection in the network	562
30.3	Influence and propagation	563
32.1	Model simplification in the single equation model of approval	604
32.2	Model simplification in the VAR of presidential approval, economy, and consumer sentiment	605
34.1	PM satisfaction, vote intention and unemployment, Britain 1997–2006	638
34.2	Hostility sent, hostility received and presidential popularity, United States 1953–78	646
34.3	PM satisfaction, vote intention and unemployment, Britain 1997–2006	652
34.4	British party support and inflation, 2010–12	653
34.5	British party support and inflation, 2010–12	654
34.6	Systematic polling house biases	655
36.1	Common choices of levels	680
36.2	Level-one error structures for longitudinal data	688
36.3	A cross-classified structure	693
36.4	Cumulative response probabilities with $\gamma_{00} = 0$ and $\theta_2 = 1$	696
37.1a	Naïve estimation	708
37.1b	Corrected estimation	710
37.2a	Naïve estimation	711
37.2b	Corrected estimation	712
37.3a	Naïve estimation	713

37.3b	Corrected estimation	714
39.1	Common spatial econometric models	734
39.2	ML estimates of covariate-coefficient estimate ( $\hat{\beta}, \beta = 2, N = 300, 1,000$ trials)	740
39.3	ML estimates of interdependence ( $\hat{\rho}, \beta = 2, N = 300, 1,000$ trials)	741
39.4	ML estimates of spillover effect ( $\hat{\theta}, \beta = 2, N = 300, 1,000$ trials)	742
40.1	SEM vs. design-based IV: a comparison of the assumptions	756
41.1	True values of $\mathbf{y}_c, \mathbf{y}_t$ and $\boldsymbol{\tau}$ , where $\tau_i = y_{t,i} - y_{c,i}$ , for the study population	772
41.2	All possible realizations of experimental data from a completely randomized study with six units and three treated units	773
41.3	Finite populations under asymptotic growth in which $h \in \{1, 2, 3, 4, \dots\}$	774
41.4	True values of $\mathbf{y}_c, \mathbf{y}_t, \boldsymbol{\tau}$ and the baseline covariate $\mathbf{x}$	778
41.5	Realization of data if $\mathbf{z}_8$ were the randomly drawn assignment	783
41.6	Null potential outcomes if $\mathbf{z}_8$ were the realized assignment and under a test of the null hypothesis that $y_{t,i} = y_{c,i}$ for all $i$	783
41.7	Null potential outcomes for all possible realizations of data when the null hypothesis is true; the observed outcomes column, $\mathbf{y}_j = \mathbf{y}_{c0} + \mathbf{z}_j \boldsymbol{\tau}'$	785
41.8	No-effects – left table, $(\mathbf{y}_c, \mathbf{y}_t)$ – and positive-causal effects – right table, $y_c^*, y_t^*$	786
41.9	Observed difference-in-means test statistics under all possible assignments for both a no-effects and positive-effects true causal effect	787
41.10	Comparing $p$ -values for tests of the strong null of no effect when no effects are true $(\mathbf{y}_c, \mathbf{y}_t)$ and false $(y_c^*, y_t^*)$	787
41.11	Values of $y_c, y_t$ and $\boldsymbol{\tau}$ when weak null hypothesis is true	790
41.12	Compliance strata	792
41.13	Realization of data if $\mathbf{z}_8$ were the randomly drawn assignment	793
41.14	Null potential outcomes for a test of the strong null of no effect if $\mathbf{z}_8$ were realized assignment	793
41.15	Attrition strata	794
43.1	Standardized differences and $p$ -values for treated to control match in the pre-treatment period for the election day registration application	830
43.2	Standardized differences and $p$ -values for treated to control match in the post-treatment period for the election day registration application	831
43.3	Standardized differences and $p$ -values for treated to control match in the pair-to-pair match for the election day registration application	831
43.4	Average log hourly wages, married women 20–40 years	832
43.5	Average log hourly wages, women over 40 and single men 20–40 years	832
44.1	Continuity-based RD analysis: effect of victory at $t$ on vote margin at $t + 1$ . Incumbent party, Brazilian mayoral elections, 1996–2012	853
44.2	Minimum $p$ -value in first 20 symmetric windows around cutoff running variable is vote margin at $t$ of incumbent party, Brazilian mayoral elections, 1996–2012	854
44.3	Local randomization RD analysis: effect of victory at $t$ on vote margin at $t + 1$ . Incumbent party, Brazilian mayoral elections, 1996–2012	854
46.1	Summary statistics: networks	884
46.2	Summary statistics: independent variables	885
46.3	Replication results	887
46.4	Plots of latent space positions for the LSMs along two dimensions	888

46.5	SBM block assignment plots	889
47.1	Summary by journals: the number of journal articles containing ‘Bayesian’ in the abstract excluding game theory-related articles	900
47.2	Summary by topics: top 15 words sorted by the topic-word probability using text2vec package (Selivanov and Wang, 2018)	901
47.3	Bayesian R Packages: Bayesian packages hosted in CRAN	901
49.1	Evaluating theories of Congress: median estimates and 95% credible intervals for Table 3 in Richman (2011) (Models 1–4) and two garbage-can models	940
49.2	BIC and approximation to Bayes factor	945
49.3	Log marginal likelihood and Bayes factor using bridge sampling	946
49.4	Information criteria for evaluating theories of Congress	951
49.5	Loo information criteria for evaluating theories of Congress with standard errors	952
49.6	Model weights based on stacking	955
50.1	Posterior quantiles, ordered model for survey of political executives	969
50.2	Posterior quantiles, GLMDM for survey of political executives	980
53.1	Equilibrium prices for first ( $\pi_n^1$ ) and second ( $\pi_n^2$ ) moving customers and average treatment effect ( $\tau_n$ ) of the customer moving first for games with $n$ possible rounds	1023
53.2	Estimation of model parameters using the correct model	1026
53.3	Estimation of average treatment effects using difference-in-means (DIM) and parameter estimation (MLE)	1026
53.4	Extrapolation	1027
53.5	Diagnosis of design without randomization	1029
53.6	Diagnosis of design with incorrect model (assume $q = 0$ )	1030
53.7	Diagnosis of design with incorrect yet observationally equivalent model	1030
55.1	Overview of machine learning methods and examples from political science	1056
56.1	Popular SVM kernels	1089
57.1	Set intersection for crisp sets	1099
57.2	Set union for crisp sets	1099
57.3	Set negation for crisp sets	1100
57.4	Per capita income (INC), calibrated to fuzzy sets membership scores (fsMS)	1103
57.5	Boolean minimization example	1113
60.1	Leadership trait analysis	1156
60.2	The ‘norming group’ of leadership trait analysis	1158



# Notes on the Editors and Contributors

## THE EDITORS

**Luigi Curini** is Professor of Political Science at the University of Milan, Visiting Professor at Waseda University of Tokyo, and chair of MetRisp (the Standing Group on Political Methods and Research of the Italian Political Science Association). His research focuses on comparative politics, valence politics and text analytics. He is the (co-)author of seven books, including *Corruption, Ideology, and Populism* (2018) and *Politics and Big Data* (2017), and more than 50 articles and chapters.

**Robert Franzese** is Professor and Associate Chair of Political Science, Director of the Program in International and Comparative Studies, Research Professor in the Center for Political Studies of the Institute for Social Research at the University of Michigan and a past President of the Society for Political Methodology. He is the (co-)author or (co-)editor of seven books or collections, including the forthcoming *Empirical Analysis of Spatial Interdependence*, and more than 50 articles, chapters and papers, including most germane to this *Handbook*, ‘Spatial-Econometric Models of Cross-Sectional Interdependence in Political-Science Panel and Time-Series-Cross-Section Data’, ‘A Spatial Model Incorporating Dynamic, Endogenous Network Interdependence: A Political Science Application’ and ‘Interdependence in Comparative Politics: Substance, Theory, Empirics, Substance’.

## THE CONTRIBUTORS

**James Adams** is Professor of Political Science at UC Davis. He received a PhD in Political Science from The University of Michigan, and a BA in economics from Princeton University. His research analyzes voting behavior, parties’ vote-seeking strategies and election outcomes. He has authored two books: *Party Competition and Responsible Party Government* and *A Unified Theory of Party Competition* (with Bernard Grofman and Samuel Merrill III). He has published over 50 papers in peer-reviewed journals including the *American Political Science Review*, the *American Journal of Political Science* and the *Journal of Politics*.

**John Aldrich** is the Pfizer-Pratt University Professor of Political Science at Duke University. His research focuses on American politics, political parties, Congress and elections, and on EITM-based research using both formal theory and quantitative empirical research.

**Pablo Barberá** is an Assistant Professor in the School of International Relations at the University of Southern California. His research agenda examines how the adoption of social media platforms is transforming different aspects of democratic politics: ideological polarisation, incivility in political discussions, politicians’ responsiveness to their constituents, the emergence of social protests, public diplomacy, etc. To answer these questions, he develops

new network science and text analysis methods that can provide researchers with the necessary tools to collect and analyse digital trace data.

**Deborah Beim** is an Assistant Professor in the Department of Political Science at the University of Michigan. She studies American politics in general and judicial politics in particular, with a focus on interactions between the US Supreme Court and the Courts of Appeals. She is also interested in applied game theory. Her work has been published in the *American Journal of Political Science*, *Journal of Politics* and other outlets.

**Ken Benoit** is a Professor, Part-Time, at the Australian National University, and he has previously held positions in the Department of Political Science at Trinity College Dublin. He received his PhD (1998) from Harvard University, Department of Government. His current research focuses on computational, quantitative methods for processing large amounts of textual data, mainly political texts and social media. His current research includes software development for text analysis (the R package *quanteda*) and the analysis of big data, including social media, and methods for text mining using combinations of crowd-sourced annotations, statistical methods and machine learning.

**Ravi Bhavnani** is a Professor in the International Relations and Political Science Department at the Graduate Institute. He received his PhD in Political Science (Comparative Politics and Methodology) and a Degree Certificate in Complex Systems from CSCS from the University of Michigan, Ann Arbor. His research explores the micro-foundations of violent conflict by means of agent-based computational modelling and disaggregated empirical analysis. More specifically, his work examines the endogenous relationships among: 1) the characteristics, beliefs and interests of relevant actors; 2) social mechanisms and emergent social structures that shape attitudes, decision-making and behaviour; and 3) patterns of conflict and violence.

**Berit Bliesemann de Guevara** is a Reader in International Politics and the Director of the Centre for the International Politics of Knowledge at Aberystwyth University's Department of International Politics. Her current research explores ways and problems of knowing in international politics, specifically in conflict and intervention contexts, as well as the use of arts-based methods in social-scientific research. She has used interpretive approaches in studies of the role of politicians' field visits in foreign and defence policymaking, myths in international politics and international peacebuilding in the Balkans. Her current hermeneutical research includes an Arts and Humanities Research Council (AHRC) project on local conflict knowledge among conflict-affected communities in Myanmar, and a Newton Fund-Colciencias project on demobilised guerrilla fighters' subjectivities in their reincorporation process into Colombian society.

**Frederick J. Boehmke** is Professor of Political Science at the University of Iowa and Director of the Iowa Social Science Research Center in the Public Policy Center. His research focuses on American state politics and quantitative research methods. Lately, this has included a particular focus on state policy adoption and diffusion, including developing the SPID database, featuring over 700 policy diffusion episodes and a series of articles evaluating methods to leverage multiple policies to better understand policy diffusion.

**Tobias Böhmelt** is a Professor of Government at the University of Essex and Associate Editor of the *British Journal of Political Science*. His current research interests are based in the areas of transnational diffusion processes, party politics, migration and population movements, environmental politics, international mediation and civil-military relations. His recent work has been published in the *American Political Science Review*, *American Journal of Political*

*Science, Journal of Politics, International Organization, International Studies Quarterly, European Journal of Political Research* and others.

**Sarah B. Bouchat** is an Assistant Professor at Northwestern University, affiliated with the Department of Political Science and Northwestern Institute for Complex Systems. They completed a PhD in the Department of Political Science at the University of Wisconsin–Madison in 2017. With research interests in political methodology, comparative political economy and authoritarian politics, Sarah's current work focuses on text-as-data, Bayesian modelling and machine learning, with applications to the study of low-information, authoritarian regimes like Myanmar.

**Jake Bowers** (<http://jakebowers.org>) is an Associate Professor of Political Science and Statistics University of Illinois at Urbana-Champaign. He was a Fellow on the White House Social and Behavioral Sciences Team (<http://sbst.gov>) and Office of Evaluation Sciences in the General Services Administration of the US Government (<http://oes.gsa.gov>) from 2015–19. He is Methods Director and a board member for the Evidence in Government and Politics network (<http://egap.org>), co-founder of Research4Impact (<http://r4impact.org>) and a research affiliate with the Center for Advanced Studies in the Behavioral Sciences (<http://casbs.stanford.edu>). His research in applied statistics and political methodology focuses on questions of statistical inference for causal effects in randomised experiments and observational studies.

**Thomas Bräuninger** is a Professor of Political Economy at the University of Mannheim and Associate Editor of the *American Political Science Review*. His area of research is comparative politics and formal theory with a focus on the effect of political institutions on policy outcomes. His current research focusses on voting behaviour, party competition, electoral systems, legislative politics and propaganda. His work has been published in, among other journals, the *British Journal of Political Science, Journal of Politics, Journal of Conflict Resolution, Legislative Studies Quarterly, Political Analysis* and *Political Science Research and Methods*.

**Klaus Brummer** holds the chair of International Relations at the Catholic University of Eichstätt-Ingolstadt. He is co-editor-in-chief of the journal *Foreign Policy Analysis* and served as president of the Foreign Policy Analysis section of the International Studies Association (ISA) in 2015–16. He has published in peer-reviewed journals such as the *British Journal of Politics and International Relations, Foreign Policy Analysis, Government and Opposition, International Politics* and the *Journal of European Public Policy*, and he is the co-editor of *Foreign Policy Analysis beyond North America* (2015) and *Foreign Policy as Public Policy?* (2019).

**Ernesto Calvo** (PhD, Northwestern University 2001), Professor of Government and Politics at the University of Maryland, Director of the Interdisciplinary Lab for Computational Social Science (iLCSS) and Field Editor of the *Journal of Politics* (JOP-Comparative). His work uses big data to study comparative political institutions, political representation and social networks. He is the author of *Non-Policy Politics* (2019), *Legislator Success in Fragmented Congresses in Argentina* (2014) and over 50 publications in Latin America, Europe and the United States. The American Political Science Association has recognized his research with the *Lawrence Longley Award*, the *Leubbert Best Article Award* and the *Michael Wallerstein Awards*.

**Charles Cameron** is Professor of Politics and Public Affairs at Princeton University. He is the author of many articles in leading journals of political science as well as *Veto Bargaining:*

*Presidents and the Politics of Negative Power* (2000), which won the American Political Science Association's Fenno Prize and William Riker Award. He has been a Research Fellow at the Brookings Institution, a National Fellow at the Hoover Institution, a Visiting Scholar at Princeton's Center for the Study of Democratic Politics and he has a recurrent visiting affiliation as Professor at New York University School of Law.

**Christopher L. Carter** is a PhD candidate in political science at the University of California, Berkeley and research associate at Berkeley's Center on the Politics of Development. His current work adopts a multi-method approach, combining natural experiments, surveys, lab-in-the-field experiments, archival data and extensive interviewing to examine indigenous-state relations in the Americas. He has also published on party systems in Latin America, decentralization in the developing world and regulation of the gig economy in the United States. He received a Master's in Latin American Studies at the University of Cambridge as a Gates-Cambridge scholar. He also holds a B.A. in Political Science and History from the University of North Carolina at Chapel Hill, where he studied as a Morehead-Cain scholar.

**Matias D. Cattaneo** is Professor of Operations Research and Financial Engineering at Princeton University. His research spans mathematical statistics, econometrics, data science and quantitative methods in the social, behavioural and biomedical sciences. Matias earned a PhD in economics in 2008 and an MA in statistics in 2005 from the University of California at Berkeley. He also completed an MA in economics at Universidad Torcuato Di Tella in 2003 and a BA in economics at Universidad de Buenos Aires in 2000. Prior to joining Princeton in 2019, he was a faculty member in the departments of economics and statistics at the University of Michigan for 11 years.

**Scott de Marchi** is Professor of Political Science and Director of the Decision Science program at Duke University. His work is currently funded by the Department of Defense, the National Science Foundation, and USAID. His research focuses on mathematical methods, especially bargaining theory, computational social science, machine learning, and mixed methods. Substantively, he examines decision-making in contexts that include the American Congress, coalition and crisis bargaining, interstate conflict, and voting behaviour. He is also collaborating with the DevLab@Duke to use machine learning models to forecast regime backsliding and civic space closures.

**Kakia Chatsiou** is a Senior Research Officer at the Institute for Social and Economic Research, University of Essex. Her research focuses on public administration, data science and natural-language processing. In her research, she studies how data can fuel positive changes to society and individuals and how the challenges of artificial intelligence, ethics, privacy and safeguarding can be overcome through compliance and transparency. Previously, she was a researcher at the Department of Government, where she worked with local authorities and voluntary-sector organisations in the UK building capacity in text analytics and data sharing as part of the Catalyst Project. She is a member of the Berkeley Initiative for Transparency in the Social Sciences and a member of the Wider Eastern Information Stakeholder Forum, UK. She holds a PhD in computational linguistics from the University of Essex.

**William Roberts Clark** is Charles Puryear Professor in Liberal Arts at Texas A&M University. He is currently a Weatherhead Scholar at the Weatherhead Center of International Affairs and Visiting Professor in the Department of Government at Harvard University. He previously taught at the University of Michigan, New York University and Georgia Institute of Technology.

He is the author, with Matt and Sona Golder, of *Principles of Comparative Politics* and the author of *Capitalism, Not Globalism: Capital Mobility, Central Bank Independence, and the Political Control of the Economy*. He has published papers in prominent journals on subjects such as the politics of monetary policy, electoral laws and party systems, the resource curse and the testing of context-dependent hypotheses.

**Scott J. Cook** is an Assistant Professor and Co-Director of the Research Methods Program in the Department of Political Science at Texas A&M University. He earned his PhD at the University of Pittsburgh (2014), where his doctoral work received the John T. Williams Dissertation Prize from the Society for Political Methodology. His research has appeared in the *American Journal of Political Science*, *Political Analysis* and *Political Science Research and Methods*, among other journals. His current work includes the forthcoming *Empirical Analysis of Spatial Interdependence* and *National Science Foundation-supported research on spatiotemporal patterns of terrorism*.

**David Darmofal** is an Associate Professor of Political Science at the University of South Carolina. His research interests focus on American politics and political methodology. He has substantive interests in voting behaviour, public opinion and political geography, and methods interests in spatial analysis, survival analysis and time series analysis. He is the author of *Spatial Analysis for the Social Sciences* and *Demography, Politics, and Partisan Polarization in the United States, 1828–2016*, the latter co-authored with Ryan Strickler. His research has also appeared in a variety of journals, including the *American Journal of Political Science*, *Journal of Politics*, *Political Geography*, *Political Behavior*, *Political Psychology*, *Political Research Quarterly* and *State Politics and Policy Quarterly*.

**Bruce A. Desmarais** is the DeGrandis-McCourtney Early Career Professor in Political Science, Associate Director of the Center for Social Data Analytics, and an Affiliate of the Institute for Cyber Science at Penn State University. His research is focused on methodological development and applications that further our understanding of the complex interdependence that underlies politics, policymaking, and public administration. Methodologically, Bruce focuses on methods for modeling networks, analyzing dynamics on networks, and experiments on networks. Primary application areas of interest to Bruce include public policy diffusion, campaign finance, legislative networks, and internal government communication networks.

**Gustavo Diaz** is a PhD candidate in the Department of Political Science at the University of Illinois at Urbana-Champaign. He studies Comparative Politics and Political Methodology, with focus on the political economy of developing countries and improving causal inference in experimental and observational studies. His research has been funded by the Lemann Institute for Brazilian Studies.

**Karsten Donnay** is an Assistant Professor in the Department of Political Science at the University of Zurich. In his research, he combines an interest in data science, (big) data analytics, statistical and computational modeling with applied interdisciplinary research across different disciplines in the social and behavioral sciences. Karsten holds a PhD from ETH Zurich and was a postdoctoral researcher at the Graduate Institute Geneva and the University of Maryland. Prior to joining the University of Zurich in 2020, he was an Assistant Professor in the Department of Politics and Public Administration at the University of Konstanz.

**Thad Dunning** is Robson Professor of Political Science at the University of California, Berkeley. He studies comparative politics, political economy and research methodology.

Dunning has written on methodological topics including causal inference, statistical analysis and multi-method research. He chaired the inaugural project of the Evidence in Governance and Politics (EGAP) group's Metaketa Initiative, which aims to achieve greater cumulation of findings from experimental research on international development and political accountability; the project resulted in an article in *Science Advances* and a book at Cambridge University Press. His methods work also includes *Natural Experiments in the Social Sciences: A Design-Based Approach* (2012, Cambridge University Press), which won the Best Book Award from the American Political Science Association's Experiments Section. He received a PhD in political science and an MA in economics from the University of California, Berkeley (2006) and was formerly Professor of Political Science at Yale University.

**Adrian Duşa** works at the Department of Sociology at the University of Bucharest, teaching social statistics, research methodology and measurement in the social sciences. He is also the President of Romanian Social Data Archive (RODA) – having been involved in data archiving for almost 20 years, in partnership with the Council of European Social Science Data Archives (CASSDA). His primary research interest is qualitative comparative analysis (QCA) – a novel methodology for social research that inspires a growing community, which he serves as a member of the Advisory Board and Steering Committee at COMPASSS, and he publishes extensively on methodological aspects, with a key interest in the programming of the minimisation algorithms. His main programming environment is R, to which he contributed a number of packages, including a highly advanced and sophisticated QCA software.

**Christopher Eddy** is a PhD student in the Department of Political Science at the University of South Carolina. His research and teaching interests are in American politics and public administration/public policy.

**Benjamin C.K. Egerod** an assistant professor at the Department of International Economics, Government and Business, Copenhagen Business School and a Bradley Fellow at the Stigler Center, Booth School of Business, University of Chicago. In his research, he focuses on business–government interactions. Because these typically are very difficult to observe, much of his research concerns the development of new tools for measuring them on a large scale – for example, through the use of political text.

**Robert A. Fahey** is a Research Associate at the Waseda Institute of Political Economy (WINPEC) at Waseda University in Tokyo. His research focuses on the use of text mining and network analysis methods to analyse behaviour and opinion-forming on social media, and the use of social media data to supplement and improve public-opinion research. His English-language work on these topics has been published in journals including *Social Science and Medicine* and the *International Journal of Information Management*, and he has presented research into the spread of populist ideas on social media in Japan at major conferences, including the *APSA Annual Meeting* and *ECPR General Conference*.

**Christopher J. Fariss** is an Assistant Professor in the Department of Political Science and Faculty Associate in the Center for Political Studies at the University of Michigan. Prior to these appointments, he was the Jeffrey L. Hyde and Sharon D. Hyde and Political Science Board of Visitors Early Career Professor in Political Science in the Department of Political Science at Penn State University. He is also an Affiliated Scholar at the Security and Political Economy (SPEC) Lab at the University of Southern California. In June 2013, he graduated with a PhD in political science from the University of California, San Diego. He also studied

at the University of North Texas, where he graduated with an MS in political science (2007), a BFA in drawing and painting (2005) and a BA in political science (2005).

**Kentaro Fukumoto** is Professor of Political Science at Gakushuin University and was a Visiting Scholar at Harvard University and at Washington University in St. Louis. He received his PhD from the University of Tokyo in 2007. His research interests include political methodology, electoral studies, legislative studies and Japanese politics. He is the author of *Nihon no Kokkai Seiji: Zen Seifu Rippo no Bunseki* [Politics in the Japanese Diet: A Statistical Analysis of Postwar Government Legislation] (2000) and *Rippo no Seido to Katei* [Legislative Institutions and Process] (2007). His articles have appeared in the *American Political Science Review*, *American Journal of Political Science*, *Behaviormetrika*, *Electoral Studies*, *Japanese Journal of Political Science*, *Journal of American Statistical Association*, *Journal of Politics* and *Legislative Studies Quarterly*. He is an editor of the *Japanese Journal of Political Science* and is on the editorial boards of *Political Analysis* and the *Asian Journal of Comparative Politics*.

**Nathan Gibson** is a PhD candidate in the Politics Department at Princeton University and is affiliated with the Center for the Study of Democratic Politics in the Woodrow Wilson School of Public and International Affairs. His research focuses on the American presidency, bureaucracy and Congress and has been published in *Congress and The Presidency* as well as the SSRC Anxieties of Democracy volume *Can America Govern Itself?* His web page and other research can be found at <https://www.ndgibson.com>.

**Jeff Gill** is a Distinguished Professor of Government, a Professor of Statistics and a Member of the Center for Behavioral Neuroscience at American University. His research applies Bayesian modelling and data analysis (decision theory, testing, model selection, elicited priors) to questions in general social science quantitative methodology, political behaviour and institutions, medical/health data analysis (especially physiology, circulation/blood and pediatric traumatic brain injury) and epidemiological measurement/data issues using computationally intensive tools (Monte Carlo methods, MCMC, stochastic optimisation and non-parametrics).

**Mitchell Goist** is a PhD candidate in political science, with a minor in social data analytics, at Pennsylvania State University. His political science research is in comparative politics, with a particular interest in extremist politics. His methodological research is in machine learning, deep learning, text-as-data and political networks. His dissertation develops a transfer learning approach to multilingual text analysis and applies the technique to the cross-national study of populist and extremist political parties. He successfully defended his PhD in 2019.

**Ezequiel González-Ocantos** (PhD, Notre Dame, 2012) is Associate Professor in the Department of Politics and International Relations at the University of Oxford, and Professorial Fellow of Nuffield College. He is the author of *Shifting Legal Visions: Judicial Change and Human Rights Trials in Latin America* (2016), which won best book awards from the American Political Science Association, the International Studies Association and the Latin American Studies Association. His research has also appeared in the *American Journal of Political Science*, *Comparative Politics*, *Comparative Political Studies*, *International Studies Quarterly*, *Law and Society Review* and *Sociological Methods and Research*, among others. In 2018, Ezequiel received the Philip Leverhulme Prize in Politics and International Relations.

**Christopher Grady** is a PhD candidate in the Department of Political Science at the University of Illinois at Urbana-Champaign. His primary interests are political psychology, intergroup

conflict, media and observational learning, and international development. His research has been funded by the National Science Foundation, International Foundation for Electoral Systems, and Evidence in Governance and Politics.

**Jim Granato** is Executive Director and Professor at the University of Houston's Hobby School of Public Affairs. His research focuses on EITM, Political Economy, and Time Series Analysis.

**Claire Greenstein** is an Assistant Professor of Political Science at the University of Alabama at Birmingham. Her research is mainly on transitional justice, primarily reparations, with a regional focus on Europe and Latin America.

**Imke Harbers** is Associate Professor of Political Science at the University of Amsterdam. Her research focuses on subnational politics, state capacity and state–society interactions, with a specific emphasis on the territorial reach of the state. Her work has appeared, or is forthcoming in, *Political Science Research and Methods*, *Political Analysis* and *Comparative Political Studies*, among others. She has held visiting positions at the University of California, San Diego, and her current research is funded by a Marie Curie Fellowship from the European Commission.

**Jude C. Hays** is an Associate Professor of Political Science at the University of Pittsburgh. His research has been published in journals such as the *American Journal of Political Science*, *Political Analysis* and *Political Science Research and Methods*, among others. He is the author of *Globalization and the New Politics of Embedded Liberalism* and co-author of the forthcoming *Empirical Analysis of Spatial Interdependence*.

**Simon Heuberger** is a PhD candidate in political science at American University in Washington, DC, concentrating on American government and quantitative methods. His research focuses on the statistical advancement of survey measurement tools, causal inference and the psychological underpinnings of public opinion. He also works as the official replicator for *Political Analysis*.

**Florian M. Hollenbach** is an Assistant Professor in the Department of Political Science at Texas A&M University. Hollenbach's research focus is political methodology and comparative politics. In his work in political methodology, he primarily studies the consequences of spatial dependence and Bayesian statistical methods. In his substantive research, he is specifically interested the development of state/fiscal capacity and fiscal policy. Dr. Hollenbach's work has appeared or is forthcoming in the *American Political Science Review*, *the Journal of Politics*, *Political Analysis*, *Sociological Methods & Research*, *the British Journal of Political Science* and *Political Science Research & Methods*.

**Macartan Humphreys** is Professor of Political Science at Columbia University and Director of the Institutions and Political Inequality group at the WZB Berlin. He works on the political economy of development, with ongoing research focusing on post-conflict development, ethnic politics, political authority and democratic development. His recent work has appeared in the *American Political Science Review*, *Journal of Development Economics*, *Science Advances* and elsewhere. Macartan has written or co-authored books on ethnic politics, natural resource management and game theory and politics.

**Matthew C. Ingram** is Associate Professor of Political Science at the University at Albany, State University of New York. His research focuses on justice reform, violence and methods. He is the author of *Crafting Courts in New Democracies: The Politics of Subnational Judicial*



*Reform in Brazil and Mexico* (2016) and co-editor of *Beyond High Courts: The Justice Complex in Latin America* (2019). His publications have also appeared in peer-reviewed journals, including *World Development*, *Comparative Politics*, *Political Analysis* and *Political Science Research and Methods* and in several edited volumes. He has held visiting positions at the University of California, San Diego, and the University of Notre Dame, and in 2019 he won a Fulbright US Scholar award for research in Mexico. He holds a BA from Pomona College and an MA, JD and PhD from the University of New Mexico.

**Slava Jankin Mikhaylov** is Professor of Data Science and Public Policy at the Hertie School of Governance. He is the Director of the Hertie School Data Science Lab. His research and teaching is primarily in the field of natural-language processing and machine learning. Before joining the Hertie School faculty, he was a Professor of Public Policy and Data Science at University of Essex, holding a joint appointment in the Institute for Analytics and Data Science and Department of Government. At Essex, Slava served as a Chief Scientific Adviser to Essex County Council, focusing on artificial intelligence and data science in public services. He previously worked at University College London and London School of Economics. Slava holds a PhD in Political Science from Trinity College Dublin.

**Luke Keele** (PhD, University of North Carolina, Chapel Hill, 2003) is an Associate Professor at the University of Pennsylvania with joint appointments in Surgery and Biostatistics. Professor Keele specialises in research on applied statistics with a focus on causal inference, design-based methods, matching, natural experiments and instrumental variables. He also conducts research on topics in educational-programme evaluation, election administration and health-services research. He has published articles in the *Journal of the American Statistical Association*, *Annals of Applied Statistics*, *Journal of the Royal Statistical Society, Series A*, *The American Statistician*, *American Political Science Review*, *Political Analysis* and *Psychological Methods*.

**Michael R. Kenwick** is an Assistant Professor in the Department of Political Science at Rutgers University, New Brunswick. He graduated with a PhD in political science from the Pennsylvania State University in the summer of 2017. He was also a Post-doctoral Research Fellow at the University of Pennsylvania's Perry World House. His research is in the area of international relations, with emphases on conflict processes, civil–military relations and border politics. Broadly, his work develops novel measurement and research-design strategies to better understand whether and how states respond to domestic and international security threats.

**Gary King** is the Albert J. Weatherhead III University Professor at Harvard University – one of 25 with Harvard's most distinguished faculty title – and Director of the Institute for Quantitative Social Science. King develops and applies empirical methods in many areas of social science, focusing on innovations that span the range from statistical theory to practical application. He is an elected Fellow in 8 honorary societies and has won more than 55 prizes and awards; he has written 175 journal articles, 20 open-source-software packages 8 books, and 14 patents. He helped reinvigorate the modern quantitative and qualitative methods subfields in political science; created the standards and methods widely used to evaluate partisan and racial redistricting; implemented 'politically robust' designs that make possible research in difficult circumstances, including the largest-ever experiments in media studies and in health policy; reverse-engineering Chinese censorship, industry–academia relationships, and methods for interpersonal incomparability in surveys. These and his many other projects are among the most widely cited in political science and across fields.

**Robert Klemmensen** is a Professor of Political Science at the University of Southern Denmark. He has published broadly on topics related to political psychology and political responsiveness. Klemmensen's work draws heavily on text-as-data methods for estimating elite policy positions and relating them to mass preferences.

**Elsa T. Khwaja** is a Doctoral Candidate in public policy at the Schar School of Policy and Government at George Mason University. Her research involves international development policy and aid effectiveness in fragile and conflict-affected zones. She holds a Master's Degree in Public and International Affairs from the University of Pittsburgh Graduate School of Public and International Affairs and a Bachelor's Degree in Global Affairs and Political Science from the University of Minnesota, Twin Cities.

**James H. Kuklinski** is Professor Emeritus in the Department of Political Science at the University of Illinois at Urbana-Champaign. He is a recipient of the Hazel Gaudet Erskine Career Achievement Award by the Political Psychology Section of the American Political Science Association. His primary interests include the nature and quality of citizen decision-making, the relationship between public opinion and legislative policymaking, and the use of experiments in social scientific research.

**Xymena Kurowska** is Associate Professor of International Relations at Central European University in Budapest and Vienna. She works within international political sociology and at the intersection of psychoanalysis and politics, with particular focus on security theory and practice, subjectivity, ideological formations and interpretive methodologies. She has applied interpretive methods in her fieldwork for a variety of projects in European security and border security policy, as well as for exploring the role of researchers' reflexivity in relational knowledge production. Her current interpretive research includes the study of norms in cyber diplomacy for the European Commission-funded project EU Cyber Direct and refining instruction in interpretive methods in International Relations at Central European University.

**Scott J. LaCombe** is a PhD candidate in political science at the University of Iowa. Scott focuses on American politics and political methodology, with an emphasis on the role of institutions in state politics, and he has published papers on policy diffusion as well as the role of direct democracy in state politics. In his dissertation, he studies the role of state institutions, measuring how states design their institutional configuration, how institutional design mediates the relationship between public opinion and policy and how institutional design affects how citizens perceive state governments. Within the diffusion literature, he researches new ways to understand state policy diffusion, including the role of state similarity, how broadband internet has changed diffusion networks and how a policy's ideological appeal alters diffusion pathways.

**Thomas Leavitt** is a PhD candidate in political science at Columbia University, where he specialises in methodology and comparative politics. His current research develops methods in design-based causal inference and Bayesian statistics. He applies these methods to substantive questions on historical political economy, racial and ethnic politics and political transitions, with a regional focus on South Africa and sub-Saharan Africa more broadly. His field experiments and qualitative research in Africa have been funded by the Center for the Study of Development Strategies (CSDS), the Center for Development Economics and Policy (CDEP) and the Abdul Latif Jameel Poverty Action Lab (J-PAL).

**Lucas Leemann** is an Assistant Professor for Comparative Politics at the University of Zurich. He obtained his PhD at Columbia University in 2014 and started as a lecturer at University College London. In 2016, he joined Essex University as a Reader. His research centres on democratic institutions, representation and data science. In comparative politics, he focuses on representation and political institutions that enable representation. Within data science, he has worked extensively on measuring attitudes. His past research has been published in the *American Political Science Review*, *American Journal of Political Science*, *Journal of Politics* and *Political Analysis*.

**Suzanna Linn** is a Liberal Arts Professor of Political Science at Penn State University. Her research focuses on time series methodology and the dynamics of American public opinion and election outcomes, with a current focus on developing tests for long-run equilibria in time series analysis and expanding theories of retrospective voting to include the effects of health and well-being outcomes. She is a Fellow of the Society for Political Methodology and president of the Society for Political Methodology (2019–21). Her work has appeared in the *American Political Science Review*, *American Journal of Political Science*, *Political Analysis*, *Statistics in Medicine* and other journals. Her book *The Decline of the Death Penalty and the Discovery of Innocence*, with Frank Baumgartner and Amber E. Boydston, received the Gladys M. Kammerer Award for the best book on US national policy.

**Lanny W. Martin** is a Professor of Political Science in the Department of Social and Political Sciences at Bocconi University, in Milan, Italy. He is also a Resident Research Fellow in the Politics and Institutions unit of the Dondena Centre for Research on Social Dynamics and Public Policy.

**Adam McCauley** is a Doctoral Candidate in the Department of Politics and International Relations at the University of Oxford. He is currently a Stipendiary Lecturer in Politics at Brasenose College (Oxford) and a Senior Lecturer at the Royal Military Academy at Sandhurst. His research focuses on insurgencies and political violence, with a particular emphasis on how violent non-state groups adapt and survive. Prior to Oxford, Adam worked as a journalist reporting on conflict and international security. His journalistic work has been nominated for numerous industry awards and has appeared in *The New York Times*, *The Atlantic*, *The New Yorker*, *TIME Magazine* and *Al Jazeera*.

**Rose McDermott** is the David and Mariana Fisher University Professor of International Relations at Brown University and a Fellow in the American Academy of Arts and Sciences. She works in the area of political psychology. She received her PhD (political science) and MA (experimental social Psychology) from Stanford University and has taught at Cornell and UCSB. She has held fellowships at the Radcliffe Institute for Advanced Study, the Olin Institute for Strategic Studies and the Women and Public Policy Program, all at Harvard University. She has been a Fellow at the Stanford Center for Advanced Studies in the Behavioral Sciences twice. She is the author of five books, a co-editor of two additional volumes and the author of over 200 academic articles across a wide variety of disciplines encompassing topics such as gender, experimentation, intelligence, identity, emotion and decision-making, cyber security and the biological and genetic bases of political behaviour.

**Adam Meirowitz** is the Kem C. Gardner Professor of Business at the David Eccles School of Business, University of Utah. He is Director of the Marriner S. Eccles Institute for Economics and Quantitative Analysis. His research focuses on applied game theory and political economy

**Samuel Merrill III** is Professor Emeritus of Mathematics and Computer Science at Wilkes University. He received a PhD in Mathematics from Yale University and an MS in Statistics from Penn State University. His current research involves mathematical and statistical modeling in political science, particularly spatial models, party competition, political cycles and polarization. He is the author of three books, including *A Unified Theory of Party Competition* (with James Adams and Bernard Grofman). He has published over 60 research papers in a number of journals, including the *American Political Science Review*, the *American Journal of Political Science* and the *British Journal of Political Science*.

**Jørgen Møller** is Professor in the Department of Political Science, Aarhus University. His research interests include the conceptualisation of democracy and the rule of law, dynamics of democratisation, conflict and democratic stability, regime change and international order, state formation and comparative methodology. His work has been published in journals such as *International Studies Quarterly*, the *Journal of Democracy* and *Sociological Methods and Research*, and in books with Routledge, Palgrave Macmillan and Oxford University Press. He is currently completing a book (with Agnes Cornell and Svend-Erik Skaaning) on *Democratic Stability in an Age of Crisis*.

**Burt L. Monroe** is Liberal Arts Professor of Political Science, Social Data Analytics and Informatics at Pennsylvania State University. He is Director of the Center for Social Data Analytics, Head of the Program in Social Data Analytics and Chief Scientist for the McCourtney Institute of Democracy's Mood of the Nation Poll. His research is in comparative politics, examining political communication and the impact of electoral and legislative institutions on political behaviour and outcomes, and methodology, especially text-as-data and other data-intensive and computationally intensive settings at the intersection of data science and social science. He is particularly interested in the development of multilingual text-as-data techniques to study democratic representation, party competition and political opposition through parliamentary speech.

**Jacob M. Montgomery** is an Associate Professor at Washington University in St. Louis in the Department of Political Science. Montgomery's research focuses on incorporating advanced computational methods into core social science tasks including measurement, causal inference and prediction. His work especially focuses on Bayesian methods including publications on Bayesian model averaging, ensemble Bayesian model averaging and Bayesian causal inference. He also researches American politics with a specific focus on American parties. He has published articles in the *American Political Science Review*, the *American Journal of Political Science*, *Political Analysis* and the *Journal of Politics*.

**James D. Morrow** is AFK Organski Collegiate Professor of World Politics and Research Professor at the Center for Political Studies, both at the University of Michigan, having also taught at the Graduate Institute of International Studies in Geneva, Stanford University, the University of Rochester and Michigan State University, and visited at Nuffield College, University of Oxford. His research addresses crisis bargaining, the causes of war, military alliances, arms races, power-transition theory, links between international trade and conflict, the role of international institutions, international law and domestic politics and foreign policy. He is the author of *Order within Anarchy*, *Game Theory for Political Scientists*, co-author of *The Logic of Political Survival* and he has contributed over 30 articles in refereed journals and 30 other publications. Professor Morrow is a member of the American Academy of Arts and Sciences. He received the Karl Deutsch Award from the International Studies Association in

1994. He was President of the Peace Science Society in 2008–9 and has held fellowships from the Social Science Research Council and the Hoover Institution.

**Rebecca Morton** is a Professor of Politics with a joint appointment between NYUNYC and NYU Abu Dhabi. She is also Associate Dean of Social Science and the Director of the Social Science Experimental Laboratory at NYUAD. She is the author or co-author of four books and numerous journal articles on experimental methods, which have appeared in outlets such as the *American Economic Review*, *American Journal of Political Science*, *American Political Science Review*, *Journal of Law and Economics*, *Journal of Politics* and the *Review of Economic Studies*. She was a co-founding editor of the *Journal of Experimental Political Science*, a co-founder of the NYU-CESS Experimental Political Science Annual Conference, held annually at NYUNYC, and the founder of the Winter Experimental Social Science Institute, held annually at NYUAD. She is currently an Advisory Editor at Games and Economic Behavior and the Chair of the Experimental Research Section of the American Political Science Association.

**Layna Mosley** is Professor in the Department of Political Science at the University of North Carolina at Chapel Hill. Her research focuses on the politics of the global economy, including multinational production and global supply chains, labour rights and sovereign borrowing. She is the editor of *Interview Research in Political Science* (2013).

**Gerardo L. Munck** is Professor of Political Science and International Relations at the University of Southern California (USC). His research focuses on democracy and democratisation, Latin America, methodology and the science of social science. His books include *Measuring Democracy: A Bridge Between Scholarship and Politics* (2009), *Regimes and Democracy in Latin America* (2007), *Passion, Craft, and Method in Comparative Politics* (with Richard Snyder, 2007) and *Authoritarianism and Democratization. Soldiers and Workers in Argentina, 1976–83* (1998). He is currently completing two books: *How Advances in the Social Sciences Have Been Made: The Study of Democracy and Democratization Since 1789* and (with Sebastián Mazzuca) *A Middle-quality Institutional Trap: Democracy and State Capacity in Latin America*.

**Eric Neumayer** is Professor of Environment and Development at the London School of Economics and Political Science (LSE). He is currently the School's Pro-Director (P-VC) Faculty Development, overseeing the School's recruitment, review, promotion, retention and pay policies, but will move to the Pro-Director Planning and Resources role in September 2019. He studied economics, political science and development studies in Germany and at the LSE. His main research interests lie in environmental economics, international political economy and research methods. He has published widely in a range of journals across different social science disciplines and he is the author of four books, most recently *Robustness Tests for Quantitative Research* (with Thomas Plümper), Cambridge University Press, 2017.

**Richard A. Nielsen** is an Associate Professor in the Department of Political Science at the Massachusetts Institute of Technology. He studies and teaches on Islam, political violence, human rights, economic development and research methods. Richard's first book, *Deadly Clerics*, offers a new explanation for why some Muslim clerics adopt the ideology of militant Jihad while most do not. His current book project explores how the internet is changing the nature of Islamic authority. Richard's other research has been published or is forthcoming in the *American Journal of Political Science*, *International Studies Quarterly*, *Political Analysis* and *Sociological Methods and Research*. He holds a PhD in government (2013) and an AM in

statistics (2010) from Harvard University, and a BA in political science (2007) from Brigham Young University.

**Dominic Nyhuis** is a Post-Doctoral Researcher at the Institute of Political Science, Leibniz University Hannover. He conducts research on parties, parliaments and subnational politics. Methodologically, he focuses on automated web data collection, quantitative methods for the social sciences and quantitative text and video analysis.

**Santiago Olivella** received his PhD in Political Science from Washington University in St Louis. He specialises in defining and implementing Bayesian latent-variable models, with a wide range of applications in electoral and legislative politics. His work has appeared in the *American Journal of Political Science*, *Journal of Politics* and the *British Journal of Political Science*, among others. He is Assistant Professor of Political Science at the University of North Carolina, Chapel Hill.

**Jong Hee Park** is a Professor in the Department of Political Science and International Relations, Seoul National University. He is the director of International Relations Data Center, Seoul National University. His research covers Bayesian statistics, time series analysis, network analysis, panel data analysis and international political economy. His work appears in many journals including *American Journal of Political Science*, *Bayesian Analysis*, *Network Science* and *Social Science History*.

**B. Guy Peters** is Maurice Falk Professor of Government at the University of Pittsburgh and founding President of the International Public Policy Association. He is also editor of the *International Review of Public Policy* and associate editor of the *Journal of Comparative Policy Analysis*. His most recent books are *Policy Problems and Policy Design* (2018), *The Politics of Bureaucracy* (7th edition, 2017), *Governance and Comparative Politics* (with Jon Pierre, 2016), and *Institutional Theory in Political Science* (4th edition, 2019).

**Maxfield J. Peterson** is a PhD candidate in Political Science at the University of Pittsburgh. His work focuses on the intersection of political economy and institutional design, with a topical emphasis on energy and environment. His dissertation explores how political incentives and patronage impact energy policy in sub-Saharan Africa through a mixed-methods approach that includes interviews, qualitative case comparisons, archival research and quantitative analysis. Max's research has appeared in *Political Studies Review* (forthcoming). Prior to graduate study, Max worked in the financial industry, and received his Bachelor of Arts in Politics with distinction from Willamette University in Salem, Oregon.

**Mark Pickup** is an Associate Professor in the Department of Political Science at Simon Fraser University. Mark is a specialist in political behaviour, political psychology and political methodology. Substantively, his research primarily falls into three areas: political identities and political decision-making; conditions of democratic responsiveness and accountability; and polls and electoral outcomes. His research focuses on political information, public opinion, political identities, norms and election campaigns within North American and European countries. His methodological interests concern the analysis of longitudinal data (time series, panel, network, etc.), with secondary interests in Bayesian analysis and survey/lab experiment design.

**Thomas Plümper** is Professor of Quantitative Social Research at the Vienna University of Economics. He studied political science and economics at the Free University of Berlin. His main research interests lie in social science methodology and in comparative and international

political economy. He has widely published in a diverse set of journals across a range of social science disciplines on research methodology and research designs, natural disasters, the political economy of international economic relations and international conflict. He has authored three books including *Robustness Tests for Quantitative Research* (with Eric Neumayer), Cambridge University Press, 2017. He also co-founded the European Political Science Association and currently serves as the association's vice-president.

**Kris Ramsay** is Professor of Politics at Princeton University. He is Co-director of the Program in Quantitative and Analytical Political Science at Princeton, Director of the PhD Program in Political Economy and the Director of the Emerging Scholars in Political Science Program. He specialises in strategic analysis and its applications to violent conflict, war and political economy.

**Mirko Reul** is a PhD Candidate at the Graduate Institute for International and Development Studies Geneva. He holds a Master's degree from the Graduate Institute and was a Fulbright Scholar at American University, Washington, DC. His dissertation project on popular allegiance in social conflicts is funded by the Swiss National Science Foundation and draws on a broad range of methods, including evidence-driven computational modelling, archival research and a lab experiment.

**Kevin Reuning** is an Assistant Professor of Political Science at Miami University in Oxford, Ohio. He graduated with a PhD in political science from the Pennsylvania State University in the summer of 2018. His research and teaching focus on political parties and social movements in the United States as well as latent-variable modelling and social network analysis.

**Chiara Ruffa** is Academy Fellow at the Department of Peace and Conflict Research at Uppsala University and Associate Professor in War Studies at the Swedish Defense University. Chiara's research interests lie at the crossroads between political science, sociology and peace and conflict research, with a specific focus on ideational variables, such as cultures, norms and frames, civil-military relations and soldiers on peacekeeping missions. Her work has been published or is forthcoming in the *European Journal of International Relations*, *Security Studies*, *Acta Sociologica*, *International Peacekeeping*, *Armed Forces and Society*, *Security and Defence Analysis*, *Small Wars and Insurgencies*, *Comparative European Politics* and several edited volumes. She is the author of *Military Cultures in Peace and Stability Operations* (University of Pennsylvania Press, 2018) and *Composing Peace* (with Vincenzo Bove and Andrea Ruggeri, Oxford University Press, 2020). She is an editorial board member of *Armed Forces and Society*.

**Andrea Ruggeri** is Professor in Political Science and International Relations and Director of the Centre for International Studies at the University of Oxford. He joined Brasenose College and the Department of Politics and International Relations in 2014. Previously, from 2010, he was Assistant Professor of International Relations at the University of Amsterdam. He holds a PhD in government (Essex, 2011), an MA in international relations (Essex, 2006) and a BA in diplomatic and international sciences (Genova, 2005).

**John P. Schoeneman** is a PhD Candidate in political science and social data analytics at Pennsylvania State University. His substantive research interest is international political economy, with a particular interest in trade and foreign direct investment networks. His methodological research is in social network analysis, machine learning, and deep learning. His dissertation applies social network analysis to better understand structural dependence in international corporate networks. He expects to complete his PhD in December 2019.

**Sooahn Shin** is a PhD student in the Department of Government, Harvard University, studying political methodology and political economy.

**Kelsey Shoub** is an Assistant Professor of Political Science at the University of South Carolina, Columbia. Her work focuses on the study of the public-policy process in the United States using big data and text analysis methods. She is a co-author of *Suspect Citizens: What 20 Million Traffic Stops Can Tell Us about Policing and Race* (2018) and has been published in *Politics, Groups, and Identity* as well as other outlets.

**Betsy Sinclair** is Professor of Political Science at WUSTL, where she specialises in the study of American political behaviour. She is the author of two books, *The Social Citizen* and *A Connected America*.

**Svend-Erik Skaaning** is Professor of Political Science at Aarhus University. His research interests include comparative methodology and the conceptualisation, measurement and explanation of democracy and human rights. His books include *Democracy and Democratization in Comparative Perspective* (with Jørgen Møller, 2013) and *The Rule of Law* (with Jørgen Møller, 2014). Among other things, he is currently completing a co-authored book, *Varieties of Democracy: Measuring Two Centuries of Political Change* and working on another book project on *The Rise of Modern Democracy*.

**Branislav Slantchev** is Professor of Political Science at the University of California, San Diego. He uses formal modeling, statistical analysis, and historical cases to study crisis escalation and coercion, war fighting and termination, stability of authoritarian regimes, and choices in international organizations. His work has been funded by the National Science Foundation, among others. He is currently working on the emergence of the fiscal state, and the domestic politics of threat perception.

**Gabriele Spilker** is Associate Professor of International Politics in the Department of Political Science and Sociology of the University of Salzburg. She holds a PhD from ETH Zurich. Before joining the University of Salzburg, she was a Postdoctoral Researcher at ETH Zurich and a Fritz Thyssen Fellow at the Weatherhead Center of International Affairs at Harvard University. Her main research interests are in the areas of international political economy, international cooperation, globalization and environmental politics. Her work has been published in major peer-reviewed journals, such as *International Organization*, *International Studies Quarterly* and the *Journal of Politics*. She is the author of *Globalization, Political Institutions and the Environment in Developing Countries* (2013).

**Marco Steenbergen** is professor of political methodology at the University of Zurich. His research interests include multilevel analysis, measurement, electoral behavior, political parties, and political psychology. His publications have appeared in the major political science journals and with the major academic presses.

**Zachary C. Steinert-Threlkeld** is an Assistant Professor of Public Policy at the University of California, Los Angeles' Luskin School of Public Affairs. He uses computational methods to study protest dynamics, with a particular interest in how social networks affect individuals' decision to protest. He has used text analysis to study mobilisation during the Arab Spring, information warfare in Ukraine and activists' online strategies, and his work with images measures how violence, social cleavages and free riding affect protest dynamics. His other work includes simulations of protest diffusion and studying how governments attempt to influence



individuals' online behaviour. USAID's Understanding Social Movements programme supported some portions of this research.

**Brandon M. Stewart** is an Assistant Professor of Sociology and Arthur H. Scribner Bicentennial Preceptor at Princeton University, where he is also affiliated with the Politics Department, the Office of Population Research, the Princeton Institute for Computational Science and Engineering and the Center for the Digital Humanities. He develops new quantitative statistical methods for applications across computational social science. He completed his PhD in government at Harvard in 2015, where he had the good fortune of working with the interdisciplinary group at IQSS. He also earned a Master's degree in statistics from Harvard in 2014.

**Jessica S. Sun** is a PhD Candidate in political science at the University of Michigan, a visitor in the Political Science department at the University of Rochester and an incoming Assistant Professor in Political Science at Emory University. Her research focuses on formal theory, civil conflict and autocratic regimes.

**Tilko Swalve** is a Post-Doctoral researcher at the University of Hamburg. He received his PhD from the Graduate School of Economic and Social Sciences at the University of Mannheim in 2019. His research areas are comparative political economy, judicial behaviour and empirical legal studies. He teaches courses on game theory, quantitative methods, the political economy of institutions and comparative judicial politics.

**Joan C. Timoneda** is a Postdoctoral Research Associate at the DevLab@Duke and the Department of Political Science at Duke University. His work focuses on the comparative political economy of regime change, with particular interest in transitions within authoritarianism and democratic backsliding. His methods research interests include text analysis, network analysis using big relational data and statistical models for panel/TSCS data.

**Rocío Titiunik** is Professor of Politics at Princeton University. She specialises in quantitative methodology for the social sciences, with emphasis on quasi-experimental methods for causal inference and political methodology. Her research interests lie at the intersection of political science, political economy and applied statistics, particularly on the development and application of quantitative methods to the study of political institutions. Rocío received her PhD in agricultural and resource economics from UC-Berkeley in 2009. Between 2010 and 2019, she was a faculty member in the Department of Political Science at the University of Michigan.

**Richard Traummüller** Professor of Empirical Democracy at the University of Mannheim. Previously, he has held positions at the Universities of Konstanz, Berne, Mannheim, Essex and Frankfurt. Previously, he held positions at the Universities of Konstanz, Berne, Mannheim and Essex. He has taught courses on data visualisation at these universities and as an instructor for the Essex Summer School in Social Science Data Analysis and the International Program in Survey and Data Science. His work has appeared in journals such as the *British Journal of Political Science*, *Comparative Political Studies* and *Political Analysis*, among others. His book project on data visualisation for the social sciences is under contract with Cambridge University Press.

**Shawn Treier** is a Senior Lecturer at the School of Politics and International Relations at The Australian National University and a visiting scholar at the United States Studies Centre at the University of Sydney. His work has been published in *American Journal of Political Science*, *Journal of Politics*, *Political Analysis*, *Public Opinion Quarterly* and elsewhere. He is also the

co-author, with Jeremy C. Pope, of the forthcoming book *Founding Factions: How Majorities Shifted and Aligned to Shape the U.S. Constitution*. His research concerns the development of Bayesian models of measurement and application to American political institutions, behavior and development, and the measurement of democracy.

**Virginie Van Ingelgom** is a Research Associate Professor F.R.S – FNRS at the Institut de Sciences Politiques Louvain-Europe, University of Louvain and an Associate Researcher at the Centre for European Studies and Comparative Politics, Sciences Po Paris. She is the author of more than 30 articles and chapters, on the issue of legitimacy, both at the national and European levels, on policy feedbacks and on the methodological issues of using qualitative comparative analysis. She is the author of *Integrating Indifference* (2014) and co-author of *Citizens' Reactions to European Integration Compared. Overlooking Europe* (2013). Her current teaching commitments include courses at the UCLouvain Master in Political Sciences (political sociology) and at the ECPR Summer and Winter School in Methods and Technics (focus groups). In 2017, she was awarded with an ERC Starting Grant for her project Qualidem – Eroding Democracies.

**Georg Vanberg** (PhD, University of Rochester, 1999) is Professor of Political Science and Law at Duke University. His research focuses on political institutions, including courts, legislatures and coalition governance.

**Gonzalo Vazquez-Bare** is Assistant Professor of Economics at the University of California, Santa Barbara. His research focuses on designing econometric methods for causal inference and policy evaluation. Gonzalo earned a PhD in economics and an MA in statistics from the University of Michigan. Prior to his PhD, he worked as a consultant for the Inter-American Development Bank and the World Bank in Washington, DC.

**Mateo Vásquez-Cortés** is an Assistant Professor of Political Science at Instituto Tecnológico Autónomo de México (ITAM). His research focuses on the political economy of conflict and development, with an emphasis on topics related to violence and crime. In particular, his work analyses the causes and consequences of violence, the determinants of the successful reintegration of ex-combatants after conflict and how emotions can affect violent and non-violent political participation. In his work, he uses both formal theory and empirics, including both experimental and quasi-experimental designs. He holds a PhD in politics from New York University.

**Tiago Ventura** is a Ph.D. Candidate in Government and Politics at the University of Maryland, College Park, and a researcher at the Interdisciplinary Lab for Computational Social Science (iLCSS). His research centers on comparative politics and computational social science, with particular attention to political economy and crime in Latin America, and political communication. His methodological interests center on the use of computational techniques for text-analysis and natural language processing, network analysis, and big data, with an emphasis on causal inference estimation in observational data. He also holds a Master's and a Ph.D. degree in Political Science from the State University of Rio de Janeiro, Brazil.

**Jennifer N. Victor** is Associate Professor of Political Science at the Schar School of Policy and Government at George Mason University. Her work has been published in the *British Journal of Political Science*, *American Politics Research*, *Interest Groups and Advocacy* and elsewhere. She is co-editor of the *Oxford Handbook of Political Networks* (2017) and co-author of *Bridging the Information Gap: Legislative Member Organizations in the United States and the European Union* (2013). She holds a PhD (2003) and an MA (1999) in political science from

Washington University in St. Louis. She is a past President of the National Capital Area Political Science Association and a past chair of the APSA organised section on Political Networks. She is a co-founding contributor to the political science blog *Mischiefs of Faction* on Vox.com, and she has blogged for *The Conversation*, *Medium*, *OUP Blog* and the *LSE US Politics* blog.

**Fabio Wasserfallen** is Professor of Comparative and European Politics at Zeppelin University in Friedrichshafen. Previously, he was Associate Professor of Political Economy at the University of Salzburg, Fung Global Fellow at Princeton University, Research Fellow at Harvard University and Guest Professor at the University of Zurich, where he also earned his PhD. His research interests include European integration, public opinion, policy diffusion, federalism and direct democracy. The findings of his research have been published, among other journals, in the *American Political Science Review*, *American Journal of Political Science*, *British Journal of Political Science* and the *European Journal of Political Research*.

**Clayton Webb** is an Assistant Professor of Political Science at the University of Kansas. His research focuses on time series methodology and the domestic political dynamics of foreign policy. His work has appeared in the *American Journal of Political Science*, *Political Analysis*, *Political Research Quarterly* and other journals.

**Anna M. Wilke** is a PhD candidate in Columbia University's Department of Political Science. Her research interests concern the political economy of development, with a focus on crime and violence in sub-Saharan Africa. Recently, she has worked on violence against women, community policing and mob justice in Uganda and South Africa. Her work draws on quasi- and field experimental methods as well as game theory. Prior to her doctoral studies at Columbia, she graduated with a Master's degree (MRes) in political economy from the University of Essex.

**Roi Zur** is a Lecturer (Assistant Professor) of Comparative Politics at the Department of Government at the University of Essex. He studies voting behavior, electoral strategies of political parties and political institutions in Western democracies. His work has been published in the *British Journal of Political Science*, *European Journal of Political Research* and *German Politics*.



# An Introduction

Luigi Curini and Robert J. Franzese

Fifty years ago, Paul Lazarsfeld and Morris Rosenberg advanced a definition of ‘methodology’ in their edited volume *The Language of Social Research*, which remains as enlightening and necessary today as then: ‘methodology consists of a reflection on empirical research, on the appropriateness of the procedures and the assumptions used in relation to the intellectual intent of the researcher’ (Lazarsfeld and Rosenberg, 1955: 4). Even as discussions of methodology elucidate the criteria for assessing the quality of research and enumerate the standards that any ‘good scientific’ study must meet, yet still these criteria and standards cannot be reduced to some list of ‘instructions’ to follow or coded into some sophisticated script to run. Notice in this regard that this definition of methods emphasizes the appropriateness of the techniques *to the intellectual intent of the researcher*. That is, the optimality of methodologies chosen is specific to the research questions and aims to which they are applied. And in a broader perspective, critical

reflection on research design and methods is crucial to the balanced and conscious development of any discipline, including political science and international relations.

In the postwar period, the study of political science and international relations began to turn from configurative description and normative evaluation toward a positive social science, a scientific discipline particularly interested in questions surrounding the establishment, maintenance, and security of democracy and peace, for the obvious historical reasons. Over the course of the ‘behavioral revolution’ of the next decade or two, this new scientific discipline’s interest in methods and methodology emerged and grew, with increasing momentum through to today. According to a *Jstor* search query, for instance, the number of articles published in the political science journals mentioning ‘method\*’ explicitly in their abstracts grew from a yearly average of 41 between 1960 and 1990 to 133 in the following decade, and to 241 from 2001 through 2017 (*Jstor* search

date: 22 November 2017)). A very similar trend, albeit of lesser magnitude, is found in the international relations journals: from an average of four abstracts per year explicitly mentioning ‘method\*’ between 1960 and 1990 to nine a decade later and an average of 15 per year since 2001.

These strongly upward trends are hardly surprising given the extraordinary vivacity of methodological development and debate in recent decades, in both the quantitative and qualitative empirical-research traditions and methodologies: from the development of parametric and non-parametric techniques for more accurate and effective causal estimation and more robust and credible causal inference, the growing interest in the Bayesian approach in both quantitative and qualitative research, the developments in experimental and quasi-experimental design, and the opportunities and challenges posed by big data to the development of ‘mixed’ designs and configurational analysis – to name just a few. These trends seem surely destined only to strengthen further in the future. The present *Handbook* aims to encompass this wealth of developments, offering brief introductions to and expositions of theoretical and empirical research methods from across political science and international relations through a series of chapters authored by leading scholars of these methods. The *Handbook* sections are organized sequentially along the lines of applied research in the discipline: from formulating good research questions and designing a good research project, various modes of theoretical argumentation, conceptualization and measurement of the variables, the moving parts, of the research contribution, and collection, representation, and preliminary exploration of these empirical data to empirical methods of quantitative and qualitative analysis, including the concluding movements: the drawing of theoretical inferences and the interpretation of substantive estimates.

The chapters are each designed to reflect the current state of thinking on their topic, but

equally to provide an accessible contribution that informs a diverse audience – graduate (and ambitious undergraduate!) students, young and established professors and academics, researchers in the private, non-profit, and public sectors, and PhDs working in other modes of teaching, research, or applied work – searching for an informed take on a topic and to understand how a method works and/or is best applied. The aim is to provide a comprehensive resource by which this diverse audience of scholars and applied researchers in political science and international relations can learn about empirical methods and how to use them most effectively.

Given its targeting of a broad audience and its structuring through the steps of a research project, agenda, and career, the *Handbook* begins, in the Preface, with a welcome address given at the common starting point that connects us all: graduate training in political science and international relations. The chapters that follow are organized along the arc of a research project, from formulating questions, to theory building, operationalization and measurement, through to quantitative and qualitative research design and empirical analysis and the drawing of conclusions. Notwithstanding its considered, sequential organization, however, this *Handbook* is not intended be read linearly, chapter-by-chapter (although one is certainly more than welcome to try that if so inclined!). Rather, we suggest the reader follow some different ‘intentional routes’ through select chapters according to her needs. For example, a reader interested in network analysis or developing a network-analytic research project could start from Chapter 45, which introduces network analysis in a relatively non-technical way, then move to Chapter 46, a more technical offering that moves the reader to advanced issues, and then to Chapter 30, which applies network analysis to social media data, with several operative examples. Similarly, a reader interested in machine-learning techniques applied to text analysis could begin with Chapters 55

and 56, two chapters that introduce a variety of machine-learning algorithms, then move to Chapter 26 for an introduction to text analytics, and then to Chapters 27, 28 and 29 for their applications to texts from a variety of sources (including social media data). Finally, a reader keen to improve her knowledge about qualitative methodologies could begin with Chapter 6, which presents a discussion on how to apply qualitative methods to theory building, before moving to the last section, several chapters devoted to presenting different approaches to qualitative methods and studies.

We will conclude this short introduction with some comments perhaps especially important for a *Handbook* of research methodology. First: research scientists do not generally deploy a methodology, however sound and sophisticated, for its own sake. Rather, the ambition is that one's scientific work will be creditable because of the soundness of its conceptualization and design and the sincerity, care, and appropriateness of its conduct and methods, and that it will be *important*, meaning that it will be 'worth to be known' (*wissenswert*: Weber, 1994). Second: also meriting heavy emphasis at the beginning of a research-methodology *Handbook* is the Weberian advice that, even though perfection in this regard is impossible to obtain, scientists must be as *objective* as possible and try to acknowledge explicitly any unavoidable limitations therein as much as possible. Third: we would notice and stress also that these aims, these prescripts, and these advanced methods of scientific work have nothing to do with the work's 'practical or policy relevance and importance', as supposedly somehow opposed to its 'academic or intellectual weight', as is sometimes alleged. If we study something important (asking and offering answers to *important* questions, meaning something people care about – 'worth their knowing', as Weber puts it) and have something positive (not normative), scientific and rigorous to say about it – which means theory developed and propounded as

tightly as possible, and empirical analysis conducted as soundly, openly, and honestly as possible – then the work will be relevant inside and also, *perforce*, outside academia (Collingridge and Reeve, 1986). And it will in this way be policy-relevant and practically important *without (necessarily)* bearing normative prescriptions.

This is because positive (not normative) political science and international relations are about how the socio-politico-economic world works – the physics and mechanics of it – not how we want (or should want) it to work (the metaphysics). It is for politicians and 'the people' to decide those desires. What we can do as (political science and international relations) researchers is work to describe the machine that produces this socio-politico-economic functioning. As scientists, we do this by posing interesting, previously unanswered questions about some aspects of this world, constructing theoretical arguments and models that offer useful understandings of how those aspects may work, and conducting careful empirical analyses to test whether and to estimate how these mechanisms actually operate empirically; that is, we provide useful theoretical and empirical simplifications. Research methodology, from this perspective, is the approaches and techniques that make political science and international relations research capable of producing these theoretically, empirically, and therefore practically, useful generalizations.

## SECTIONS AND SECTION EDITORS

We extend, in closing, our enormous gratitude and appreciation to the section editors, without whose close and wonderfully constructive readings of their section's contributions this tremendous collection of masterworks would not have been possible. In particular, we want to thank Branislav Slantchev for the section 'Formulating Good

Research Questions and Designing Good Research Projects', Kris Ramsay and Adam Meirowitz for the section 'Methods of Theoretical Argumentation', Lucas Leeman and Robert Klemmensen for the sections 'Conceptualization and Measurement' and 'Large-Scale Data Collection and Representation Methods', Vera Troeger and Richard Nielsen for the section 'Quantitative-Empirical Methods', and, finally, Chiara Ruffa for the section 'Qualitative and "Mixed" Methods'.

## REFERENCES

- Collingridge, Davide and Reeve, Colin. *Science speaks to power. The role of experts in policy making*, London, Francis Pinter, 1986.
- Lazarsfeld, Paul F. and Rosenberg, Morris (eds.). *The Language of Social Research: A reader in the methodology of social research*, Glencoe, IL, The Free Press, 1955.
- Weber, Max. Max Weber-studienausgabe: *Wissenschaft Als Beruf 1917/19. Politik Als Beruf 1919*. Tübingen, Mohr Siebek Ek, 1994.

# Foreword

We begin our *Handbook's* tour of the scientific research project/agenda with the starting point that connects us all, which is graduate training in political science and international relations. Whether we are entering graduate school, working through it, just starting on our career in political science or international relations research and teaching, or continuing to advance it after many years, Gary King's encouraging welcome address to entering graduate students about how to succeed in graduate school on the road to becoming a research scientist in political science and international relations provides excellent advice about preparing for and doing research (and is a very pleasant read as well).



*This page intentionally left blank*



# So You're A Grad Student Now? Maybe You Should Do This

Gary King<sup>1</sup>

Congratulations! You've made it to graduate school. This means you're in a select group, about to embark on a great adventure to learn about the world and teach us all some new things. This also means you obviously know how to follow rules. So I have five for you – not counting the obvious one that to learn new things you'll need to break some rules. After all, to be a successful academic, you'll need to cut a new path – and so if you do exactly what your advisors and I did, you won't get anywhere near as far, since we already did it. So here are some rules, but break some of them, perhaps including this one.

First, you're probably wondering how in the world you can write a dissertation – something like 250 pages – from scratch. Well, remember this: a dissertation is both *easy* and *irrelevant*. It's *easy* because a dissertation is the equivalent of maybe three to five papers, and you must have written that number every year for at least the last ten. So write a paper, then another, and then another; at worst,

you'll wind up with a series of articles as a dissertation, which often works out great; at best, you'll initiate a whole research program with a sequence of papers that sums up to more than the parts, or possibly a great book (which is something like four articles' worth of effort and maybe six in terms of credit). After all, you probably haven't the slightest idea how to write a book; so start writing articles and see where you wind up. Maybe a book will pop out naturally, but there's no reason to force it. (The same applies after grad school: all eight of the books I've written started out as articles that I couldn't figure out how to fit into 40 or so pages.)

A dissertation is also *irrelevant*, because this assignment is not about writing 250 pages; it's about reorienting your life, making the transition from a student taking classes – and doing what you're told – to being an independent, active professional, making regular contributions to the collective enterprise, competing and cooperating with your colleagues in pursuit of common

goals. To do all this, you need to arrange your whole life, or at least the professional portion of it, around this goal. You should not try to change into a dissertation writer (or a dissertator!) but into a professional academic, looking for opportunities to make contributions to the scholarly community that make a difference in the world. If you do that successfully, you'll get a dissertation for free along the way.

Second, in graduate school, *never shoot for the immediate goal; aim for the one after that*. Let's start with the dissertation prospectus, the text of which will not matter five minutes after it's approved. No one will ever ask you whether you did what you promised in your prospectus, and even you are unlikely to read it again. One reason for this is that writing a prospectus is itself almost logically impossible: you are supposed to convince three experienced faculty that you will discover something that will surprise them and they do not now believe to be true. And you're supposed to do that how? By speculating about what you will find, and how important it will be, if you ran some hypothetical analysis on an imaginary dataset you do not even have access to yet and may not even exist.

So don't write a traditional prospectus; instead, write an article or chapter and bring it to your prospectus committee (stapled to a one-page outline of your imagined dissertation to meet the formal goal). Then you will have made some progress on the goal after the next one and, at a minimum, will switch the conversation from armchair speculation to a productive discussion and useful advice about your work.

The same idea applies to dissertations, which are also useless five minutes after approval. Instead, try to write papers that will work as publishable articles or a manuscript that will make a book publisher happy. Just skip the step of writing a dissertation (and certainly do not use the word 'dissertation' in your dissertation; just refer to your 'work' or 'manuscript'). Similarly, don't waste your time attending dissertation defenses (except

for your friends' – and especially your own!), but go to all the job talks you can and imagine yourself standing at the front of the room, thinking of how you might respond to each question.

Third, everything you write from now on must answer this one question: *whose mind will you change about what?* This means you are not choosing a 'dissertation topic'. You've already done that by your choice of subfield and maybe even your graduate program. You should instead lead with a finding, discovery, result, or argument (and should at least be able to begin with 'In this work, I demonstrate that...'). Then rigorously organize your work to answer this key question. Remove *every* point, section, sentence, or paragraph that does not directly answer this question or address your argument. (Keep deleted portions in a folder for other projects to avoid separation anxiety, but get them out of this work.) The point of your dissertation is not (or not only!) to show how smart you are; it's to prove your point, make your argument, or solve a problem. Everything else that gets in the way of your contribution goes.

Here's a measure of whether you've succeeded: your argument (and its structure) should be crystal clear from your table of contents, without reading the text. Keep the table of contents as a separate file and keep editing it as you write. The advice you got in eighth grade about writing the outline before the text is a nice theory but doesn't work because you learn about your argument by writing it out. (You know how authors of fiction explain that they wanted to end the story in one way but the characters caused them to end it in a different way? Pretty much the same thing applies to nonfiction. The story takes on a life of its own. It is one reason we write down ideas.) Although you probably can't satisfy your eighth-grade teacher now, keep iterating between the text and table of contents. When you're done, the table of contents should be so clear it can tell the story on its own, and the text will then be unencumbered by

scaffolding or the vestige of old monsters trying to distract readers from your argument.

That monumental throat-clearing exercise called a 'literature-review section' is a great example. Fuhgeddaboutit. Those people have their own books and articles, where they make their own points; they don't get to be in your dissertation unless they help you make your point. I know you're accustomed to writing literature reviews, but that's for your teacher, testing your knowledge for class. You've now passed that test and don't need to keep taking it: do not include a literature-review section. At the same time, be professional: leave out gratuitous or fawning citations to your professors or anyone else; tell them how great they are in person if you like, but don't let them get in the way of making your contribution clear.

A final way to focus on your point is to not insist that your dissertation have 'symmetric' evidence: present all available observable implications of your theory, even if one is an ethnography of a restaurant and another is a cross-national quantitative study. Any good evidence or argument can help you evaluate your claim and demonstrate whose mind you're going to change about what. Avoid selection bias, but do not distract your audience with forced symmetry that sends you off collecting the same data from every state merely for aesthetic reasons that do not help support the evidence.

Fourth, you obviously need to get the social science right, but present your results so others not only understand what you are saying but have *no choice* but to read your work. Until graduate school, at least one person was always paid to read what you wrote. After graduate school, if you don't write so that others find they must read your work, it could be the case that no one ever reads it. You could even write a great paper, get it published in a top journal, and the only person who ever reads it is you.

The job of an academic, and the mission of the university, is the creation, preservation, and dissemination of knowledge. If no one

reads your work, you will change no one's mind, make no difference, and get no credit. Modern political reinterpretations notwithstanding, Christopher Columbus would have gotten tenure, but Lief Erikson would have had to go back on the market. Thus, your title needs to grab readers by the lapels and yank them into the page so they feel they must read your abstract; your abstract needs to interest readers enough so that they feel their own work is at risk, or they are so interested that they immediately read your introduction; and so on. Doing 'good work' is no longer good enough.

Imagine two dissertations, identical in all respects except that the title, abstract, and introduction in one is rewritten so that it resonates with your audience. The author of that one will get a great job and have a great future. The other not so much. You might as well be the one to learn this. Try out your idea on your grad-school colleagues, your friends, your parents, and non-academics. If they don't get it, it isn't because they weren't trained. Figure out how to convey what you're doing so anyone can understand it.

Graduate school is a transition from being a private citizen taking classes to a public figure writing for a big amorphous, ill defined audience that it is your responsibility to define, find, and engage. This is not easy, and it accounts for most of the frustration scholars have with the peer-review process. It will take more time than you think (even after adjusting for this sentence). It will require rewriting, recasting your argument, reconceptualizing your theory, recollecting your evidence, remeasuring your variables, or reanalyzing your data. You'll have to revise more than you want and you thought possible. But try not to get discouraged; they call it *research*, not *search*, for a reason! Be your usual relentless self and get it done.

In my experience, almost all dissertations are written in about four months, even though it takes many people years and a spark of motivation like a job offer or graduation deadline to get started. In the end, it

is simple: educate your advisor that you're ready to graduate (yes, that's your responsibility!) and then she who shows up with the pile of paper gets the degree.

Finally, the process may sometimes seem like drudgery, but remember one last rule: you're allowed to have a life. Go have some fun. And also do not forget that you are tremendously privileged to participate in science and academia and discovery and learning – by far the most exciting thing to 99 percent of the faculty at your university. The thrill of discovery, knowing you're part of something bigger, the adrenalin-producing ah-ha moments, the feeling of learning something that no one in the history of the world

has ever known before but, because of you, many will now know are more exciting than all the skiing and mountain biking you could pack into a lifetime. Don't miss how intoxicating and thrilling it all really is.

### **Note**

- 1 This preface comes from a talk at the Graduate School of Arts and Sciences at Harvard University on a panel called 'The Dissertation: Strategies for Getting from the Beginning to the End of the Process'. My thanks to Cynthia Verba for the inspiration and for arranging this panel, and to my own dissertation committee for help getting me started – Leon Epstein, Art Goldberger, Barbara Hinckley, and Bert Kritzer.

PART I

# Formulating Good Research Questions and Designing Good Research Projects



*This page intentionally left blank*



# Asking Interesting Questions

William Roberts Clark<sup>1</sup>

Good research is driven by impatience with bad answers to interesting questions. But where do interesting questions come from? Since this is the opening chapter of a handbook on research methods, it is imperative to point out at the start that there is no ‘method’ to asking research questions, in the sense of a cookbook you can follow that will lead, inexorably, to scientific ‘discovery’. There *may* be a scientific method for evaluating answers, but there is certainly no scientific method for asking questions or generating answers. And there is certainly room for a lot of creativity in developing interesting and enlightening research designs, and serious shortcomings to ‘cookbook’ approaches.<sup>2</sup> Karl Popper (1962, 2003), for example, argued that science begins after a scientist has conjectured an answer to a question. The scientific method, therefore, is more (perhaps only) useful in evaluating answers to questions. Generating questions and answers, in contrast, is as much an art as it is a science.

But that is not to say that the process is random or lacks structure. Thomas Kuhn

(1962: 763) says that episodes of scientific discovery begin with an individual with the ‘skill, wit, or genius to recognize that something has gone wrong in ways that may prove consequential’. But, he hastens to add, ‘anomalies do not emerge from the normal course of scientific research until both instruments and concepts have developed sufficiently to make their emergence likely and to make the anomaly which results recognizable as a violation of expectation’.

In the parlance of social media, scientific discovery begins with a ‘WTF’ moment. Scientific discovery begins when a scholar observes something contrary to expectations and recognizes that this anomalous observation ‘may prove consequential’. Note that the motivating fact may be an observation about the world, but it may also be about what others have said about the world.<sup>3</sup>

But not just any surprise will do. Anyone who has ever parented a young child is familiar with the questions, born out of wonder, such as those that our children asked my partner and me: ‘Why is the sky blue?’ ‘Where does



the sun go (at the end of the day)?’ ‘If my brain controls my body, why do I have to go to the doctor to find out what’s wrong with me when I am sick?’ Answers to all of these questions (assuming they are consistent with what scientists currently believe) are discoveries for the inquirer because they change what *they* know, but they do not lead to *scientific* discoveries unless they change what *we* know. The fields of optics, astronomy and neuroscience have their respective answers to the questions above (although the last question is probably less settled than the other two).

So, questions often begin with surprise, but good research questions begin with well-informed surprise. If you alone are surprised by an observation, the answer to your ‘WTF moment’ is likely to be personally rewarding. If most well-informed observers are surprised by an observation, then an answer is likely to be socially and, therefore, *scientifically* valuable.<sup>4</sup>

But sometimes science proceeds when an individual recognizes that the answers embodied in what ‘we know’ about a subject are not very good. For example, for millennia ‘we’ knew that the answer to the question ‘where does the sun go’ was something like ‘the sun circles a stationary earth, so at a certain point each day it leaves our sight while shining on the other half of the planet only to return the next morning’. Eventually, however, scientists with ‘the skill, wit, or genius’ to recognize the mounting anomalies created by models based on a geocentric view of the universe came to the conclusion that a better answer was needed. At first these better answers came in attempts to modify the heliocentric view with elaborate patches meant to explain away anomalous observations. In addition to skill, wit, and genius, it required a great deal of courage to challenge the existing view in a more fundamental fashion.

So, good questions come from knowing what ‘we’ know. But they also come from thinking deeply about what we know

and being sufficiently unsatisfied with bad answers to take the risk of thinking differently about a problem. As with all the arts, good science seems to come from individuals and groups that engage in a certain kind of practice. I would like to begin this chapter by commenting on what I see as a common structure of many great contributions to political science and international relations. Specifically, I will put forward a list of five questions that, when answered well, are likely to produce work that asks and answers interesting and important questions and gives us a reason to be confident in those answers. In the second half of the chapter I will ruminate on the kind of practice that I expect to lead to good question asking and good answer giving.

## FIVE QUESTIONS

When I was in graduate school, one of my professors, D. Michael Shafer, taught me how to read. He did so by encouraging me to employ a template he created so students could record the key parts of what they read: ‘What is the dependent variable?’ ‘What are the independent variables?’ ‘What is the logic that ties them together?’ etc. I found this enormously helpful in getting through the ridiculous amount of reading required in my graduate classes. When I began teaching I shared this list with my students and over the years I have refined it for various reasons. I have come to believe that this list of questions is useful not just in focusing our reading efforts, but also in our research efforts. If you ask what the author’s answer to each of the following questions is, you will have a good summary of most articles or books in our discipline.<sup>5</sup> If you ask whether the author has a good answer to each of the questions, you will have a good critique of the paper in question. And if you are impatient with any bad answers provided by the author, and develop better ones, you will be on your way

to making your own contribution to the literature. Consequently, I have come to believe that these questions can also serve as an excellent guide when designing a research project. If you have good answers to these five questions (and at least one of these answers is an improvement over existing work), you will have a good paper, dissertation or book. These questions also correspond to the organization of the modal paper in our discipline: 'Introduction', 'Literature Review', 'Theory', 'Research Design' and 'Findings'.

It is important to add that research questions need not be generated by reading. They can just as easily, and perhaps more profoundly, be provoked by our interaction with and observation of the social world. We might observe behavior and ask: 'Why does that happen?' It is good practice to offer one's tentative answer to such a question unencumbered by 'the literature'. But it is imprudent to spend very much time on such activity before evaluating existing answers to your question.

### **Question 1: What Do I Wish To Explain? (The Introduction)**

Following Kuhn's description of scientific revolutions, most good work begins with a puzzling observation. Beginning with observation is important because good readers would like to be convinced that the phenomenon you are explaining actually occurs (though it is frequently fruitful to engage in thought experiments about things that have

not occurred). This step is by no means trivial and considerable methodological sophistication may be necessary to accurately describe the real world events or, better still, patterns of events which you wish to explain.

Samuel Huntington's classic *Political Order in Changing Societies* (1968) seeks to explain the rising political instability he observed around the world. As evidence of this rising instability, on page 4 of this 462-page book, the author presents US Department of Defense data showing that the number of nations around the world experiencing military conflicts of various types rose almost monotonically from 34 in 1958 to 57 in 1965 (Table 1.1). This is a dramatic increase: in less than a decade the number of conflicts nearly doubled! The problem, however, is that, as a result of decolonization, the number of independent countries in the world also grew rapidly during this period. If one takes Huntington's numbers and divides them by the number of independent countries in each year (as a measure of the opportunity for military conflict), the *relative frequency* of military conflict actually declined over this period. Since military conflict was just one proxy for political instability, it is entirely possible that political instability actually increased during the observed period. But if you believe that the relative frequency of conflict is a better indicator of political instability than the raw frequency, you would be justified in wondering if the phenomenon explained in the subsequent 400 or so pages actually occurred.

The first order of business, therefore, in demonstrating that something that may

**Table 1.1 Military conflicts, 1958–65**

	1958	1959	1960	1961	1962	1963	1964	1965
Prolonged, irregular or guerrilla insurgency	28	31	30	31	34	41	43	42
Brief revolts, coups, uprisings	4	4	11	6	9	15	9	10
Overt, militarily conventional wars	2	1	1	6	4	3	4	5
Total	34	36	42	43	47	59	56	57

Source: U.S. Department of Defense.

prove consequential has happened is to demonstrate that *that* thing has happened. This crucial task is often best accomplished with clearly presented, well thought out, descriptive evidence. While this often requires a fair amount of methodological skill, sometimes it simply requires numeracy – which, unfortunately, is often in short supply. Effectively presenting evidence for one's explanandum is perhaps best described in the breach. For example, you can read newspaper headlines on almost a daily basis that purport to capture some important change in the world that is, in fact, not supported by the text of the accompanying article. Would that it were the case that these mistakes were rare in academic work.

One common mistake is to make a claim about inter-temporal change in a variable by citing only current values of that variable. 'Tenure-track jobs are disappearing', reads the title of an article, but the article makes no reference to the number of such jobs that were available in the past. How do we know that change has occurred? A related issue that requires a bit more methodological skill to avoid is to point out a difference between the values of a few recent values of a variable and preceding values and claim that they are evidence of a new trend, without comparing the new observations with a long enough trend of data to determine whether they represent a meaningful deviation from the trend or, as is often the case, just typical variation within the trend.

Another common error is what might be called 'the denominator problem' – the failure to choose a denominator that would transform the data into a variable appropriate for the conceptual comparison relevant to the discussion at hand. We already saw an example of this when Huntington confused a trend in the raw frequency of a variable for a trend in the relative frequency of the data, which I argued would have been more appropriate. But it is also possible that the raw number is what most interests us – in which case we should not be distracted by

an apparently related ratio. To return to the 'disappearing tenure-track jobs' problem we often hear about in the popular press. In the rare instances where inter-temporal data is presented in an attempt to establish this trend, the quantity presented is typically the ratio of tenure-track jobs to the total number of college teaching jobs. This is problematic because it is entirely possible for the *share* of tenure-track jobs to be declining when the *number* of tenure-track jobs is increasing (as has been the case in the United States for decades). And it is probably the latter number that is of interest to most readers (for example, current doctoral students hoping to forecast future demand for people with the credentials they are working hard to obtain).

### **Question 2: Why Does It Need To Be Explained? (The Literature Review)**

Having explained that *this* thing has occurred, it is important for authors to demonstrate that (a) this thing violates expectations in some way (i.e., 'something has gone wrong') and (b) this violation may 'prove consequential'. In other words, in the words of Miles Davis, 'so what?'

Once again, it might be easier to say what one should not do. I once attended a practice job talk where a smart, hard-working and, subsequently, very successful scholar, when pressed to say what he was trying to explain, said that he was trying to explain why a particular variable varies. Being less supportive than I should have been, I asked, 'do you have a theory that leads us to expect this variable to be a constant?' Variables vary. It is even in their name. Observing that variation, therefore, hardly constitutes a surprise. So if variation in a variable does not constitute a violation of expectations, what does?

As a comparative politics scholar, it pains me to say that I have attended many seminar talks over the past few decades, most given

by successful and influential senior scholars, where the work in progress is motivated by an assertion that is some variant of the following ‘puzzle’:

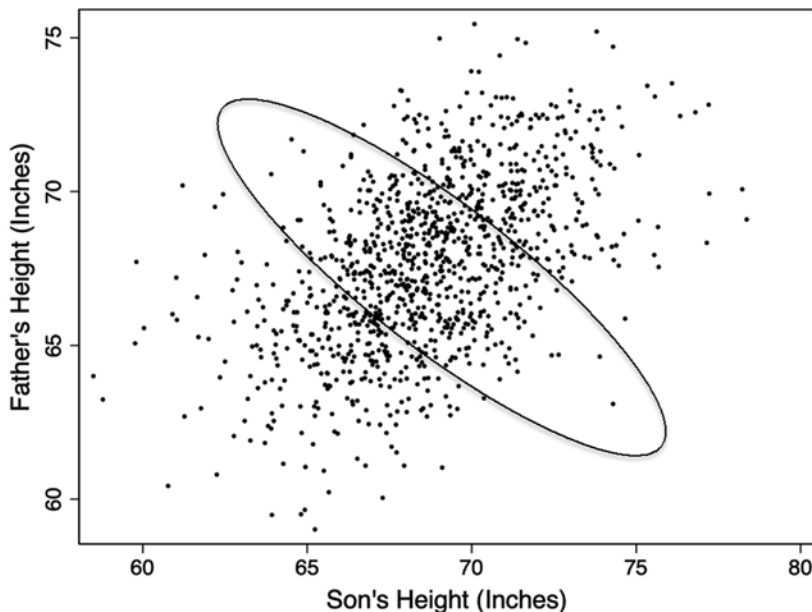
Theory  $Q$  claims that high levels of variable  $X$  should cause  $Y$  to happen, but in country  $i$  at time  $t$ ,  $X$  was very high, and  $Y$  did not occur.

The problem with this ‘puzzle’ is that once the misunderstanding on which it is based is cleared up, it is no longer a puzzle. The misunderstanding is this: with very few exceptions (I cannot think of one), the empirical implications of social scientific theories are best treated as probabilistic (Lieberson, 1991). Whether one traces the reasons to the intrinsically probabilistic nature of all human behavior deriving from human agency, the limitations of our understanding, the fact that most (all?) social phenomena have multiple, context-dependent causes or the possibility of classification error (did  $Y$  occur or did it not? was  $X$  really high or low? and compared to what?), it is best to think of our hypotheses as probabilistic. This means the most that theory  $Q$  can claim is that ‘high levels of variable  $X$  should make  $Y$  more likely to happen’. Consequently, the fact that  $Y$  did not occur in country  $i$  at time  $t$ , despite the fact that  $X$  was very high, is not, at least to my ear, particularly puzzling. Unlikely events are expected to happen occasionally. Consequently, one cannot reasonably call a probabilistic conjecture into question with a single null case. Doing so is like being puzzled about one’s uncle who lived to a ripe old age despite being a heavy smoker. This is not puzzling, because the best scientific evidence is that smoking increases the *likelihood* of cancer, not that it *always* leads to cancer. In contrast, it would be surprising to find an entire sub-sample of the population that appears to be immune to the deleterious effects of smoking, or that, after controlling for income or education (or any other potential confound), smokers are not more prone to cancer than non-smokers. In sum, since our theories typically justify expectations about patterns of

data, it takes observations about patterns of data, not discrete data points, to violate those expectations.

While recognizing a pattern in the data is often necessary for generating surprise, it is by no means sufficient. Going back to the many comparative politics seminars I have attended: be wary of the scholar who selects a small sample of observations and demonstrates that a widely corroborated empirical regularity, such as the incumbency advantage, the democratic peace, Gamson’s Law, Duverger’s Law or the resource curse, ‘doesn’t hold’ in that subsample. Why? Because social behavior is probabilistic, so even highly predictive empirical models yield predictions with non-zero errors. As a result, one can always find a sub-sample of data where the broader pattern does not hold. Take any ‘football shaped’ scatter plot, such as the famous (Freedman, Pesani, and Purves, 2007) scatter plot shown in Figure 1.1.<sup>6</sup> One can select out a sub-sample of cases, such as those in the ellipse, to suggest that the regression line is flat or even negative even though there is clearly a positive relationship in the sample on the whole.

Recall that I said to ‘be wary’ of a scholar who motivates their study with a sub-sample of cases that appear to run contrary to a well-corroborated set of expectations. But I would not encourage you to dismiss such a scholar. It is, for example, entirely appropriate to show that there are boundary conditions on even the most well-corroborated empirical regularities. But the mere existence of such a sub-sample does not constitute a puzzle until one can convince the reader that the sub-sample constitutes a comprehensible category and is not just the result of felicitous (from the standpoint of the author seeking something to write about) case selection. Further, if one does take as their project the task of explaining why a well-corroborated regularity does not apply to a particular sub-sample, it is incumbent upon them to develop an explanation for why the sub-sample is different that yields new predictions other



**Figure 1.1 Relationship between the height of fathers and sons**

Source: Freedman, Pisani and Purves (2007) added random noise to data from Pearson and Lee (1903) who only had data to nearest inch: <http://myweb.uiowa.edu/pbreheny/data/pearson.html>

than the fact that the sub-sample is different. Otherwise, they are engaged in both post-hoc and ad-hoc reasoning.

Yet another problem can arise when one generates their research project by gazing at a scatter plot. Many will look at a figure such as Figure 1.1 after estimating a regression line and be disturbed that so many observations fall far from the regression line. It is okay to want the model to fit the data well, but given the probabilistic, multi-causal nature of our hypotheses, it is not puzzling that some observations fall far from the regression line. My father was six feet tall, while I, ahem, am not. That is not surprising because other factors enter into height at adulthood other than my genetic inheritance from my father – diet and contributions from my mother’s genetic make-up come to mind. Being puzzled in this way is a slightly more sophisticated version of the ‘if  $X$  is high in country  $i$  at time  $t$ , why do we not observe  $Y$ ’ problem. Both methods are frequently used to

justify the claim that ‘existing explanations are incomplete’. The problem is that any explanation the author comes up with is likely to be susceptible to the same criticism.

I want to be clear: there is nothing wrong with being unsatisfied with explanations that do not fit the data well. However, if the only result of pointing out observations that fall off the regression line is a new model that marginally increases measures of goodness of fit, do not be surprised if readers fail to see this as ‘consequential’. *Ceteris paribus*, papers that are motivated by the identification of unclear, misleading or incorrect understandings in the existing literature are more consequential than those that point to merely ‘incomplete’ understandings because the former causes us to revise (that is, to ‘look at again’) rather than merely supplement our current understanding.

So far, we have been seeking to identify violations of expectations that are consequential

for our understanding of the world, but one might also place a priority on consequences that are more practical. One way of asking the ‘so what’ question is to ask, ‘if you were successful in explaining your anomalous observation, how would the world be different?’ Unless one is entirely naïve, this is a very tough question to answer. But since most of us became political scientists and international relations scholars because we wanted to make the world a better place, it is still worthwhile. One reason to think about the ‘normative’ implications of the questions we ask is that an even passing familiarity with the literature in political science and international relations is enough to unearth a seemingly endless supply of unclear, misleading or incorrect understandings. In light of this, it is not unreasonable to try to tackle first those that are tied to issues about which we care deeply.

Nobel laureate Robert Lucas, in his Marshall lectures at the University of Cambridge, said: ‘Once you start thinking about economic growth, it is hard to think about anything else.’ (Ray, 1998) I suspect that is because it is not hard to see the real world, stick to your ribs, consequences of economic growth. Likewise, immigration, environmental regulation, political violence, economic inequality, government corruption, racial and ethnic discrimination, financial instability, authoritarianism, gender bias, illiteracy, failing schools or a host of other policy issues are of interest because of their impact on matters of justice and human well-being. Explaining observations that violate our expectations can be quite consequential when doing so sheds light on these and other social problems.

Marx’s last and most famous thesis on Feuerbach is that ‘the philosophers have only *interpreted* the world, in various ways. The point, however, is to *change* it’, and it is interesting that this is etched on his tomb despite having never been published while he was alive.<sup>7</sup> It captures the frustration of many scholars who would like to ‘make a difference’. It certainly captured my romantic

heart when I first read it as a young man (not much younger than Marx was when he wrote it) at the start of graduate school. But I was not in graduate school long before I realized the complexity of ‘interpreting’ the world and the dangers that could result if one sought to change the world without having interpreted it correctly. Understanding the world is a prerequisite for changing it in a responsible manner.

While it is desirable – perhaps even noble – bridging the gap between studying the world the way it is and using this information to improve social conditions is difficult – particularly when people, and, therefore, politics are involved. One problem is that if social ills have political roots, even accurate explanations of their causes are likely to be insufficient for mitigating them. One reason for this is the fact that the hallmark of politics is conflicting values. Explaining to prisoners confronted with plea deals that reward them for incriminating each other that they collectively benefit by keeping mum will not solve the prisoner’s dilemma because they will still have individual incentives to rat on their co-conspirators.<sup>8</sup>

So, while understanding the world may be a necessary condition for (responsibly) changing it, it is not likely to be sufficient. And, conversely, changing the world can make it a lot harder to understand. One of the things that makes social science difficult is that the entities we study can read what we write and change their behavior in ways that make our models less predictively accurate.<sup>9</sup>

Something like this may have been at work in the writings of Marx. The phrase ‘workers of all lands, unite!’ also appears on Marx’s tomb. In contrast to his theses on Feuerbach, this phrase was published during his lifetime. Marx and Engels closed one of the most influential political pamphlets ever written with it, three years after bemoaning the irrelevance of prior philosophers.<sup>10</sup> In an 1890 appendix to *The Communist Manifesto*, Engels admits that few heeded the call in 1848 but suggests many eventually did so over time, including those who

were organizing in support of the eight-hour workday in 1890. It is not unreasonable to suggest that Marx's analysis of an internal logic to capitalism (that the inexorable immiseration of the proletariat would lead to revolution) helped fuel the formation of labor unions and the creation of social programs that improved the material conditions of workers. But in doing so, this made them less revolutionary – thereby reducing the probability of the revolution he predicted.

Another example of how it is hard to have both influence in the real world and predictive accuracy comes from the recent literature on 'the happiness curve' – the robust empirical regularity that reported life satisfaction tends to decline when people are in their forties and rise consistently starting in their early fifties (Rauch, 2018). One explanation for this empirical regularity is that because human psychology is biased towards overly optimistic forecasts, young people overestimate how much their lives will improve in their thirties and forties. This results in disappointment during their middle years even if individuals' lives have improved considerably, but not by as much as they had expected. This disappointment also leads people to update their expectations and make grim forecasts for the future. Consequently, when life in their fifties, sixties and beyond turns out to be not as bad as expected, they report high levels of life satisfaction. If this process is truly at work, people who read this literature might be inclined to make more realistic predictions about future life satisfaction. If they did so in large numbers, the 'happiness curve' could disappear.

Notice that to the extent that Marx changed history, it may have been in ways that frustrated both his predictive accuracy and his social desires (for revolution), but if happiness researchers turn out to have the same degree of impact on society they might be perfectly willing to trade predictive accuracy for tangible improvements in people's life satisfaction.

In sum, we would like to answer questions that, when answered, would prove

consequential. These consequences can be either for the way we think about the world, or for the way people behave. While, all else equal, we would like our research to lead to improvements in human well-being, the strategic nature of politics means that even when we provide good answers to questions that are important to us, it may not lead directly to improvements in social outcomes. That is not to suggest that we should stop trying.

### **Question 3: What Is the Explanation? (Theory)**

A good explanation will take an observation that is sufficiently surprising to justify your study, and turn it into something that, in retrospect, should have been expected all along. In what remains one of the few books I know of that attempts to teach people how to explain things, the authors of *An Introduction to Models in the Social Sciences* (Lave and March, 1975) describe explanation as a process in which one imagines a prior world such that, if it existed, the surprising fact(s) would have been expected. Technically, any set of statements that logically imply the occurrence of the anomalous observation constitutes an explanation. But good explanations have additional attributes, and we would like to produce the *best* explanation. A satisfying explanation will give the reader an understanding of the process or mechanism that is likely to produce the previously anomalous observation. Readers want to know how surprising events came about, and explanations should tell them. Good explanations are efficient – the ratio of things they explain (implications) to things they require you to believe (assumptions) is high.

There is an optimal degree of novelty to an explanation. An explanation should be interesting, yet sound. By 'interesting' I mean that an explanation should cause us to see the world in a new way. By 'sound' I mean an explanation should fit in with other things we know about the world. An explanation that

causes us to see everything in a new way is likely to be wrong. An explanation that does not require us to change our mind at all is probably just a corollary of things we already knew (and, by extension, our motivating puzzle must not have been much of a puzzle).

Finally, explanations must be logically consistent. I have had empirically minded political scientists and international relations scholars tell me that formal theory is not important because they are sophisticated enough to live with theories that contain contradictions. This is nonsense. It can be shown with elementary logic that anything follows a contradiction. Consequently, if your theory contains a contradiction, anything can be said to follow from it. As a result, a contradictory theory rules nothing out and, therefore, no amount of empirical information will be sufficient to falsify it. Since potential falsification is the hallmark of science, a theory that contains a contradiction is not a scientific theory.<sup>11</sup>

One way to increase the likelihood that your explanation is logically consistent is to try to capture it with a formal model. Formal models allow us to demonstrate that our explanation's conclusions follow from its assumptions – most importantly, that our previously puzzling observation is not surprising in light of the world that our explanation posits. Also, by making the assumptions of our explanation explicit, we are more likely to notice if they contradict each other.

While these benefits of formalization are undeniable, it does not follow that every explanation should be formalized. I typically encourage my students to first articulate their explanations as a story that reveals a process that produces the previously unexpected observation. Formalization is only necessary when one hears such a story and asks ‘why would people do that?’ or, equivalently, ‘that doesn’t sound like an equilibrium’, or ‘isn’t there a tension between this part of the story and that part of the story?’ When one is confronted with such questions, a good formal model can often

provide answers. Thus, I tell my students to learn how to write down formal models not because they will always need one, but because, like fire insurance policies, they are always at risk of needing one.

Another reason to begin with an informal statement of one’s theory is to avoid the trap of thinking that a game theoretic model will generate a theory for us. Formal models help us interrogate certain aspects of our theory; they do not produce the theory for us. We must start with some theoretical intuition about what explains the phenomenon in question before we can begin to model the process.

#### ***Question 4: If the Explanation Is True, What Else Should We Observe? (Research Design)***

If you offer a view of a theoretical world that has the previously puzzling observation as one of its implications, you have offered an explanation. And while there are various ways to evaluate that explanation, to be scientific your answer to your original question must provide an answer to the following question: ‘if your explanation is correct, what else ought to be true?’ Good scientific explanations provide lots of answers to this question. If your explanation only implies the facts that you set out to explain, then there is no way to empirically evaluate your answer. You cannot use the fact that democracies seldom fight each other, or the fact that there is a lot of corruption in presidential democracies, to evaluate your explanation of these things, because it was those facts that led you to develop your explanation in the first place.

This part of the research process is a stumbling block for many researchers when they are attracted to a subject rather than a question. I once had a student who visited Brazil, was shocked by the level of corruption in the government there and developed an explanation that pointed to aspects of the large



district magnitude proportional representation electoral system as a cause. The student was surprised when I said I thought the argument had merits, but that returning to Brazil to collect data was not a promising avenue for evaluating the argument: we already knew that Brazil fit the argument! Perhaps data on corruption levels in countries with different electoral laws (such as the United States) would be more useful, I suggested. The student, however, responded that he did not want to study corruption in other countries – after all, he was interested in Brazil!

A similar problem is found in a very famous book by Theda Skocpol, *States and Social Revolutions* (1979). In it, the author wishes to explain the occurrence of social revolutions, and argues that her subject dictates her empirical strategy. Given her definition, there are only five historical cases of social revolution. She argued that as a consequence of this fact, structured focused comparison (specifically, Mill's Method of Agreement) was the only possible method for evaluating her explanation. That is not true.

The chief problem here is that if an explanation for a set of rare events only has implications about those rare events, the author does not have a data problem, they have a theory problem. If an explanation for global warming only predicts the general rise in the temperature that motivated the explanation, then it is not a very useful explanation. Cosmologists have offered explanations for the creation of the universe, but they do not choose their methodology for evaluating their explanations based on the fact that the object of their study only happened once. Instead, they ask: 'if my explanation for this unique event is correct, what else ought to be true?' They then think about how best to carefully observe the implications of their argument.

The goal of empirical research, therefore, should be to examine as many implications of one's explanation as possible. Because many, many scholars restrict their attention to the empirical puzzle that motivated their

study to begin with, many important papers can be written by simply asking of existing explanations, 'if this argument is true, what else ought we observe?'

One reason why scholars often restrict their attention to the data that generated the question is that it can often take considerable creativity to think about the implications of an explanation. There is no cookbook-like approach that can be applied that will automatically reveal to the scholar that seemingly unrelated events might be instantiations of a single social process. But one practice that Lave and March recommend is to try to see your answer to a particular question as related to a more general process.

For example, in her critical review of Skocpol's book, Barbara Geddes (2003) suggests that one element of Skocpol's explanation of rare social revolutions had implications for the occurrence of peasant revolts. Geddes suggests that a statistical model examining the conditions under which peasant revolts do and do not occur would, therefore, be useful in evaluating the empirical relevance of Skocpol's explanation of social revolutions.

Notice that when we ask 'what else ought to be true', we separate the question 'what is the author's explanandum?' from 'what is the author's "dependent variable?"' The explanandum is a statement of what the author develops a theory to explain. The 'dependent variable' is the endogenous variable in a model testing one or more of the implications of the author's theory. There are times when these might be the same, but there is no reason to assume they will be. In fact, when they are, we should wonder if the author is engaged in post-hoc reasoning – 'have they observed the dependent variable and its covariates and constructed a causal story after the fact?' Doing so would constitute a 'test' of the theory only to the extent that the lion's share of the observations could be thought to have been appreciably different from those that were observed before the theory's formulation. Conversely, a theory that produces a lot of

novel implications helps assuage the reader's suspicion that the author is merely engaged in a curve-fitting exercise.

In sum, it is typically more helpful to think of empirical work as testing the implications of a theory, rather than testing the theory directly. One reason why this is true is that testing the theory directly can easily descend into more or less complicated versions of curve-fitting and post-hoc reasoning. Instead, spend time thinking about the implications of your explanation for observations other than those that motivated your question in the first place. The more varied those implications, the better, because it is only those observations that are made after the construction of your theory that run the risk of being false and therefore actually constitute an empirical check on your explanation. And remember: if your theory only has implications for a set of events too small to use standard inferential tools to evaluate, you do not have a data problem – you have a theory problem.

### ***Question 5: Do We Observe the Implications of Our Explanation? (Findings)***

Determining if evidence is consistent with one's theoretical expectations is the primary focus of research methodology, and is therefore the central focus of the remainder of this Volume. Here I will merely stress the following: many, many studies present, often in dizzying detail, reams of information that is either irrelevant to or inconsistent with theoretical expectations. Typically, however, it is presented in a manner that suggests that this information confirms the author's expectations. Distinguishing when this is the case is a large part of what is meant by learning to read critically.

As I said, all of the collective wisdom of research methodologists is relevant for becoming a critical reader and producer of knowledge, but I will focus on one

admonition: present clear estimates of the quantities of interest as well as a statement about the degree of confidence one has in those estimates.<sup>12</sup> There are a few ways in which this admonition is frequently violated, and I would like to briefly draw your attention to them.

At least in the social scientific papers I read, explanations typically produce claims about the association between variables. Even when one is engaged in what looks like a descriptive exercise, like Huntington's attempt to demonstrate rising political instability, one is engaged in demonstrating that variables are related to each other in a particular way. If one wants to demonstrate that a phenomenon is changing over time, one must look at the relationship between that variable and time. If one wants to demonstrate that a particular behavior or attitude is more prevalent in some places or among some groups, one must look at the relationship between that variable and group membership or spatial location. Consequently, most of our empirical claims are about the relationship between variables. In a linear model we think of this quantity of interest as a slope coefficient, so I will use that terminology here, though the term 'derivative' might be even more appropriate.

A common way in which scholars become distracted from presenting the quantity of interest is by presenting something other than an estimate of a slope, when that is the quantity they are concerned with. For example, it has become common for scholars to plot the predicted probabilities from a logit model on the y-axis with some variable of interest on the x-axis when the quantity of interest is the association between a change in that predicted probability and a meaningful change in some variable of interest. The problem with doing so is that it requires the reader to infer the slope of that relationship from the picture. While it is true that slopes are not constant in non-linear models such as logit, and therefore the quantity of interest does not reduce to a single number, it would be better

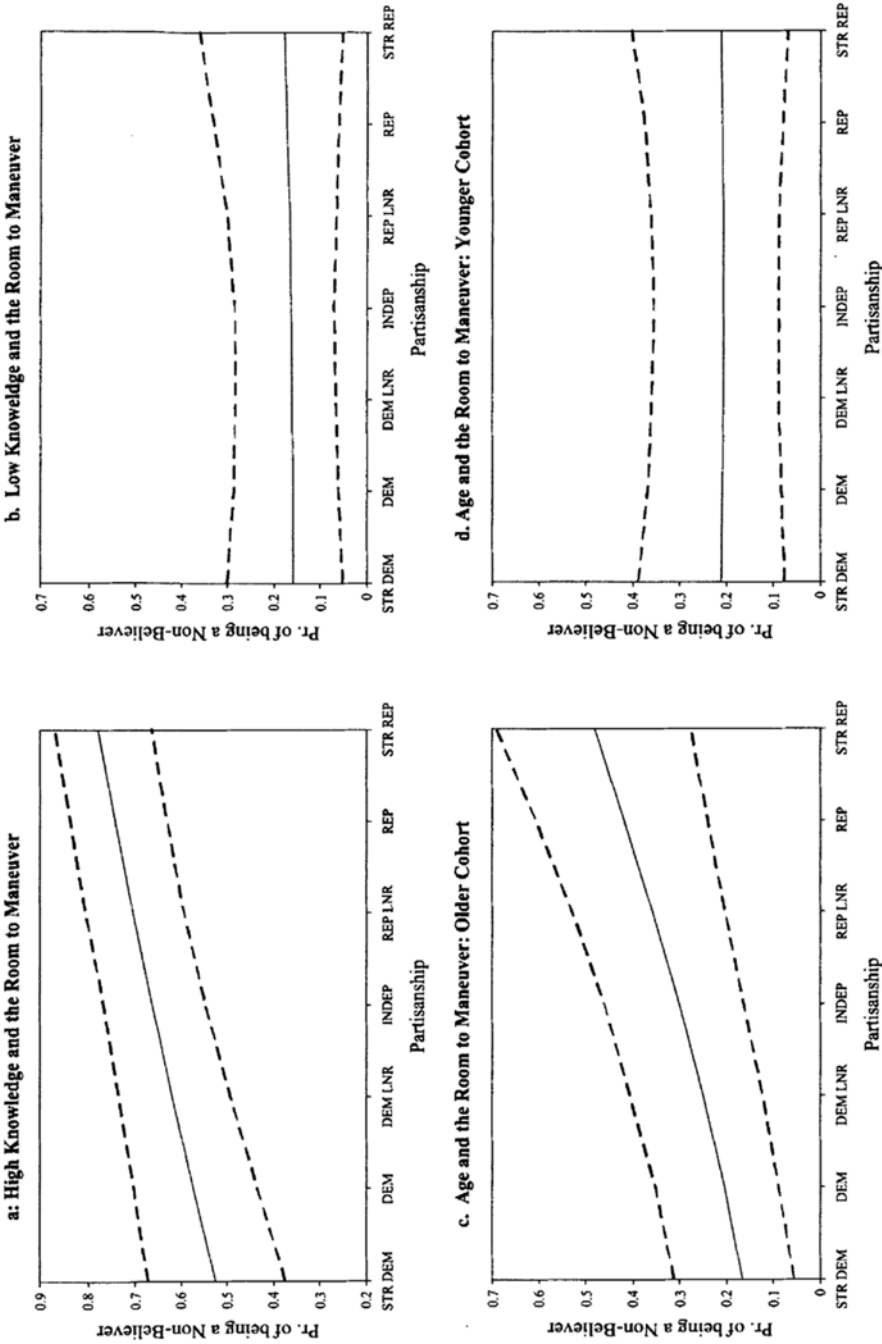
to plot the marginal effect of the variable of interest across a meaningful set of values of that variable of interest.<sup>13</sup> Adding confidence intervals around the predicted probability does not help because that tells the reader if the predicted probability is significantly different from zero, which is typically not the hypothesis being tested.

For example, Hellwig, Ringsmuth and Freeman (2008) present the graphs in Figure 1.2 as evidence for citizens' propensity to believe governments have little room to maneuver policy in a globalized economy. Each panel plots the predicted probability (and 90% confidence intervals) that a survey respondent said they did not believe the US government retains the 'room to maneuver' policy against the respondent's partisanship. The authors interpret the apparent difference between the slope of the plots in the left hand panel and the right hand panel as evidence that partisanship has an effect on respondent beliefs among respondents with college degrees (panel a) but not among those with high school degrees or less (panel b), and among respondents above the age of 59 (panel c) but not below the age of 40 (panel d). But what is the basis of this conclusion? The slopes on the right clearly look to be close to zero and, in comparison, the slopes on the left appear to be positive. But we are offered neither an estimate of the slopes for any degree of partisanship, nor an estimate of our uncertainty about that estimate. We can try to calculate the slope at different points on the line by estimating the 'rise over run' and we can kind of compare that estimate with the uncertainty implied by the error bars, but why make the reader construct a t-test from the picture rather than present that information for the reader by plotting marginal effects with their associated confidence intervals? Neither do the authors provide any evidence whether the slopes in the left hand panels are different from the slopes in the right hand panels. As a consequence, these pictures, and ones like them that appear frequently in

the literature, provide almost no *quantitative* evidence about the quantity of interest (under what conditions, if any, a change in partisanship is associated with a change in citizen beliefs about the government's 'room to maneuver').

Another common way of obscuring the quantity of interest is by presenting 'marginal effects' that are not marginal. It is commonplace for authors to say things like 'to gain some substantive understanding of these results, I note that a one standard deviation change in X is associated with a 0.056 change in Y'. The problem with this is that there is nothing typical or representative about a standard deviation – in data approximating a normal distribution, about two-thirds of all observations will be less than a standard deviation away from the mean. As a consequence, a change of a standard deviation in the variable of interest is not a particularly meaningful counterfactual to consider. This is particularly true where this practice is most frequently found – when interpreting the results of a non-linear model. Under this circumstance, the marginal effect of a variable is extremely sensitive to where it is being evaluated. The slope described by a 'marginal effect' the size of a standard deviation is likely to be very far from the slope of any estimated marginal effect within this interval. Another reason why this is not a particularly useful counterfactual comparison is that marginal effects are interpreted under a *ceteris paribus* clause where other factors are held constant – something which is not likely to be approximated in the real world when the variable of interest experiences an unusually large change the size of a standard deviation.<sup>14</sup>

Another common way in which scholars present information that is not the quantity of interest is when they have a hypothesis that is conditional in nature and either present results from an unconditional model or, equally common, estimate a conditional model but go on to interpret some of its results as if they were unconditional.<sup>15</sup>



**Figure 1.2 Partisanship and beliefs about 'room to maneuver': the conditional effects of knowledge and age**

Source: Hellwig, Ringsmuth and Freeman (2008, figure 2, p. 875.)

## Summary

My claim, up to this point, is that a paper, book or dissertation that has good answers to the five questions above will be a useful paper, book or dissertation. It does not follow that a paper, book or dissertation must have an *innovative* answer to all five of those questions. Progress can be made as long as one of the answers is better than existing answers and none are worse.

Which questions are ‘most important’ and, therefore, which ones should be the focus of your efforts to innovate? It is hard to say, though I believe that it is probably not best to try to explain something that no one has explained before. This is an important point. I have had many graduate students inform me gloomily that someone has beaten them to their ‘question’. My standard reaction is to say, ‘well, I doubt they have come up with the definitive answer, so what are you worried about?’ Since any question worth asking is likely to be difficult to answer, it is highly unlikely that another scholar is likely to beat you to the punch and have the last word on a subject. Indeed, if you are asking a question that no one else has asked, it should give you pause. Maybe it is not a very interesting question, or maybe there is something about asking the question in that way that led other scholars to believe productive answers were not forthcoming. That said, the mere fact that other smart people have asked the question does not mean it is a great idea for you to try to answer it.

Graduate students are told that they need to make an original contribution, which leads them to believe that they must ask a question that has never been asked, or at least never been answered, before. That is not true. Rather, an ‘original contribution’ requires only that the student provide a better answer to at least one of the questions mentioned above. So, if a student at the prospectus stage is going to attempt to offer a novel explanation, then part of their answer to question 2 should contain a statement about what they

bring to the table that might allow them to make progress where others have failed. What theoretical insight, methodological advantage or historical knowledge puts the author in a position to simultaneously recognize that ‘things have gone wrong’ with existing explanations and offer a solution that pushes the field in a promising direction?

Since ‘theoretical innovation’ is often thought to be the most prized contribution a political scientist can make, scholars often believe that a good paper should offer a novel explanation. I believe this comes, in part, from physics envy combined with the notion that theoretical physicists have a higher status than experimentalists. I believe that the idea that every important contribution must contain a theoretical innovation has greatly hampered the progress of our discipline. How is the accumulation of knowledge possible if, every time a scholar puts pen to paper, they have to offer a new explanation? Given frequently imperfect research designs and flawed empirical methods, I often think the opposite is true. We might be tempted to declare a moratorium on the development of new explanations until the discipline has reached consensus about empirical tests of the implications of existing explanations. As my critique of Huntington suggests, if we do not get at least some of the empirics right, how do we even know if our observations violate current theoretical expectations enough to warrant new explanations? One reason to resist such a temptation is that new theories do more than explain anomalies. For example, they also address conceptual and logical problems with existing explanations.

## PRACTICES THAT ENCOURAGE GOOD QUESTION ASKING

Following Kuhn’s line of reasoning above, it is worth asking what is likely to promote the skill, wit, and genius capable of recognizing when things have ‘gone wrong in ways that

may prove consequential'. Of Kuhn's three desiderata, 'skill' seems the least constrained by natural ability and, therefore, the most responsive to the environments we create. While artistic creation involves many aspects, a degree of craftsmanship is typically involved and craftsmanship is derived largely from practice. Extensive training in game theory and statistics is now commonplace in most graduate (and some undergraduate) programs in political science and international relations, and this is what is typically thought of when scholars evaluate the 'skills' of job applicants. These skills are important because without them scholars might ask questions based on faulty reasoning based on formal or informal fallacies such as the ecological fallacy, ad hominem attacks, hasty generalization, confusing correlation with causation, ignoring strategy induced selection effects, and failing to recognize the presence of confounds.

But while methods training is extremely helpful, it is not sufficient to produce scholars who ask and answer interesting questions. The problem sets typically assigned in quantitative methods and formal theory classes do help build the skills necessary to execute sophisticated research, just as playing scales and arpeggios builds the techniques necessary to execute sophisticated music. But there is more to training a musician than playing scales and arpeggios, because as important as scales and arpeggios are, they are not music. I have heard musicians criticized for having sufficient technique that they 'know how to say things on their instruments, but they do not seem to have anything to say'. The analogous criticism is frequently leveled at newly trained political scientists and international relations scholars.

So, what is to be done? To play good music, students have to listen to good music and they have to have a lot of experience making good music. Most graduate programs provide students with the equivalent of listening to music. When I was a newly minted PhD I heard Bruce Bueno de Mesquita give a lecture

at the Hoover Summer Program in Game Theory and International Politics at Stanford University. He built a game theoretic model based on the assumptions of hegemonic stability theory – seemingly on the fly, based on comments shouted out by my classmates. I had an epiphany. Of course, if developing social scientific explanations is an art, then it must be taught as the arts are taught! I was watching the master at the easel – engaged in the very craft I was trying to learn. It suddenly occurred to me that much of my graduate training amounted to the equivalent of sitting in a room listening to recordings of music, and then when it was time to write my dissertation it was as if a door had been flung open and I was handed an instrument I had never played (I imagined a cello) and pushed out onto a stage where I was expected to perform. Most graduate programs in political science teach people the equivalent of playing scales in methods classes and music history or appreciation in substantive classes, leaving them to figure out on their own how to put this together to make music.

The missing piece in most of our graduate education is what musicians call *etudes*. These exercises are designed to be music-like (so students can begin to think about interpretation and expression) but are artificially designed to allow for a degree of repetition of particular techniques (articulation, vibrato, dexterity) that allows those skills necessary for musical expression to seep into the student's muscle memory. Many doctoral programs emphasize that students should write publishable papers, but I believe that success is unlikely if this is attempted before students have engaged in many repeated attempts to explain things or think about what observations are implied by their explanations. Students need to practice asking and answering the five questions outlined above, and writing a single paper in each seminar does not give them the 'reps' to develop muscle memory. Virtually no skill worthy of the name can be developed after a dozen or so attempts.

Consequently, I have argued that problem sets in 'substantive classes' can help students become proficient at asking and answering the questions that will make for innovative research. An analogy to the visual arts might be useful. When students are learning to draw, they are not handed a blank sheet of paper and told to 'think of something interesting to draw, that no one else has drawn'. Rather, a bowl of fruit, or perhaps a wooden model of a human figure, is placed on a table. Then, everyone in the class draws the same thing, after receiving instruction from the instructor about how to do so. In contrast, many political science departments do the equivalent of handing their students a blank sheet of paper and telling them to 'draw something interesting'. Problem sets in substantive classes can be the equivalent of a bowl of fruit. The instructor can assign students to a question related to a particular research area: 'Explain why X occurs under Z circumstances.' 'If P explains Y, what else ought we observe?' 'Why is Q an interesting question?' 'Does Figure 2 count as confirming or disconfirming evidence for hypothesis 2, and why?'

Students need a lot of experience 'making music' before they 'have something to say'. If the analogy to the arts does not resonate with you, consider the following. Political science and international relations can take a lesson from the so-called bench sciences, where students work on many projects as members of large teams before they are tasked with the responsibility of deciding on the topic of the group's next project. Experience and repetition helps students learn what works and what does not.

While graduate pedagogy is important for stimulating creative question asking and answering, the broader climate and culture that we create in our departments and research centers is equally important. In particular, it is extremely important to create an environment where it is safe to play with ideas and challenge orthodoxy. I once had a colleague who, while walking down the hall, read a passage from a book that he

thought was incorrect and loudly declared the author an 'idiot'. Creativity and risk taking are not encouraged by a culture that suggests that only stupid people say stupid things. Instead, it is important to create the idea that the smartest among us are capable of error and that there is a big difference between *saying* something that is stupid and *being* stupid. To that end, I think it is extremely important for senior scholars to be transparent about the errors they have made. Young scholars need to learn that if they have made a mistake, they are in very good company, and if the requirement for admission was never making a mistake, the building would be empty.

While a culture of support for individual risk taking is vital to any scientific or artistic community, there is an optimal degree of individualism behind scientific discovery. If you don't read what everybody else reads and fail to train like everyone else trains, you will ask naïve questions that the rest of your community knows the answers to. But if you only read what everyone else reads, and only train like everyone else trains, you are unlikely to experience that moment when you see something that has gone wrong which no one else has seen.

Jazz bassist Scott LaFaro started playing the bass in 1954, when he was 19 years old, and in the few short years before he was killed in a tragic car accident in 1961, he completely changed the world's conception of what could be accomplished on a double bass and what role the instrument could play in a piano trio. Prior to playing the bass he had played the clarinet and saxophone for years, and many have attributed his phenomenal technical prowess to the fact that he practiced the bass by playing *etudes* composed for the clarinet by Hyacinthe Klosé in the 19th century (LaFaro-Fernández, 2009). The lesson LaFaro taught the world, in addition to the general benefits of interdisciplinarity, was: 'if you want to sound like everyone else, practice like everyone else; but if you want to sound like no-one else, practice like no one else'.<sup>16</sup>

Just as there is an optimal degree of individuality that is likely to produce scholars with the skill, wit and genius to determine when something has gone wrong in ways that may prove consequential, communities that strike the right balance between conformity and diversity are likely to encourage the habits that lead to scientific breakthrough.

On the one hand, it is important for a scientific community to share a commitment to the growth and dissemination of knowledge and a common understanding of the logic of inference and the standards of evidence. Without this shared understanding, criticism is likely to fall on deaf ears. But on the other hand, it is important for a community to be as diverse and eclectic as possible. People from different cultural, class, linguistic and religious backgrounds are likely to see the social world differently because they are likely to have had different experiences. These different experiences are likely to lead to diverse moral, political and social intuitions that lead them to raise questions that a more homogeneous group might not consider (Page, 2007).

In addition, diverse groups are less likely to fall prey to what I call ‘strategic confirmation bias’. Confirmation bias occurs when an individual embraces an idea uncritically because it conforms to their prior beliefs. When confirmation bias is at work, people are less likely to scrutinize the research practices that produced the claim in question. They are less likely to look for confounds, to ask about the details of data collection or to think critically about either the micro-foundations or the moral implications of a claim because the results confirm what they have long suspected about the world.

Strategic confirmation bias occurs when an individual is able to overcome first order confirmation bias and think critically about the claim being made, but is deterred from voicing the criticism because they believe that others are refraining from criticism as a result of confirmation bias. Under such circumstances, critically engaging the claim in public might signal to others that the critic does not share their beliefs on the matter.

Strategic confirmation bias is most likely to be a problem in communities where ‘everybody’ shares particular beliefs. In such an environment, thinking critically about a result that confirms the community’s beliefs could result in ostracism, or, at the very least, fewer dinner invitations. A community composed of individuals from diverse educational, class, religious and ideological backgrounds is less likely to produce the kind of monolithic views that encourage strategic confirmation bias. Individuals are more likely to say something when they see something wrong that may prove consequential because the set of taken for granted shared beliefs is likely to be smaller. Diversity is most likely to be helpful in this regard when the multiple dimensions of identity are relatively uncorrelated. If gender, race or ideology are heavily correlated, then dissent on one dimension can be seen as defection on another. Thus, in ideal circumstances, communities would have as much within group diversity as between group diversity.<sup>17</sup> Of course, diversity has to be sufficiently developed to give individuals confidence that speaking up under such circumstances will not simply confirm that one is an ‘outsider’. If a community promulgates the norm that in a multidimensional space we are all, on one dimension or another, outsiders, the cost of revealing that one ‘thinks differently’ about something is likely to be less costly. The daunting thing about strategic confirmation bias is that it is mostly likely to occur around issues about which scholars feel passionately. As a result, there is a danger that a research community will be least scientific about the matters that it cares most deeply about and most scientific about matters which its participants view as largely inconsequential.

## CONCLUSION

Good scientists ask interesting questions and are unsatisfied, even impatient, with bad



answers. I have argued that most work in political science and international relations can be understood through the lens of five questions and that contributions can be made to the literature by improving on a research community's answer to any of the five questions.

Since coming up with better answers to questions is as much art as it is science, I have argued that the best way to train good social scientists is to learn from the way in which artists are trained. Musical and visual artists learn their crafts through structured repetitive practice. The implication of this insight for the social sciences is that scholars should be given materials to work with that allow them to engage in the daily practice of asking and answering the five questions outlined in the first section of the chapter. I have suggested that the best way to encourage this is through the use of problem sets in our substantive courses. I have also hinted that there are great benefits to interdisciplinarity. By bringing habits, techniques and insights that are normal in one discipline into a setting where they are rare, individuals are more likely to recognize when something has 'gone wrong in ways that may prove consequential'. Finally, I have argued that diverse communities are more likely to produce good question askers, in part because they are less likely to fall prey to strategic confirmation bias.

## Notes

- 1 The author wishes to thank Branislav Slantchev and Laurie Clark for thoughtful comments and useful suggestions on an earlier draft of this chapter.
- 2 The fad around 'clever identification strategies' is but the most recent instantiation of this phenomenon.
- 3 In the words of Branislav Slantchev (in personal communication), 'theoretic innovation does not have to begin with an empirical observation but with a potential flaw in the logic, inconsistency of the assumptions, or an insight about a general claim (e.g., the impossibility results)'.
- 4 It is fashionable in many top graduate political science programs for faculty to say that 'substantive courses' are a waste of time and enterprising students should have an almost single-minded focus on methods training. It is also commonplace for professors to complain that their students are not adept at identifying interesting questions. I suspect that these phenomena are not unrelated.
- 5 The questions would have to be adapted to serve this purpose for literature reviews, methods papers and purely theoretical papers.
- 6 The scatter plot is based on data from an example from Pearson and Lee's (1903) examination of the correlation between the adult heights of fathers and sons.
- 7 Marx (1950 [1888]).
- 8 In contrast, if the only problem is a co-ordination problem then the mere dissemination of information is likely to be sufficient. But such problems are about as political as getting drivers to stay on their side of the road.
- 9 Though sometimes this works in the opposite direction. For example, experiments have shown that students who take economics classes behave much less cooperatively, and therefore more in line with the models learned in these courses.
- 10 *The Communist Manifesto* (Marx and Engels, 1996) had little immediate impact on embryonic socialist movements, but its long-run influence is undeniable.
- 11 In the possibly apocryphal words of theoretical physicist Wolfgang Pauli, it is 'not even wrong'.
- 12 King, Keohane and Verba (1994).
- 13 In the language of calculus: if the quantity of interest is  $dy/dx$ , then plot  $dy/dx$  against  $x$ , not  $y$  against  $x$ . The former tells the reader what they need to know. The latter makes the reader try to infer what they need to know from the picture.
- 14 See King and Zeng (2006) on 'the Dangers of Extreme Counterfactuals'.
- 15 See Brambor, Clark and Golder (2006) or Kam and Franzese (2007) for a fuller discussion.
- 16 At the same time, nearly every innovative jazz musician learned their craft by memorizing performances of musicians that came before them.
- 17 The connection between 'intersectionality' and cross-cutting cleavages should be explored further.

## REFERENCES

- Brambor, Thomas, William Roberts Clark and Matt Golder. 2006. 'Understanding Interaction Models: Improving Empirical Analysis', *Political Analysis* 14(1): 63–82.

- Freedman, David, Robert Pisani and Roger Purves. 2007. *Statistics*. 4th edition (New York: W.W. Norton).
- Geddes, Barbara. 2003. *Paradigms and Sand Castles: Theory Building and Research Design in Comparative Politics*. (Ann Arbor: University of Michigan Press).
- Hellwig, Timothy T., Eve M. Ringsmuth and John R. Freeman. 2008. 'The American Public and the Room to Maneuver: Responsibility Attributions and Policy Efficacy in an Era of Globalization', *International Studies Quarterly* 52(4): 855–80.
- Huntington, Samuel P. 1968. *Political Order in Changing Societies* (New Haven: Yale University Press).
- Kam, Cindy D. and Franzese, Robert J. Jr. 2007. *Modeling and Interpreting Interactive Hypotheses in Regression Analysis* (Ann Arbor: University of Michigan Press).
- King, Gary and Langche Zeng. 2005. 'The Dangers of Extreme Counterfactuals', *Political Analysis* 14(2): 131–59.
- King, Gary, Robert O. Keohane, and Sidney Verba. 1994. *Designing Social Inquiry: Scientific Inference in Qualitative Research* (Princeton: Princeton University Press).
- Kuhn, Thomas S. 1962. 'Historical Structure of Scientific Discovery', *Science* 136(3518): 760–4.
- LaFaro-Fernández, Helene. 2009. *Jade Visions: The Life and Music of Scott LaFaro* (Denton, TX: University of North Texas Press).
- Lave, Charles A. and James G. March. 1975. *Introduction to Models in the Social Sciences* (New York: Harper and Row).
- Lieberson, Stanley. 1991. 'Small *N*'s and Big Conclusions: An Examination of the Reasoning in Comparative Studies Based on a Small Number of Cases', *Social Forces* 70(2): 307–20.
- Marx, Karl. 1950 [1888]. 'Theses on Feuerbach' in Friedrich Engels, editor, *Ludwig Feuerbach and the End of Classical German Philosophy*. (Foreign Languages Publishing House: Moscow).
- Marx, Karl and Friedrich Engels. 1996. *The Communist Manifesto* (London: Pluto Press).
- Page, Scott E. 2007. *The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies* (Princeton: Princeton University Press).
- Pearson, K. and Lee, A. 1903. 'On the Laws of Inheritance in Man: I. Inheritance of Physical Characters', *Biometrika* 2(4): 357–462.
- Popper, Sir Karl. 1962. *Conjectures and Refutations: The Growth of Scientific Knowledge* (New York: Basic Books).
- Popper, Sir Karl. 2003 [1959]. *The Logic of Scientific Discovery* (New York: Routledge).
- Ray, Debraj. 1998. *Development Economics* (Princeton, NJ: Princeton University Press).
- Rauch, Jonathan. 2018. *The Happiness Curve: Why Life Gets Better after 50* (New York: St. Martin's Press).
- Skocpol, Theda. 1979. *States and Social Revolutions: A Comparative Analysis of France, Russia, & China* (New York: Cambridge University Press).



# From Questions and Puzzles to Research Project

Adam McCauley and Andrea Ruggeri

*Nam placuisse nocet*  
(Perfection brings destruction)  
*Metamorphoses*, Ovid

Regardless of perspective, philosophy, approach or assumptions, strong projects require a good research question. The process can be laborious, and the researcher will likely spend significant time in draft, making the question clear and explicit before it emerges in its polished form in the research project. For this reason and those outlined below, a scholar will face many of the core issues discussed in this chapter and its companion Chapter 1. For the lucky ones, the chapters on different methods will be eye-openers – whether on case studies and process tracing (see Chapters 59 and 62), formal theory (see Chapters 3 and 11) or the exploration of estimators in multivariate statistics (Chapters 33–57). However, each of these journeys begins with a singular prompt: what are you going to study?

This is the vital query of academic life. Any researcher, reviewer or dissertation committee

member will wonder the following: what main question is this research aiming to answer? What is the relevance of the inquiry? What contribution will the potential answers make to the wider field? How conclusive can the answer be? The first three interrogations help scholars generate crucial research questions and phrase them most cogently. The final interrogation urges scholars to consider whether their questions are likely to lead to neat or complete answers. Potential indeterminacy should not dissuade the researcher, but should embolden her to explore and explain as much, as rigorously, as possible.<sup>1</sup> Thus, this chapter explores how we can approach, imagine and generate research questions in international relations (IR) and how our answers can open pathways to related research projects. Our mission is to explore the craft of formulating research questions. Adopting the position of the craftsperson, beyond the idealized realm of the scientist, this chapter provides insights from published work and offers ‘rules

of thumb' to think systematically about the puzzles that motivate academic inquiry. We stop short of discussing different philosophies of social science and how they relate to the discipline, as there are excellent works on the epistemology of IR (Hollis and Smith, 1990; Fearon and Wendt, 2002; Jackson, 2016). Further, while we do not aim to dismiss other styles of research in IR, space limitations have forced us to be selective, and we have settled on the styles with which we are most familiar.<sup>2</sup>

### **WHERE RESEARCH QUESTIONS MIGHT COME FROM**

While the research question is a prerequisite for any academic project, for all the time spent on methodological training, theoretical instruction and historical exploration in our field, we devote few formal resources to teaching the art of question development. In handbooks and courses on research design, 'there is rarely any mention [...] of the specific challenges students face when devising a research question' (Bachner, 2012: 2).

Some ideas seem to emerge at random or through communing with colleagues in one's field. Often, though, questions emerge as the net results of unseen processes of thought – a mind's work tying together different ideas (old and new). These ideas then inform research questions, which, like their subsequent answers, become a product of continual evolution. Scholars can find it difficult to explain the origins of their interests in topics such as sanctions, international organizations, military occupation or civil wars. Even if the origins of their interest remain unknown, we argue that the practice of question formation can be looked at systematically, and pay careful attention to how we refine, specify and clarify research questions. Insights emerge from the working (and reworking) of one's inquiries and, as Bertrand Russell (2009) put it, 'I do not pretend to start with precise questions. I do not

think you can start with anything precise. You have to achieve such precision as you can, as you go along.' This intellectual journey is as important as the destination.

To this end, patience and commitment are vital for the generation of knowledge.<sup>3</sup> Scott Berkun challenges the very notion of an 'epiphany' – a word which suggests that inspiration comes from supernatural forces or beings – and suggests ideas emerge from a lifetime of hard work and personal sacrifice. What might at first glance appear to be driven by intuition or creative accomplishment is most often the result of systematic commitment to scholarly review. In this way, real innovation emerges from 'an infinite number of previous, smaller ideas' (Berkun, 2010) – it is not produced in splendid isolation. Alex Pentland (2014) illustrates the importance of social context for idea generation and finds intellectual strength in numbers, particularly when these crowds involve free-flowing ideas and a community engaged in the process of knowledge generation. These characteristics are essential for innovative and productive societies – and these insights ought to hold for smaller epistemic communities as well.

Steven Johnson (2011) agrees: 'World-changing ideas generally evolve over time – slow hunches that develop, opposed to sudden breakthroughs.' He argues that ideas may be aided by our better, stronger communication platforms, which allow us access to real-time information in ways never previously thought possible. Our age of communication also facilitates connections between researchers – and these networks retain intellectual weight to be leveraged in search of evocative questions. Through studying the complexities of innovation, Johnson highlights the importance of collaboration while stressing the value of preparation. Scholars must 'do the work' if they want to be in position for good fortune to strike. Finally, Johnson knows the route to understanding is not carved by successes alone, but by the failures that lead to mid-range solutions, and often to

refashioning something old to create something new.

For our purposes, we find the ‘literature review’ serves as the tangible product of this systematic approach: it aids scholars in identifying extant claims in the literature and helps orient their question to explore where the potential cause and effect might be identified. By knowing as much as we can about what works and what does not, or how things are assumed to work and might not, we are better able to generate pressing and potentially catalyzing research questions.

Connecting previously under-appreciated insights, however, demands that a scholar engage deeply across the broad expanse of IR scholarship. These scholars should explore beyond their sub-field into political science more generally (Putnam, 1988), and beyond their discipline into cognate areas of sociology (Wendt, 1999), psychology (Levy, 1997a), the detailed accounts of history (Gunitsky, 2014; Levy, 1997b) and models born of the study of economics (Hafner-Burton et al., 2017). Scholars will also benefit from their appetites beyond traditional scholarship, looking to fiction, television and films to leverage these created worlds to their advantage. Each of these creative spaces can convey important insights or stimulate questions that one’s discipline might not. With this background – and a notebook filled with potential topics – researchers will be quicker to identify unexplored perspectives. The novelty of these topics will also be readily ascertained through participation in conferences and targeted (subject-specific) workshops.

Burkus (2014) also argues that creativity emerges when an individual can connect different types of knowledge. In the academic context, this creativity emerges when researchers incorporate relevant knowledge in parallel (but perhaps under-connected) disciplines in new and novel ways. Further, scholars also benefit from working between different *sub-fields* within their discipline, where insights and research puzzles are often found. Returning to the practicality of

accomplishing this, scholars should look to the literature review as a forum for this considered and engaged intellectual exploration.

To this end, there are important differences between using and writing literature reviews, and the best illustrate how critical this initial survey can be. We can break this down into two versions of the literature review. The *private* review is for the scholar’s own use. This survey connects assumptions, inconsistencies and insights, while highlighting possibilities for improvement. The *public* review is for the consumers of the finished scholarly product. This explains and situates the research project within the discipline. The private review can (and often should) be partially sacrificed for the latter, where only the critically important items are cited. All too often, early career scholars want to signal their commitment by adding extra entries to these public literature reviews, failing to distinguish between the logics that underpin the exercise: the logic of exploration and the logic of explanation.<sup>4</sup> The final literature review, which adopts the logic of explanation, should serve as a storefront display displaying only the essential and most compelling items: neither customers nor your readers should have to search through the storage room to find what they came for.

The literature review embodies our philosophy for idea generation. This initial stage often focuses on identifying potential puzzles, and these early queries must be analyzed, challenged and sharpened before they can be polished for use. This process of discovery, through systematic reading and questioning, does not preclude creativity or emotional investment. In fact, these features are assets, given ‘the role of curiosity, indignation, and passion in the selection and framing of research topics’ (Geddes, 2003: 27). It is possible to be systematic and analytic while being motivated by indignation and passion, but it takes practice (Blattman, 2012). All ideas benefit from intellectual clarity, and *good* ideas, specifically, are the product of intellectual clarity. These good ideas will

only be stronger if their discovery is motivated by inspiring research questions.

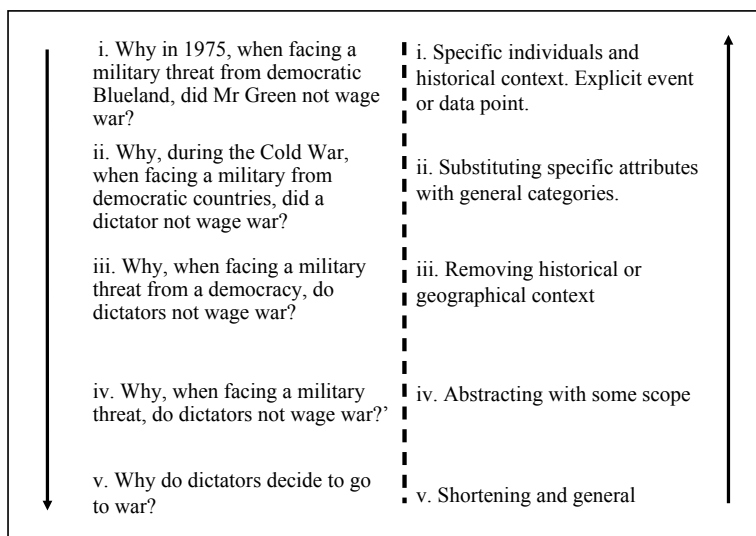
To this end, that clarity begins with simplicity.<sup>5</sup> Scholars should begin with the most basic and essential assumptions of how causes, effects and/or processes unfold, and construct their research queries in response. Consider this approach akin to Occam's razor: start from simple premises and proceed with the most basic explanation to build your initial theory. By removing all premises that are clearly incorrect, we are left with propositions that might plausibly capture elements of the answer. Simplicity distinguishes superior inquiries and research designs.<sup>6</sup>

## SPECIFIC AND GENERAL RESEARCH QUESTIONS

Academics often face the hedgehog or fox dilemma (Berlin, 2013): do you want to know a lot about one thing (hedgehog) or a little about many things (fox)? Are you interested in why the Mau Mau rebellion emerged in Kenya between 1952 and 1960, or are you interested in how civil wars begin? These

questions are related, and their relationship is important. While a scholar eager to explain why civil wars emerge might use the Mau Mau as a case of interest, a scholar focused only on the Mau Mau can generate lessons learned about this case alone. Solving the hedgehog and fox dilemma demands a scholar decide what questions she can answer, how comprehensive these answers can be and how she imagines her insights might apply to the wider universe of cases. Fundamentally, it requires that the researcher decide whether to privilege causal identification, which limits the questions she can ask, or whether to ask an unbounded question, while remaining cognizant of any methodologically imposed limits to the resulting answers.

Figure 2.1 presents a schematic of the process of abstraction for a research question. Although most scholars use this process intuitively and implicitly, visualizing the practice can be pedagogically useful. This process mirrors the form of Sartori's (1970) *ladder of abstraction* and provides a step-wise sequence for a researcher to explore alternative levels of inquiry that may emerge within the research agenda.



**Figure 2.1 From specific to abstract (and vice versa)**

Consider a fictitious dictator, Mr Green, who faced a military threat from a democratic country ('Blueland') in 1975 but decided not to wage war. The event is quite specific, as it focuses on only one person in a given year and within a single country. The second step calls for substituting the specific details of the case with more general categories.<sup>7</sup> We replace Mr Green with the general term 'dictator', change the specific date to a historical period ('the Cold War') and replace the country with the category of regimes to which it belongs ('democratic countries'). Of course, this process assumes that these different conceptualizations are possible, and, further, that their operationalizations are plausible. These issues (conceptualization, operationalization and measurement) should not constrain the question-generating process but should be tackled directly in later stages of the project. The third step up the ladder of abstraction sheds the temporal and geographical dependence. Not all research questions are ahistorical or devoid of geographical context. In fact, as Figure 2.3 will make clearer, we believe quite the opposite: historical and geographical contexts have complex interaction effects (Eckstein, 1998; Tilly and Goodin, 2006). However, developing a research question demands an early attempt to simplify by eliminating layers of specificity,<sup>8</sup> while recognizing that these layers can be added later when necessary. The fourth step focuses on the principal agent, the dictator, and drops the opposing regime type. In that way, we end up with a general research question and the project shifts from a study of interaction (dyadic type) that might explain war, to a study of a single regime type (monadic type) and its effects. This step retains some scope conditions by qualifying the circumstances that we are most interested in, namely, when the dictator is under a military threat. The final step reformulates the question so that instead of asking why something did not happen, we ask why it does. Although the puzzle was suggested by the absence of an event (i.e. the dictator did not go to war), posing

the question in positive terms provides clarity, especially when there might be a complex set of causal paths that explain the absence of any phenomenon.<sup>9</sup> This also echoes the preferred analytical posture of defining concepts in positive instead of negative and residual terms (Sartori, 1970). Thus, we settle on our final version: Why do dictators decide to go to war?<sup>10</sup>

Reading through case studies and historical accounts can be a useful starting point for exploring puzzles. Figure 2.1, viewed in reverse, illustrates a similarly valuable process of evolution, essentially 'stepping down' the ladder by adding specificity and context to the research question. This concentration of focus – shifting from the general to their specific elements – is a reflex of intellectual development and can be seen in many research agendas (see Figures 2.4 and 2.5 later in this chapter).

While ingenuity and originality are important for scholarly approach, all work lives within the ecosystem of extant scholarship. Researchers must not only motivate their specific question in the context of that scholarship but position their contribution among these existing studies. This means explaining how the findings clarify or qualify previous research; how their research remains relevant to the scholarly or policy communities; and – often overlooked – whether the aim of the project is feasible. One would do well to heed Merton's admonition from 1959 – still relevant today – that research agendas ought to: (1) focus on interesting and important phenomena in society; (2) lead to new studies of these social phenomena; (3) point to fruitful approaches for such studies; and (4) contribute to further development of the relevant research fields for these studies.

## RESEARCH QUESTION PILLARS

There are four 'pillars' that are often discussed in terms of framing and motivating a

research question: (1) a puzzle; (2) a gap; (3) a real world problem; and 4) feasibility (Figure 2.2).<sup>11</sup> Research does not have to speak to all of these pillars, but it is important not to overlook an important fifth pillar: passion for the topic. This final pillar can mean the difference between success and failure.

### **The Puzzle**

Most inquiries begin with a compelling puzzle. As Zinnes (1980: 318) puts it, ‘puzzles are questions, but not every question is necessarily a puzzle’. Gustafsson and Hagström (2018) propose a formula that succinctly captures what research puzzles look like: ‘Why x despite y?’, or ‘How did x become possible despite y?’, where x and y are two variables that can be isolated and studied. A puzzle formulated in this fashion is, admittedly, a research question, but one requiring much closer familiarity with the state of the art than the basic ‘why-x question’. Importantly, you will only discover such a puzzle by exploring prior research. During this exploration, scholars should remain open to the unexpected: did you find conflict where you expected cooperation, or order when you expected disorder? Most puzzles present after reorganizing prior arguments and findings, when this review begins to highlight the logical tensions and empirical contradictions. This, alone, ought to encourage systematic reading of the literature to ‘connect the dots’.

A puzzle that is particularly timely can be additionally appealing. Many scholars have been able to start new research agendas on ‘hot topics’ in contemporary politics. These scholars might gain a ‘first mover’ advantage, but the potential benefits must be weighed against the risks of investing in a research agenda with little systematic prior work. In the early 2000s, the ‘new kids’ on the academic block were the pioneers of the latest generation of civil war study (Collier and Hoeffler, 2004; Fearon and Laitin,

2003). Subsequently, those ‘new kids’ were replaced by the disaggregation generation. Today, the first mover advantage may go to scholars studying cyber security and artificial intelligence.

### **Filling a Gap**

As the second pillar in Figure 2.2 suggests, extant gaps in the literature present a clear starting point for new research projects. These gaps traditionally emerge at the weak points in previous theoretical frameworks or empirical designs. Researchers will benefit from challenging the assumptions that define either the piece of research or the related research agenda. For instance, this might include questioning how actors form preferences (Moravcsik, 1997) or whether actors adopt outcome-oriented rationality or process-oriented rationality (Hirschman, 1982). Perhaps extant research has focused too much on material incentives (capabilities, monetary aspects, etc.) and missed important dynamics associated with ideational aspects such as norms, ideas or emotions (Sanín and Wood, 2014; Checkel, 1998; Petersen, 2002). Perhaps there are stubborn, and under-analyzed, assumptions about the utility of specific strategies for a given outcome – e.g. that violence or force is an effective means of driving change. By challenging these assumption, scholars have developed pathbreaking work on the importance and influence of non-violent protest and its influence on regime change (Chenoweth and Stephan, 2011).

Challenging assumptions has a long history of sparking new research agendas in IR: where does action take place (Singer, 1961)? Where is the agency (Wendt, 1987)? Waltz (1959) wrote one of the core IR texts focusing mostly on this issue, asking: at what level should we analyze international politics – at the level of the individual, the subnational/local, the state/regime or international structure? Therefore, when we think about research questions and aim to formulate



What's the puzzle?	<ul style="list-style-type: none"> <li>• Are there contradictory findings?</li> <li>• Is an outcome unexplained?</li> <li>• Is the question puzzling per se?</li> </ul>	<ul style="list-style-type: none"> <li>○ clashing findings</li> <li>○ unexpected outcome</li> <li>○ question attractiveness</li> </ul>
Filling a gap?	<ul style="list-style-type: none"> <li>• What do you know?</li> <li>• How main contributions about this differ?</li> <li>• Strong/wrong assumptions in previous work?</li> <li>• Where is agency?</li> <li>• Is this gap due to non-interesting topic?</li> </ul>	<ul style="list-style-type: none"> <li>○ previous knowledge</li> <li>○ contradictions</li> <li>○ challenge assumptions</li> <li>○ level of analysis</li> <li>○ interesting</li> </ul>
Real-world problem?	<ul style="list-style-type: none"> <li>• How relevant is this?</li> <li>• Would a non-academic care?</li> <li>• Could it be translated into policies?</li> <li>• Are there ethical implications?</li> </ul>	<ul style="list-style-type: none"> <li>○ societal relevance</li> <li>○ audience reach</li> <li>○ policy implications</li> <li>○ academic ethics</li> </ul>
Methodological rigor?	<ul style="list-style-type: none"> <li>• Can you answer this question?</li> <li>• Ho can you answer this question?</li> <li>• Can you use new methods?</li> <li>• Are new data necessary?</li> </ul>	<ul style="list-style-type: none"> <li>○ scope</li> <li>○ method awareness</li> <li>○ method sophistication</li> <li>○ data awareness</li> </ul>
Will you enjoy it?	<ul style="list-style-type: none"> <li>• Is this topic 'yours' or imposed?</li> <li>• Can you add a creative twist?</li> </ul>	<ul style="list-style-type: none"> <li>○ autonomy</li> <li>○ creativity</li> </ul>

**Figure 2.2 Where good research questions come from**

new research agendas, it can help to identify whether the phenomenon can be scrutinized either at the micro level (i.e. individuals), at the macro level (i.e. structural features) or as an interaction between both. Increasingly, scholars have identified the value of an intermediate layer, called the meso level, where different organizational configurations (an identity group, the state and international organization) can connect both the micro and macro levels of analysis.

### ***Real-World Problems***

Enterprising research often aims to respond to 'real-world problems'. In part, this speaks to the utility of the intellectual practice: 'Our task is to probe the deeper sources of action in world politics, and to speak truth to power – insofar as we can discern what the truth is' (Keohane, 2009: 708). Karl Deutsch (1970), when discussing a new edition of Quincy Wright's *A Study of War*, wrote: 'War must be abolished, but to be abolished must be studied.' Pathbreaking research on security and strategy by Schelling (1960) clearly engaged with the

issues in his contemporary period. The researcher ought to ask whether their research question has consequences for wider society and what (or whom) might be influenced by their findings. Another way to orient our thinking is to explore whether, and what kind, of non-academic may find value in the work.<sup>12</sup> For non-academics, the incentive structure of publications and tenure-track preparation are replaced by a more tangible perspective of the work's value in terms of policy implications, implementation strategies, effectiveness and efficiency. These non-academics may be government analysts, politicians, practitioners employed by NGOs or civil servants at major inter-governmental bodies. These considerations ought to inspire scholars to avoid the crutch of academic jargon and to shy away from focusing on niche issues with unclear implications. In addition to crafting strong questions around specific themes or topics, researchers should strive for clarity in thought and communication. Empirical sophistication is no excuse for sloppy, senselessly convoluted style. Researchers would be well served to invest energy in crafting clear and 'economical' prose (McCloskey, 1985).

Despite perennial and contentious debates over the discordant responsibilities of the professional academic and the policy maker, students of IR have learned a few lessons. First, if authors are clear about possible policy implications, those policy makers will be more apt to read and incorporate their research. Second, clarity of argument and assertions should be considered a duty, particularly if your research resonates with contemporary policy challenges. If the policy implications are not made explicit, policy makers are likely to infer their own conclusions, potentially misusing the research. One debate around the 2003 invasion of Iraq focused on the Bush administration's misuse of the democratic peace literature in legitimizing the military intervention (Ish-Shalom, 2008). This point speaks to our question in Figure 2.2: what are the ethical implications of study? Researchers should be honest and self-reflective about potential challenges when it comes to the methodological approach and the practice of gathering evidence. Further, they should reflect on the effect of publishing publicly on findings with clear social costs: it is incumbent on the researcher to fully understand the consequences of their work, for themselves and everyone they have involved. Academic institutions have strict procedures and processes to assure that people who participate in a research project will not be facing any risks. Supervisors and senior colleagues should assist and advise early career scholars given the intrinsic dangers of studying certain phenomena in international relations.

### **Feasibility**

This pillar represents an important concern in political science and IR. Our first question is trenchant, though problematic: will you have the necessary data and methods to find an answer to your new research puzzle? This question is a necessary first inquiry, as it may

also be the last: supervisors and more experienced scholars are always concerned with the feasibility of the project and might try to refocus the core question if the answer demands resources and skills beyond what is currently available. It is possible that these individuals are wrong – perhaps they are not creative or visionary enough to see the potential – but one ignores this concern at one's own peril. It is difficult to know early in the process what methods and data will be most useful, and whether they will be available. Effective review of methods and existing data is a critical step (and one with which this *Handbook* is designed to assist). Scholars should find out whether the requisite data exist, but not despair immediately if the answer is unclear. A range of successful research projects involve data collection, sometimes exclusively so (Singer and Small, 1994; Sarkees and Schafer, 2000; Vogt et al., 2015; Sundberg and Melander, 2013; Tierney et al., 2011). However, one should be careful, as large data collection is quite demanding and expensive – most of the existing datasets are the product of years of research and were obtained by small armies of coders or through use of advanced automated data gathering tools. PhD students and early career scholars could face serious challenges if the data collection is not constructed carefully. Pragmatism is an asset in assessing the realities of obtaining the necessary data.

### **Enjoying Your Research**

Undertaking a research project comes with associated opportunity costs – time, energy and interest expended on any research project are resources a scholar cannot use elsewhere. Thus, it may matter whether the research project is strictly one of your choosing or one suggested (or even imposed) by someone else. Suggested topics may, at times, come from a supervisor or academic peers, and a particular approach may be imposed according to the availability of data.

You may also have stumbled into your research question through previous study or thanks to an essay or assignment you have already written. You need to ask yourself: does the topic continue to inspire you? Do you enjoy working on it?

In the end, a research project is 10 percent inspiration and 90 percent perspiration, and while a creative mind is an asset, the eventual result will depend on your commitment, hard work and effective time management. Further, scholars should also be aware that starting a completely new or novel research agenda is comparatively rare in an academic life – and perhaps a luxury, as well. However, understanding your personal interest and stakes in a project is important for deriving joy from your work and will be helpful in motivating scholars to finish their projects.

## DEVELOPING RESEARCH QUESTIONS

We now consider a range of possible research questions, as shown in Figure 2.3. This list is not exhaustive, but it does provide a starting point for crafting them. Broadly speaking, there are two types of research question. The first focuses on the main phenomenon to be explained (Y, the dependent variable, or *explanandum*). The second focuses on factors that can explain variations in that Y (X,

the independent variables, or *explanans*). Both types are legitimate queries for research, although they might involve different methods and approaches.<sup>13</sup>

The first ideal type – ‘what is Y?’ – is the most direct of the explanandum-oriented research questions. This question is highly theoretical, involves heavy conceptualization and is likely to produce typologies that might be useful for follow-up questions (Collier et al., 2012; Gerring, 1999). ‘What is Y’ queries also demand data to describe different trends and types (Gerring, 2012). Although we use ‘power’ as our example, we could easily have picked something else from a wide range of similar forms of inquiry in IR: security (Baldwin, 1997), human security (Paris, 2001), securitization (Buzan, 2008), or civil war (Sambanis, 2004; Kalyvas, 2005).

The second ideal type remains focused on Y, but instead of seeking to provide a conceptual definition, it tracks how Y has changed over time and space. Research agendas such as those charting countries’ Polity IV scores (Marshall and Jaggers, 2002) or calculating their position on the V-Dem spectrum (Coppedge et al., 2017) belong to this family of research questions.

The third type also starts with a definition of Y, but proceeds to provide potential explanations of the data-generating process while maintaining that definition. These questions are very useful for elaborating and studying

1. What is Y?	→ What is power?
2. How has Y changed?	→ How has trade increased?
3. Why Y?	→ Why war?
4. Under what conditions Y?	→ Under what conditions peace?
5. Do Y and X co-vary?	→ Do democracy and peace correlate?
6. Does X cause Y?	→ Do international organizations cause cooperation?
7. What is the effect of X on Y?	→ What is the effect of aid on civil war?
8. Is the effect of X on Y mitigated by Z?	→ Is the effect of democracy on war conditional on trade?
9. Why Y varies across G or T?	→ Why level of cooperation varies across regions?
10. Why X affect Y in T but not in T-1?	→ Why alliances influence risk of war differently over time?

**Figure 2.3** On research question types

mechanisms (Tilly, 2001; Hedström and Ylikoski, 2010). One of the most cited articles on processes leading to war, ‘Rationalist Explanations for War’ (Fearon, 1995), belongs to this style of research. What is an appropriate way to define war? What can account for the occurrence of war thus defined? The conceptual framing of the question and the provision of several tentative mechanisms that could answer the puzzle posed by the definition have made this article foundational in the study of war.

The fourth type is the natural extension of the third and involves starting with the classic phrase ‘*under what conditions...*’. This framing is useful because it pushes the researcher to think about variations of Y, their central variable of study, as well as the influence owing to variation of possible explanatory factors – the Xs. Hence, *under what conditions* suggests we should be conscious of co-variation.

The next three types of question seek to make this co-variation more explicit. Here, the choice of adjective or verb suggests the nature of that correlation. Specifically, in Figure 2.3 question type six outlines the *causes of effects*, while question seven refers to the *effects of causes*. This distinction is easy to state, but harder to distinguish in practice. Effects may have multiple causes and these causes inevitably give rise to a range of effects. Situating your research within this space is important and should be done consciously. Moreover, these ideal types are particularly important for our contemporary study of IR, and have attracted the attention of contemporary scholars (Van Evera, 1997; Samii, 2016; Lebow, 2014). Researchers should be cautioned against slipping into unfounded causal claims, however. Are democracies richer? Do democracies fight each other less than they fight other regime types? Do countries that are members in many IOs trade more than countries with fewer memberships?

For this, one could ask directly: does X cause Y? It used to be that quantitative scholars were less concerned about the differences between correlational and causal claims. Many incorrectly assumed that endogenous effects (those

that emerge within the phenomenon of study) could be suitably controlled by lagging (modulating the expected influence) the relevant variables, and that omitted causal effects could be discerned by imposing specific scope controls. The emergent consensus about the importance of causal identification has changed all that (Angrist and Pischke, 2008; Samii, 2016). Projects that had employed apparently solid quantitative causal identification strategies can now be challenged with qualitative and archival information that casts doubt on the supposed exogenous identification (Kocher and Monteiro, 2016). Qualitative scholars are also increasingly aware of the high bar for causality, which has led to the development of sophisticated approaches to methods such as process-tracing (Bennett and Checkel, 2014).

The last three ideal types of questions add layers of complexity. They ask whether the effect of X on an outcome is conditional on, or mitigated by, another factor Z. These queries tend to make research agendas more dynamic and expansive since they push researchers beyond studying X and Y alone. This can involve theoretical and methodological improvements. For example, quantitative scholars have been developing better ways of studying interactions between variables (Brambor et al., 2006; Franzese and Kam, 2009), while qualitative scholars have provided systematic approaches to conditionality and contextuality (Tilly and Goodin, 2006). The additional factors might also involve time (different historical periods or temporal moments) or space (regions or other appropriate geographical variables).

## EXAMPLES FROM IR: DEMOCRATIC PEACE AND CIVIL WAR

We now sketch some lessons learned by looking (briefly) at two research agendas in IR: democratic peace and civil war. Readers who are not specifically interested in the

topic of security in IR could seek to draw parallels with other areas of IR or more broadly in Political Science.<sup>14</sup>

Figure 2.4 sketches a (non-comprehensive) trajectory of research questions and topics about the democratic peace. This is one of the final IR debates in which paradigms remain in contention.

The research agenda began with a broad research question, with subsequent studies providing more stringent scope conditions and improved accuracy of empirical analysis. Subsequent debates have centered on how to make sense of the correlational findings: if democracies do not fight each other, we need mechanisms to explain why this might be the case (Maoz and Russett, 1993). In response, several papers, mostly from a realist perspective, challenged the pattern of causation and the conceptualization, while hinting toward possible omitted variable bias (Rosato, 2003). Given the rarity of the phenomenon (not just war, which is already rare, but war that involves democracies on opposite sides, which should be even rarer, as the central assertion

in this literature would have it), critics argue that it is difficult, and perhaps foolhardy, to make inferences from such a small sample. In response, scholars offered more elaborate explanations, finding cause for the seemingly pacific relations between democracies in variables such as trade or international organizations (Oneal and Russett, 1999; Russett and Oneal, 2001).

Subsequent developments adopted non-realist perspectives, and either attempted to shift the focus to variables other than regime type (e.g. capitalism (Gartzke, 2007), although see Dafoe (2011) for a methodological critique), or to dig deeper into the microfoundations of the proposed mechanisms (Tomz and Weeks, 2013).

Figure 2.5 offers a schematic of how research on civil wars has evolved. The study of civil wars has an established research pedigree but it experienced something of a renaissance in the 2000s (Kalyvas, 2010), as research became more systematic and publications more abundant (Fearon and Laitin, 2003; Collier and Hoeffler, 2004).

- Causes of war?
  - Specification : Are democracies more peaceful?
    - Further specification: Do democracies not fight each other?
- Mechanisms : Why democracies do not fight? Norms, structure, endogeneity
- Criticisms: Correlation no causation, conceptualization problem, omitted variable bias, rare event, non-representative sample
  - Reactions: larger samples, further controls, different methods
  - Extensions: international organizations, trade and democracies
- Alternative within: Is it capitalism or democracy?
  - Technical discussions : specifications, operationalization...
  - New methods: survey experiments...

**Figure 2.4 Stylized democratic peace research agenda**

- i. What does explain variation in civil war (cw) onset?
  - plethora of Xs
- ii. Refocus on cw facets: durations, intensity, civilians' victimization, outcomes, legacy...
- iii. Analytical and empirical refocus: dyad, transnational, groups, local data, leaders...
- iv. Refocus on actors' actions: governance, alliances, splintering, displacement, crime...

**Figure 2.5 Stylized civil war research agenda**

This period of academic exploration offered a range of arguments about the variables related to the onset of civil wars, moving beyond the traditional economic or ethnic factors. Many studies sought to explain the variation in civil wars with capabilities (of host states and non-state actors), countries' exports, domestic shares of natural resources and demographic patterns, among others. Scholars unpacked different facets of civil wars by looking carefully at their duration, intensity and level of violence against civilians (Kalyvas, 2006). Some proposed different levels of analysis to match the theoretical concepts and empirical operationalizations (Cederman and Gleditsch, 2009). This work has benefited from intensive data-collection exercises about the actors involved (Cunningham et al., 2009), the transnational linkages between these actors (Gleditsch, 2007), other relevant local features (Sundberg and Melander, 2013) and leader attributes (Prorok, 2016). Qualitative work has not lagged behind, either, with research on rebel governance (Arjona, 2016), systems of alliances (Christia, 2012), and their splintering and their influence on displacement (Steele, 2009). More recently, scholars have started to explore how the processes leading to political violence might be related. We now have studies of the substitution effects between civil wars and terrorism (Polo and Gleditsch, 2016), and between civil war and military coups (Roessler, 2016). The field of inquiry has also expanded to incorporate insights from the study of organized and transnational crime (Kalyvas, 2015; Lessing, 2017).

As both our brief examples illustrate, research agendas progress through systematic refinement. Scholars challenge each other's intuitions, assumptions and findings, and – in the process – reassess the models and methods being employed. Progress is built on failure, but it is failure forward, toward better understanding. The initial conceptual thought about the nature of a phenomenon ('what is security?') provides a center of gravity that attracts subsequent research that unpacks the

latent assumptions, alters the levels of analysis, and critically engages the conclusions reached. These improvements often come from leveraging new methods or specifying new mechanisms that connect the variables. These interventions often offer competing explanations of empirical patterns, which, in turn, pique the interest of new generations of scholars who jump in to analyze under-explored or emergent puzzles associated with these phenomena.

## CONCLUSION

Can you pass the 'elevator test'? That is, can you succinctly explain your research question and project aims in less than 45 seconds? The time limit is arbitrary, but the question helps sharpen the clarity of your work. If you cannot pose a short coherent question and provide the gist of the answer, then there is something amiss with your project. One often achieves this clarity after repeatedly writing and rewriting one's ideas until they are thoroughly understood. You cannot effectively communicate things you only dimly understand. All acquisition of knowledge relies on commitment, iteration and a significant amount of writing, thinking, and writing again. 'You do not learn the details of an argument until writing it up in detail, and in writing up the details you will often uncover a flaw in the fundamentals' (McCloskey, 1985: 189).

Clear communication is crucial because journal editors and book publishers ask reviewers to assess where a project fits between four categories: a relevant contribution to a relevant topic; an irrelevant contribution to a relevant topic; a relevant contribution to an irrelevant topic; or an irrelevant contribution to an irrelevant topic. While it is usually easy to avoid falling into the last category, careless write-ups and hidden flaws in the research process can often strand submissions in the second and third

categories. Depending on the patience of referees, this might also lead to outright rejection. Always be conscious of how to position yourself in that first category.

Thus, research projects are not simply assessed in terms of good or bad. They can be good or bad across a set of characteristics – from analytical rigor to methodological appropriateness – and should be evaluated across multiple dimensions.<sup>15</sup> It is also important to ask whether research can be ‘creative and inspiring in a swashbuckling way’, as scoring high on this dimension often signals that the research agenda can make a substantive difference. Of course, being rigorous and innovative is great, but every researcher should strive for intellectual humility, remaining cognizant of the potential costs and risks involved.

We conclude our chapter with a brief ‘Questions & Answers’ section that might be instructive to readers.

### ***How Specific Should a Research Question Be?***

From ‘why do civil wars start?’ to ‘why did the Mau Mau rebel against British rule?’, there are varying degrees of generality. However, as we emphasized in our discussion, scholars can ask general questions and explore them through specific cases or start with specific cases and build toward more general abstractions. The answer depends on what scope conditions you select and how generalizable you wish to be. There are tradeoffs on either end of this spectrum.

### ***Shall I Only Select Questions I Can Fully Answer?***

This is an important question and there might be disagreements about the answer. When discussing Figure 2.2, we stressed the trade-off between originality and feasibility. However, we also suggested that there are

myriad ways of assessing the potential and value of research agendas. Our advice is to be more curious than concerned in the exploratory phase, while rationally assessing the feasibility before you commit to the project over the longer period. Creativity and innovation are important features of any research project.

### ***How Useful Is a Literature Review When Posing a Research Question?***

Researchers with a thorough knowledge of previous research (in content and method) will have a demonstrable advantage over those without. Remember the distinction between a logic of exploration and a logic of explanation when writing literature reviews. The former is crucial for posing a proper research question; the latter is vital when writing in a piece of research and situating its contribution and relevance.

### ***Do I Need to Justify My Research Question?***

Researchers must do more than justify their topic – they need to situate their research questions. What previous research are you engaging with or challenging? What research streams does your project bring together? What is your intended contribution (what will we learn within IR or Political Science?) and how is it relevant (why should we care?)

### ***What Do I Need to Define in My Research Question?***

Definitions are central for research, but their importance varies with the type of research question that you are engaging with. If your question is a Y-focus query, most of the research will be about conceptualization, defining ideal-types and discussing typologies. X-focused research questions will also

require clear definitions, but this will likely occur at a later stage in the research design.

### ***How Many Research Questions?***

A good rule for any piece of writing or research is one paragraph, one idea. Applied to research, a reasonable rule of thumb might be one dominant research question for each paper, where a book or thesis might have more – though usually these questions are closely related. This predominantly takes the form of a large research question with different, but reinforcing, sub-questions that can improve the comprehensiveness of the project. Remember, the more research questions you pose, the harder it will be to answer them. Less is more, particularly if those questions are comprehensively studied and answered.

### ***Do I Need to Keep the Same Dependent Variable?***

Usually, yes. Switching among things you are trying to explain can be disorienting for a reader unless your central argument is about a selection process ( $Z \rightarrow X$ , then  $X \rightarrow Y$ ) or a recursive one ( $X \leftrightarrow Y$ ). Moreover, when you are formulating hypotheses, it is good practice to keep the Y consistent. Consistency improves project clarity.

### ***Do I Need Only One Main Explanatory Variable?***

Not necessarily. International relations, and politics in general, are complex, and several factors are often necessary to account for the variation of IR phenomena. However, scholars often mix main explanatory factors (what should be the core contributions of the research project) with controls (other factors that could be affecting the dependent variable). This is especially prone to happen in the

early stages of the research projects. Not all Xs should be placed under the spotlight. Focus on those Xs that make a contribution and remain essential to your research question.

### ***Can I Change My Research Question?***

There are two ways to interpret this question: as a change of topic, or as a change of the specific question being asked. The former is beyond our scope, as it comes down to the individual researcher or to their relationship with the supervisor or PhD advisor. The latter, however, may simply happen as you read more of the literature, strengthen your theoretical argument and advance your empirical analysis. Not all change is bad, but be conscious of why you are doing it.

### ***Do I Need to Specify Actors, Preferences, Strategies and Other Relevant Components of My Theory in the Research Question?***

Not necessarily, but they can assist in formulating the research question and project directives. Understanding at which level the analysis is being conducted, where agency is located, what preferences are involved and how they are formed can be crucial. These details help us explain the strategies actors adopt and help identify the structural constraints and ideational factors that might influence them. This information is essential for theorizing and can help refine your research question.

### ***When Should I Ignore Your Advice?***

We, like everyone else, are limited by our own subjectivity. In any domain where creativity is central and where the routes to



knowledge are many and varied (and sometimes deeply personal), there is a risk of dissuading researchers from following their own intuition and insights. We have tried to be systematic and egalitarian in our treatment of differences, but we cannot possibly account for all situations. We hope that our advice highlights common issues and potential solutions, so take our advice as formative, not conclusive.

## ACKNOWLEDGMENTS

We thank Branislav Slantchev, Juan Masullo, Francesco N. Moro for their comments on this chapter. With this chapter we want to remember the tragic loss of Giulio Regeni in Cairo. He was a graduate student conducting field research.

## Notes

- 1 The authors would like to thank Juan Masullo for his comments on this distinction between conclusiveness and rigor.
- 2 As Rob Franzese's saying goes: we summarize and report the best we know without implying that it is the best out there.
- 3 Procrastination and perfectionism are its enemies.
- 4 We owe this helpful distinction to a conversation with Monica Duffy Toft.
- 5 As Kristian Skrede Gleditsch would put it, simplicity is the simpler version of the term parsimony.
- 6 Bruno Munari (1981) compares the passages of research design building to the steps of cooking.
- 7 In this way, our proposed strategy is consonant with that of others who advocate taking inspiration from a specific event or actors before formulating the wider research question (Kellstedt and Whitten, 2018).
- 8 Here the parallel is with sculpture – getting to the gist of it. Hence, the research question-generating process should be similar to the craft of sculpture.
- 9 This could be labeled the 'Anna Karenina problem', where all happy families are alike and all unhappy ones are unhappy in their own way.
- 10 Weeks (2014) studies this very question.
- 11 King, Keohane and Verba, 1994; Gustafsson and Hagström, 2018; Bachner, 2012.
- 12 There is a difference between the 'grandma test' and the 'policy-maker test'. The former is about clarity and communication of a research project; the latter is about its implications and its answer to the non-trivial question 'so what?'
- 13 Stathis Kalyvas, in a personal communication, suggests that there are Y-focus and X-focus research projects, and often scholars stick to this divide in their careers. It is the case that certain methods work more effectively for Y-focus questions and others – qualitative methods and quantitative methods – for X-focus questions. Again, though, this only serves as a rule of thumb and is not a law.
- 14 Keohane (2009) points out, however, that '[t]he study of world politics begins with the study of war' before turning to the vital question: 'Why is war a perennial institution of international society and what variable factors affect its incidence?'
- 15 Personal communication with Brian Burgoon, whose elaboration is from Charles Sabel's categorization.

## REFERENCES

- Angrist, Joshua D., and Jörn-Steffen Pischke. 2008. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.
- Arjona, Ana. 2016. *Rebelocracy*. Cambridge University Press.
- Bachner, Jennifer. 2012. 'The Common Mistakes Students Make When Developing Research Questions.' Unpublished manuscript, Johns Hopkins University.
- Baldwin, David A. 1997. 'The Concept of Security.' *Review of International Studies* 23 (1): 5–26.
- Bennett, Andrew, and Jeffrey T. Checkel. 2014. *Process Tracing*. Cambridge University Press.
- Berkun, Scott. 2010. *The Myths of Innovation*. Updated and expanded paperback edition. O'Reilly.
- Berlin, Isaiah. 2013. *The Hedgehog and the Fox: An Essay on Tolstoy's View of History*. 2nd edition. Princeton University Press.
- Blattman, Christopher. 2012. 'Children and War: How "Soft" Research Can Answer the Hard Questions in Political Science.' *Perspectives on Politics* 10 (2): 403–13.

- Brambor, Thomas, William Roberts Clark and Matt Golder. 2006. 'Understanding Interaction Models: Improving Empirical Analyses.' *Political Analysis* 14 (1): 63–82.
- Burkus, David. 2014. *The Myths of Creativity: The Truth about How Innovative Companies and People Generate Great Ideas* [Electronic resource]. First edition. Jossey-Bass.
- Buzan, Barry. 2008. *People, States & Fear: An Agenda for International Security Studies in the Post-Cold War Era*. ECPR Press.
- Cederman, Lars-Erik, and Kristian Skrede Gleditsch. 2009. 'Introduction to Special Issue on "Disaggregating Civil War".' *Journal of Conflict Resolution* 53 (4): 487–95.
- Checkel, Jeffrey T. 1998. 'The Constructive Turn in International Relations Theory.' *World Politics* 50 (2): 324–48.
- Chenoweth, Erica, and Maria J. Stephan. 2011. *Why Civil Resistance Works: The Strategic Logic of Nonviolent Conflict*. Columbia University Press.
- Christia, Fotini. 2012. *Alliance Formation in Civil Wars*. Cambridge University Press.
- Collier, David, Jody LaPorte and Jason Seawright. 2012. 'Putting Typologies to Work: Concept Formation, Measurement, and Analytic Rigor.' *Political Research Quarterly* 65 (1): 217–32.
- Collier, Paul, and Anke Hoefler. 2004. 'Greed and Grievance in Civil War.' *Oxford Economic Papers* 56 (4): 563–95.
- Coppedge, Michael, John Gerring, Staffan I. Lindberg, Svend-Erik Skaaning, Jan Teorell, David Altman, Michael Bernhard et al., 2017. 'V-Dem Dataset V7.'
- Cunningham, David E., Kristian Skrede Gleditsch and Idean Salehyan. 2009. 'It Takes Two: A Dyadic Analysis of Civil War Duration and Outcome.' *Journal of Conflict Resolution* 53 (4): 570–97.
- Dafoe, Allan. 2011. 'Statistical Critiques of the Democratic Peace: Caveat Emptor.' *American Journal of Political Science* 55 (2): 247–62.
- Deutsch, Karl W. 1970. 'Quincy Wright's Contribution to the Study of War: A Preface to the Second Edition.' *Journal of Conflict Resolution* 14 (4): 473–8.
- Eckstein, Harry. 1998. 'Unfinished Business: Reflections on the Scope of Comparative Politics.' *Comparative Political Studies* 31 (4): 505–34.
- Fearon, James D. 1995. 'Rationalist Explanations for War.' *International Organization* 49 (3): 379–414.
- Fearon, James D., and David D. Laitin. 2003. 'Ethnicity, Insurgency, and Civil War.' *American Political Science Review* 97 (1): 75–90.
- Fearon, James, and Alexander Wendt. 2002. 'Rationalism v. Constructivism: A Skeptical View.' In Carlsnaes, Walter, Thomas Risse, and Beth A. Simmons, eds. *Handbook of International Relations*, 52–72. Sage.
- Franzese, Robert J., and Cindy D. Kam. 2009. *Modeling and Interpreting Interactive Hypotheses in Regression Analysis*. University of Michigan Press.
- Gartzke, Erik. 2007. 'The Capitalist Peace.' *American Journal of Political Science* 51 (1): 166–91.
- Geddes, Barbara. 2003. *Paradigms and Sand Castles: Theory Building and Research Design in Comparative Politics*. University of Michigan Press.
- Gerring, John. 1999. 'What Makes a Concept Good? A Criterial Framework for Understanding Concept Formation in the Social Sciences.' *Polity* 31 (3): 357–93.
- Gerring, John. 2012. 'Mere Description.' *British Journal of Political Science* 42 (4): 721–46.
- Gleditsch, Kristian Skrede. 2007. 'Transnational Dimensions of Civil War.' *Journal of Peace Research* 44 (3): 293–309.
- Gunitsky, Seva. 2014. 'From Shocks to Waves: Hegemonic Transitions and Democratization in the Twentieth Century.' *International Organization* 68 (3): 561–97.
- Gustafsson, Karl, and Linus Hagström. 2018. 'What Is the Point? Teaching Graduate Students How to Construct Political Science Research Puzzles.' *European Political Science* 17 (4): 634–48.
- Hafner-Burton, Emilie M., Stephan Haggard, David A. Lake, and David G. Victor. 2017. 'The Behavioral Revolution and International Relations.' *International Organization* 71 (S1): S1–S31.
- Hedström, Peter, and Petri Ylikoski. 2010. 'Causal Mechanisms in the Social Sciences.' *Annual Review of Sociology* 36: 49–67.
- Hirschman, Albert O. 1982. *Shifting Involvements: Private Interest and Public Action*. Princeton University Press.

- Hollis, Martin, and Steve Smith. 1990. *Explaining and Understanding International Relations*. Oxford: Oxford University Press.
- Ish-Shalom, Piki. 2008. 'Theorization, Harm, and the Democratic Imperative: Lessons from the Politicization of the Democratic-Peace Thesis.' *International Studies Review* 10 (4): 680–92.
- Jackson, Patrick Thaddeus. 2016. *The Conduct of Inquiry in International Relations: Philosophy of Science and Its Implications for the Study of World Politics*. Routledge.
- Johnson, Steven. 2011. *Where good ideas come from: The natural history of innovation*. Penguin Audio.
- Kalyvas, Stathis N. 2005. 'Warfare in Civil Wars: Stathis N. Kalyvas.' In Isabelle Duyvesteyn and Jan Angstrom, eds. *Rethinking the Nature of War*, 99–119. Routledge.
- Kalyvas, Stathis N. 2006. *The Logic of Violence in Civil War*. Cambridge University Press.
- Kalyvas, Stathis N. 2010. 'Civil Wars.' In Carles Boix and Susan C. Stokes, eds. *The Oxford Handbook of Comparative Politics*, 416–434. Oxford University Press.
- Kalyvas, Stathis N. 2015. 'How Civil Wars Help Explain Organized Crime – and How They Do Not.' *Journal of Conflict Resolution* 59 (8): 1517–40.
- Kellstedt, Paul M., and Guy D. Whitten. 2018. *The Fundamentals of Political Science Research*. 3rd edition. Cambridge University Press.
- Keohane, Robert O. 2009. 'Big Questions in the Study of World Politics.' In Christian Reus-Smit and Duncan Snidal, eds. *The Oxford Handbook of International Relations*, 708–15. Oxford University Press.
- King, Gary, Robert O. Keohane and Sidney Verba. 1994. *Designing Social Inquiry: Scientific Inference in Qualitative Research*. Princeton University Press.
- Kocher, Matthew A., and Nuno P. Monteiro. 2016. 'Lines of Demarcation: Causation, Design-Based Inference, and Historical Research.' *Perspectives on Politics* 14 (4): 952–75.
- Lebow, Richard Ned. 2014. *Constructing Cause in International Relations*. Cambridge University Press.
- Lessing, Benjamin. 2017. *Making Peace in Drug Wars: Crackdowns and Cartels in Latin America*. Cambridge University Press.
- Levy, Jack S. 1997a. 'Prospect Theory, Rational Choice, and International Relations.' *International Studies Quarterly* 41 (1): 87–112.
- Levy, Jack S. 1997b. 'Too Important to Leave to the Other: History and Political Science in the Study of International Relations.' *International Security* 22 (1): 22–33.
- Maoz, Zeev, and Bruce Russett. 1993. 'Normative and Structural Causes of Democratic Peace, 1946–1986.' *American Political Science Review* 87 (3): 624–38.
- Marshall, Monty G., and Keith Jagers. 2002. 'Polity IV Project: Political Regime Characteristics and Transitions, 1800–2002.' Unpublished manuscript and Codebook, University of Maryland.
- McCloskey, Donald. 1985. 'Economical Writing.' *Economic Inquiry* 23 (2): 187–222.
- Moravcsik, Andrew. 1997. 'Taking Preferences Seriously: A Liberal Theory of International Politics.' *International Organization* 51 (4): 513–53.
- Munari, Bruno. 1981. *Da Cosa Nasce Cosa*. Il Mulino.
- Oneal, John R., and Bruce Russett. 1999. 'Assessing the Liberal Peace with Alternative Specifications: Trade Still Reduces Conflict.' *Journal of Peace Research* 36 (4): 423–42.
- Paris, Roland. 2001. 'Human Security: Paradigm Shift or Hot Air?' *International Security* 26 (2): 87–102.
- Pentland, Alex. 2014. *Social physics: How good ideas spread-the lessons from a new science*. Penguin Audio.
- Petersen, Roger D. 2002. *Understanding Ethnic Violence: Fear, Hatred, and Resentment in Twentieth-Century Eastern Europe*. Cambridge University Press.
- Polo, Sara M. T., and Kristian Skrede Gleditsch. 2016. 'Twisting Arms and Sending Messages: Terrorist Tactics in Civil War.' *Journal of Peace Research* 53 (6): 815–29.
- Prorok, Alyssa K. 2016. 'Leader Incentives and Civil War Outcomes.' *American Journal of Political Science* 60 (1): 70–84.
- Putnam, Robert D. 1988. 'Diplomacy and Domestic Politics: The Logic of Two-Level Games.' *International Organization* 42 (3): 427–60.
- Roessler, Philip. 2016. *Ethnic Politics and State Power in Africa: The Logic of the Coup-Civil War Trap*. Cambridge University Press.

- Rosato, Sebastian. 2003. 'The Flawed Logic of Democratic Peace Theory.' *American Political Science Review* 97 (4): 585–602.
- Russell, Bertrand. 2009. *The Philosophy of Logical Atomism*. Routledge.
- Russett, Bruce, and John Oneal. 2001. *Triangulating Peace: Democracy, Interdependence, and International Organizations*. Norton.
- Sambanis, Nicholas. 2004. 'What Is Civil War? Conceptual and Empirical Complexities of an Operational Definition.' *Journal of Conflict Resolution* 48 (6): 814–58.
- Samii, Cyrus. 2016. 'Causal Empiricism in Quantitative Research.' *The Journal of Politics* 78 (3): 941–55.
- Sanín, Francisco Gutiérrez, and Elisabeth Jean Wood. 2014. 'Ideology in Civil War Instrumental Adoption and Beyond.' *Journal of Peace Research* 51 (2): 213–26.
- Sarkees, Meredith Reid, and Phil Schafer. 2000. 'The Correlates of War Data on War: An Update to 1997.' *Conflict Management and Peace Science* 18 (1): 123–44.
- Sartori, Giovanni. 1970. 'Concept Misformation in Comparative Politics.' *American Political Science Review* 64 (4): 1033–53.
- Schelling, Thomas C. 1960. *The Strategy of Conflict*. Harvard University Press.
- Singer, J. David. 1961. 'The Level-of-Analysis Problem in International Relations.' *World Politics* 14 (1): 77–92.
- Singer, J. David, and Melvin Small. 1994. 'Correlates of War Project: International and Civil War Data, 1816–1992 (ICPSR 9905).' Ann Arbor, MI: Inter-University Consortium for Political and Social Research.
- Steele, Abbey. 2009. 'Seeking Safety: Avoiding Displacement and Choosing Destinations in Civil Wars.' *Journal of Peace Research* 46 (3): 419–29.
- Sundberg, Ralph, and Erik Melander. 2013. 'Introducing the UCDP Georeferenced Event Dataset.' *Journal of Peace Research* 50 (4): 523–32.
- Tierney, Michael J., Daniel L. Nielson, Darren G. Hawkins, J. Timmons Roberts, Michael G. Findley, Ryan M. Powers, Bradley Parks, Sven E. Wilson and Robert L. Hicks. 2011. 'More Dollars than Sense: Refining Our Knowledge of Development Finance Using AidData.' *World Development* 39 (11): 1891–1906.
- Tilly, Charles. 2001. 'Mechanisms in Political Processes.' *Annual Review of Political Science* 4: 21–41.
- Tilly, Charles, and Robert E. Goodin. 2006. 'It Depends.' In Robert E. Goodin and Charles Tilly, eds. *The Oxford Handbook of Contextual Political Analysis*, 3–32. Oxford University Press.
- Tomz, Michael R., and Jessica L. P. Weeks. 2013. 'Public Opinion and the Democratic Peace.' *American Political Science Review* 107 (4): 849–65.
- Van Evera, Stephen. 1997. *Guide to Methods for Students of Political Science*. Cornell University Press.
- Vogt, Manuel, Nils-Christian Bormann, Seraina Rügger, Lars-Erik Cederman, Philipp Hunziker and Luc Girardin. 2015. 'Integrating Data on Ethnicity, Geography, and Conflict: The Ethnic Power Relations Data Set Family.' *Journal of Conflict Resolution* 59 (7): 1327–42.
- Waltz, Kenneth N. 1959. *Man, the State and War*. Columbia University Press.
- Weeks, Jessica L. P. 2014. *Dictators at War and Peace*. Cornell University Press.
- Wendt, Alexander. 1987. 'The Agent-Structure Problem in International Relations Theory.' *International Organization* 41 (3): 335–70.
- Wendt, Alexander. 1999. *Social Theory of International Politics*. Cambridge University Press.
- Zinnes, Dina A. 1980. 'Three Puzzles in Search of a Researcher: Presidential Address.' *International Studies Quarterly* 24 (3): 315–42.



# The Simple, the Trivial and the Insightful: Field Dispatches from a Formal Theorist

Branislav L. Slantchev<sup>1</sup>

This advice begins with a disclaimer. I have a view of formal modeling that – while being shared by many fellow modelers and philosophers of science – is at variance with what seems to be the prevalent one in the discipline. All my thoughts on the use of formal models in research are bound up with that view, and are probably not useful if one does not share it.

Contrary to popular opinion, the biggest hurdle to effective modeling is not the absence of advanced mathematical skills. Instead, the problem lies with a hazy conception – shared by both proponents and critics – of what models are supposed to accomplish. The received view in the discipline seems to be that the primary purpose of (formal) models is to make predictions that are then ‘tested’ empirically.<sup>2</sup> On this account, models are hypothesis-generating machines – insert assumptions, crank through solutions, spit out predictions – and the benefit of formalism is to make the process more rigorous, and thus more scientific.<sup>3</sup> Since ‘research practice

in political science currently revolves around theory testing’, the sole value of the model is nearly always taken to lie in its ability to withstand empirical scrutiny.<sup>4</sup>

This view is incoherent in logic and unimplementable in practice.<sup>5</sup> It offers an impoverished interpretation of the role of models in research by denying outright their *raison d’être*: efficient and effective communication. Overcoming the fundamental hurdle to modeling requires one to recognize that models are merely specific arguments.

## MODELS ARE ARGUMENTS

A model is not evaluated if its predictions are not analyzed, regardless of how true the assumptions of the model are believed to be.<sup>6</sup>

Models are closed deductive systems, which simply means that their conclusions follow from their premises. Given these premises, the conclusions are true irrespective of

empirical referents. Models cannot be incorrect when their inferential rules are followed. No amount of ‘testing’, no matter how carefully designed, can alter this basic fact. The notion that hypothesis-testing somehow confers validity on a model or that it constitutes the core element in scientific practice has been roundly debunked.<sup>7</sup>

Moreover, all premises in these deductive systems are almost invariably false, in the sense that they do not correspond exactly to anything in the real world. This is true of any argument that purports to explain any social or natural phenomenon, not just models, and certainly not just the formal ones. For example, one might criticize a formal model for having the ‘unrealistic’ assumption that uncertainty over a parameter is represented by, say, the uniform distribution. But the non-formal argument that relies on the concept of uncertainty is not ‘more realistic’ because it does not make that assumption. It is exactly the opposite: without being specific about the concept, it might be impossible to evaluate as an argument. The virtue of the modeling exercise is that it can establish that the claim holds for the uniform distribution, which in turn could be used to establish whether it holds for a class of distributions or even arbitrary ones. Being non-specific and vague does not make an argument more general or ‘realistic’. If one is going to critique a model because of ‘unrealistic’ assumptions, one might as well give up any attempt to explain anything.

The proof of the model is not in its empirical consummation.<sup>8</sup>

If one thinks of models as arguments, then it quickly becomes clear what their role in research must be: give specific expression to a line of thinking, communicate it effectively and persuade the audience of its usefulness. It is the ability to perform these tasks well that defines a good formal model, so let us unpack them a bit.

First, formal models force one’s argument to be *specific*. That is, in order to represent abstract concepts with the mathematical

formalism of a model, one is invariably forced to define more or less precisely a specific representation of that abstract concept. For instance, we often use the rather abstract concept of ‘power’ in our research. In international relations, the term is ubiquitous, and yet it is never quite clear what it means. There are hundreds of articles about what ‘power’ could mean, and how it is to be understood in different contexts, and it is nearly impossible to evaluate research that deploys this concept (theoretically or empirically) because its meaning is protean. And yet we feel quite confident that something called ‘power’ is important in understanding international relations, not the least because policy makers seem so concerned about it.

When one writes down a formal model that wishes to use ‘power’, the abstract must become concrete. In the standard models of crisis bargaining, for example, ‘power’ is related to an actor’s expected payoff from war. It could remain undifferentiated – the higher the payoff, the more powerful an actor is – or become even more specific, relating to particular other concepts. For instance, it could be modeled as relating to the *costs* of war (the lower one’s costs, the more powerful one is), to the *probability of victory* (the more likely one is to win the war, the more powerful one is) or to the *valuation of the stakes* (the higher the stakes, the more powerful one is). Each of these specific formalizations of ‘power’ is subtly different from the others even if they all affect the expected payoff from war. For example, the *costs* only determine an actor’s own war payoff and nothing else, whereas the *probability of victory* determines the war payoffs of both actors (if one becomes more likely to prevail, the other one must be less likely to do so). *Valuation*, on the other hand, still relates to one of the actors, but determines both its war and peace payoffs.

At first blush, it appears to matter little how power is formalized: some of the most basic results in the crisis bargaining literature have been derived with each specification.

However, research soon showed that the specific formalization does, in fact, matter for many important conclusions. Uncertainty about the opponent's 'power' is a very common ingredient in theories of war, but once we have specified a precise formal definition of power, we are forced to adopt a precise formal definition of uncertainty as well. Thus, one could distinguish among uncertainty about the costs (independent private value), the probability of victory (interdependent value and uncorrelated types) and the undifferentiated value of war (correlated types). Fey and Ramsay have shown in a very general framework that the different sources of uncertainty have very different implications for the probability of war and the possibility of war-avoiding mutually acceptable settlements in crisis bargaining.<sup>9</sup> In other words, it matters very much what specific conceptualization of the abstract notion of 'power' one uses for one's argument.

Thus, contrary to the oft-repeated allegation that models are 'too abstract', models are in fact quite specific. They force us to give a particular expression to our line of thinking. The resulting clarity reduces the definitional burden of arguments and enables sharper communication.

Second, formalization lays bare the *structure* of the argument, which ensures its internal validity and simplifies communication. Consider the canonical crisis bargaining model used by Fearon to identify the risk–return trade-off as an important cause of war.<sup>10</sup> Amid its myriad of simplifying premises, the model assumes that one of the actors can make a take-it-or-leave-it demand, whose acceptance ends the game peacefully but whose rejection leads to war. Ultimatum games are quite popular in both economics and political science not because they are particularly representative of real-world situations but because modeling a bargaining situation opens a Pandora's box of additional assumptions about time horizons, discounting, inside and outside options, sequencing of moves, timing of offers and responses and

many others. (These usually remain rather buried in non-formal negotiation models where they are often tacitly assumed.)

One problem is that strategic bargaining models (ones where the analyst specifies the structure of the interaction; e.g., only one actor makes demands or the two alternate) that incorporate uncertainty tend to produce infinitely many solutions, as opposed to the ultimatum game, which usually yields sharp unique results. Because of that, the results of Fearon's analysis had to be tentative because it was not known just how dependent they were on that assumption. Powell extended the model to an alternating-offers bargaining framework and showed that, unlike the common models of that type, his also yielded a unique result that mirrored exactly the risk–return trade-off identified by Fearon's ultimatum game.<sup>11</sup>

This appeared to give us a stronger warrant to accept the argument until Leventoglu and Tarar showed that both the result's uniqueness and its risk–return trade-off aspect were dependent on another structural assumption: a proposer whose offer gets declined is not permitted to attack (only an actor responding to an offer could do so).<sup>12</sup> This premise is inconsistent with the standard assumption in international relations – that any actor can take military action whenever it chooses to do so – and relaxing it wipes out the canonical result. Their model shows that crisis bargaining can involve significant delays without escalating to war, and so the risk–return trade-off mechanism is less robust than previously thought. By varying only specific premises, scholars have been able to gain a much better understanding of how a particular argument works.

Third, formalization forces one to confront one's own demons of *unstated assumptions*. I have often heard the breezy dismissal of formal models with 'you can always concoct a model that yields any result you want', an assertion with the Schrödinger quality of being apparently both true and false at the same time.

It is true that anything can be described by some model when there are no restrictions to the premises one is permitted to make. It is also true that nothing in the modeling technology itself restricts the premises except perhaps to ensure that they are not mutually exclusive. The supposed (but famed) 'rigor' of models does not extend to the definition of the system itself or to the interpretation of its parameters and results.<sup>13</sup> Models can readily produce conclusions that are absurd (these are usually easy to spot) or trivial (because the premises assume them). This happens most often when people try to reverse-engineer an argument by working from the desired conclusion to the premises needed to generate it. So, the charge is correct, at least when it describes sloppy modeling practices.

The charge, however, is almost comically wrong when leveled against models designed from first principles and with careful attention to detail, as any formal theorist worth their salt would tell you. Constructing a valid argument that does not beg the question can be surprisingly difficult; the analysis can often be startling, and the entire process quite edifying. There is much to be learned from attempts to formalize one's intuition. There is a lot of trial and error involved in getting the argument right (meaning, to have a model that is both non-trivial and solvable), and often the conclusions are not the same as the ones the analyst expected to obtain. Models are great disciplinarians, and they can teach us when our intuition has gone astray or when our 'straightforward' argument turns out to require a small army of auxiliary premises to sustain.

My favorite example of the process in print (in addition to my own trials and tribulations) comes from Schelling.<sup>14</sup> He was dealing with the problem of surprise attack and reasoning from an intuition about an armed burglar surprised by an armed homeowner whose house he has broken into. Presumably, both prefer that the burglar just leave quietly, but the problem is that neither is sure whether the other might shoot. If he thinks that the

homeowner might shoot, the burglar becomes more likely to shoot first. But the homeowner knows that and now becomes even more fearful about getting shot, so he becomes more likely to shoot first. But the burglar also knows that his own fears are compounding the homeowner's, which makes him even more trigger-happy. This escalating spiral of mutual fears causes one of them to pull the trigger, ending the interaction with an outcome both would have preferred to avoid.

Schelling's intuition is compelling, but he decided to formalize it to see how such a 'multiplier effect' could arise. He then proceeds through a series of different formalizations, none of which yields the desired result. The failure in each iteration teaches him something about the problem that he had tacitly assumed without knowing how crucial it was for the inference he was making.

Even if his resolution leaves something to be desired from a modern standpoint, the exercise was clearly useful.<sup>15</sup> Not only did the author learn a lot about the subtleties of his intuition, but the exposition of the unsuccessful attempts to formalize it has great value for scholars who wish to build upon the insight. Knowing where the blind alleys are is crucial to progress.<sup>16</sup>

It is for these three main reasons that I almost invariably ask graduate students to formalize their arguments. Even the practice of constructing a model without solving it can be enormously beneficial, as it forces one to at least identify the basic premises and overall structure of its logic. Sometimes this is enough to expose a fatal weakness, as even rudimentary models can identify problematic lapses in reasoning.

Because models are correct by definition, specifying the premises and working systematically through their implications toward the conclusion has many benefits. We can tell whether a specific conceptual definition is consistent with a set of inferences, whether a collection of premises really does yield a claimed consequence and whether seemingly disparate arguments share a common



core. We can easily agree on what a definition means, on what the precise assumptions are and on how the results are obtained. This means we can build on each other's work more effectively and transmit that knowledge with a smaller chance of miscommunication.

None of these benefits are unique to formal models. People can make, and have made, arguments that are specific, well-constructed and non-trivial without formalizing them. What I am saying is that formalization facilitates the process for the analyst and democratizes it by making the argument instantly accessible to anyone with a modicum of training. To give an analogy, it was not formal musical notation that made Beethoven's Ninth possible. But it was this notation that allowed him to transmit what was in his mind efficiently, that has permitted generations of composers to build on his approach and that has enabled audiences to enjoy the results.<sup>17</sup>

Think of your model as an argument, and of yourself as a persuader.<sup>18</sup>

## WHEN TO MODEL, OR HOW TO GET INSPIRED

Although every argument can be modeled, formalization can be especially appropriate for certain kinds of arguments. Since published work almost never describes the inspiration behind the models and the evolution of the arguments before they end up in print, in this section I will give examples from my own experiences. In general, I resort to formal modeling when I have some intuition about a phenomenon of interest but find myself asking, 'How does this argument work?' Here are several instances of that question that produced formal models.

### *The Puzzling Case*

Logic, especially when human beings are involved, is often no more than a way to go wrong with confidence.<sup>19</sup>

I teach US foreign policy to undergraduates. One of the lectures is about the Korean War, and I never felt I understood why the US and China ended up in a war over North Korea. Political scientists who have tackled this question usually frame it as 'Why did the US miss clear warning signals from China and extend the war beyond the 38th parallel?' My reading of the history, however, indicated that the signals were anything but 'clear' – the Chinese chose an Indian intermediary dismissed by its own government as biased, they failed to intervene when it would have made most sense militarily (right after Inchon) and, perhaps most critically, they appeared to have failed to prepare for war.

The last of these was crucial: nobody will take you seriously if you threaten war but do nothing to get ready for it. Assured that the Chinese were bluffing, the Americans forged ahead, only to run into a vast mass of Chinese troops who swept them back south. Unbeknownst to anyone, China had entered North Korea with enough strength to shatter the UN advance and eventually stalemate the war despite reinforcements sent by the United States. If the Chinese had been serious about their threats, why not show that they were preparing to fight? This would have been a clear signal, but they chose not to make it, which in turn misled the United States into invading the north. Neither side wanted to fight the other, and yet war is what they got. Why?

This case was specially puzzling considering existing crisis bargaining theory. According to all our models, an actor who is willing to fight should make that preference known by some sort of costly diplomatic or military move. This is the only way to convey resolve and hopefully convince the opponent to make war-avoiding concessions. The Chinese behavior ran contrary to all these models, as they had deliberately concealed the one move that could have conceivably had a deterrent effect on the Americans. Why would they do this?

I was reading a science fiction novel by David Weber at the time, and there was a

scene in it where a very powerful military vessel on patrol detected a pirate ship in the distance. The captain's problem was that if she chased the pirates openly they would probably have the time to make their escape. So, she disguised the engine emission signature to mimic a merchant ship and lured the pirates into coming close enough for her to attack them. This was a fictional scenario, but one could think of any number of historical episodes involving such an ambush, so I wondered if the Chinese behavior had such an element to it.

The key was the possible self-denying aspect of revealing military strength: what if doing so gave the opponent an opportunity to do something that eroded your advantage? In the fictional story, the pirates would run away, thwarting the purpose of the patrol. In the Chinese case, the United States could use the information to target the troops with devastating effect because of its superior firepower and control of the skies (as it happened, Mao was still wrangling with Stalin over Soviet planes, since the Chinese air force was non-existent). Thus, if one believed that the encounter was likely to end in war, it could make perfect sense to conceal one's strength and gain the advantage of surprise.

This now presents a dilemma that the fictional captain did not have, but an actor engaged in crisis bargaining would: failure to reveal strength makes it more likely that the opponent would not offer significant concessions, which in turn would make war more likely. So, while feigning weakness could potentially be useful in war, it might not necessarily be optimal if it also made war more likely. This is the kind of problem – with incentives pointing in opposite directions – that is especially well suited for game-theoretic modeling.

I started with the canonical ultimatum crisis bargaining model and merely modified it so that after the initial demand was rejected, both actors chose how much to mobilize for the war, with their mobilizations being costly but also improving their chances of victory. Since a strong actor (one with low costs of

arming) would mobilize more aggressively, it could cause the other to respond with similarly aggressive mobilization. This, in turn, would weaken the incentive to reveal one's strength through the initial demand. The analysis revealed that this was indeed the case: there were circumstances in which a strong actor would not make a risky but revealing high demand but would instead pretend to be weak by mimicking the demand of the weak type. The cost of that was a lower peace payoff but the benefit was improved war payoff. The dilemma could produce behavior contrary to the canonical models but consistent with what the Chinese had done.

The analysis also showed that the assumption of two-sided incomplete information was unnecessary: it complicated the math without changing the basic insight, which could be obtained under asymmetric information about the actor making the demand. When I sent the revised version to the journal, a referee pointed out that the fundamental result might be had with an even simpler setup. He/she was right, and the published version centers on a much simpler model where only the informed responding actor can mobilize additional resources at a fixed cost.<sup>20</sup>

The model also yielded several surprises. First, there was the unexpected application: Jeff Ritter pointed out to me after a presentation that the mechanism could explain the puzzle of secret defensive alliances – they enhance one's fighting potential but deny one any deterrent advantages, the precise dilemma that could end in a feint. Second, there was the implication about the venerable explanation for war as being caused by mutual optimism. If an actor holds optimistic beliefs about the expected payoff from war, then it would be loath to make concessions, which in turn would make war more likely because it would weaken the opponent's incentive for peace. The crisis signaling literature had suggested that this optimism could be reduced by costly signals of resolve. One's apparent will to fight should lead the opponent to revise their estimate about war,

and if they are still willing to fight, then this in turn should lead the original actor to lower their estimate as well. The feigning weakness argument showed that a strong actor could deliberately foster false optimism in the opponent, which short-circuits the inferences: this actor could not use the opponent's willingness to fight as evidence to lower their own estimates about war since that willingness is based on the wrong inference induced by the feint. This would strengthen mutual optimism and make war more likely than the signaling literature would suggest.

Thus, my inability to offer a coherent explanation of a historical case to my undergraduates, coupled with the coincidental reading of a science fiction novel, led to novel insights about crisis bargaining.

I find history both fascinating and often puzzling. It offers a tremendous menu of opportunities to not understand something and is a fertile source of inspiration for research. I teach an entire graduate course on the history of international relations that is designed to puzzle students and generate ideas for study. It is important to realize that historical episodes (and empirical patterns) can only be puzzling considering existing explanations: one must know enough to understand that one does not understand something.

Being puzzled productively requires quite a bit of preparation, which is why it is often difficult for scholars tackling unfamiliar areas of research. Sometimes one's puzzlement is due to ignorance of existing work and is easily resolved by a literature review. This is why I am not a big fan of advice that tells scholars to avoid the literature review until after they have clarified their ideas. In fact, searching the literature with a particular idea in mind can be especially productive and efficient.

### ***The Inconsistent Assumption***

Varian contends that academic journals 'really aren't a very good source of original

ideas'.<sup>21</sup> I think what he means by that is that it is hard to get inspired by academic work to create something 'original'. I am using scare quotes because I never quite understood the emphasis on originality in the profession – sometimes there is a very good reason an idea did not appear in print before you had it, and it is not because you are a genius. Be that as it may, here is an example motivated by published research.

The study of crisis bargaining and escalation has relied on stylized models neatly summarized and analyzed by Fearon.<sup>22</sup> He distinguishes between behaviors that involve sunk costs (paid irrespective of the outcome) and those that involve audience costs (paid only if an actor backs down after making a threat). As an example of the former, Fearon gives 'building arms or mobilizing troops'. He is very careful to note that these actions 'may affect the state's expected value for fighting' (p. 70) and that it would be 'more realistic to have the probability that the defender wins in a conflict depend on  $m$  [the level of mobilization]' (p. 72). He opts not to do this in order to keep sunk costs and tying hands analytically distinct.

I wondered whether this distinction was distorting. It was very difficult to conceive of a military move that did not alter the distribution of power between the actors. This meant that any such move would alter not only the incentives of the actor making it but also those of its opponent: tying one's hands might simultaneously untie the opponent's, affecting the overall probability of war in the crisis in ways not anticipated by the original models.

To check this intuition, I modified that standard escalation game so that both actors could choose their mobilization levels during the crisis, with the probability of victory dependent on these allocations.<sup>23</sup> Sure enough, there were important differences in the inferences. For example, in contrast to Fearon's results, the military threat model (MTM) shows that bluffing is possible. It also undermined the long-established result

in the crisis bargaining literature that militarily stronger actors are likely to obtain better deals but must run a higher risk of war in order to do so.<sup>24</sup> In the MTM, these actors still get better deals, but their risk of war might be *lower* because they can compel the other side by a more aggressive mobilization (the untying hands effect). The MTM also showcased the importance of considering the costs of maintaining peace through mutual deterrence with high military allocations, which led me to another paper.

### ***The Tacit Assumption***

An essential, but often overlooked, assumption in the canonical crisis bargaining model is that peace is costless.<sup>25</sup> The high mobilizations without war in the MTM alerted me that this assumption might be problematic. I should have known this from Powell's earlier contribution that showed how the possibility of an armed peace depended on the long-term costs of deterrence, but since the point had been peripheral to the goals of his article, I had missed both it and its implications for crisis bargaining.<sup>26</sup>

Clearly, if peace were to be costlier than war, then the bargaining puzzle would fall apart: actors would fight because war was preferable to any negotiated outcome for at least one of them. This would shift the focus from the now trivial problem of war under these circumstances to the unexamined problem of how actors would create these circumstances in the first place. In other words, if they knew that making peace too costly would lock them into war, would they pursue strategies that do so anyway?<sup>27</sup>

I took the direct approach: since arming is costly irrespective of the outcome, I decided to look at how actors paid for military power. Since I had been reading a lot about the emergence of centralized government in early modern Europe, my head was full of examples of kings who could not tax very effectively but who borrowed a lot and sometimes

failed to pay back their debts. War finance through borrowing instead of taxation also seemed appropriate because funds would be instantly accessible (unlike tax proceeds), and this appeared important considering the fact that the vast majority of interstate wars only last for a few months.<sup>28</sup>

As usual, I started with the canonical model, allowed both actors to determine the distribution of power through their military mobilizations and merely allowed one of the actors to borrow to increase the resource base, thereby permitting larger military allocations. I assumed that the actor was committed to repaying the debt if the bargaining ended peacefully or if the war ended in victory, but that the debt was repudiated if the war ended in defeat.<sup>29</sup> The analysis then revealed conditions under which the actor would incur a debt so high that the other would not be willing to concede enough to enable its peaceful repayment, and as a result the interaction would end in war under complete information. Further analysis revealed the importance of the actor's efficiency in converting financial resources into military capability, a topic never explored formally before.

The review process beefed up the article substantially, since the referees wanted me to allow both sides to borrow and wanted me to account for interest on the debt that was somehow related to the risk of default. These analyses took several months to complete and showed that the fundamental insight was not dependent on these simplifying assumptions in the original. And so the article conveyed a new, and different, argument about the causes of war.<sup>30</sup>

### ***The Unsatisfying Argument***

Among the most fertile sources of inspiration are attempts to explain your arguments to others. Teaching students, discussing with colleagues and sometimes even just chatting with friends and family have all, at one point or another, stopped me dead in my tracks in

the sudden realization that my argument does not quite work, either because there seem to be missing steps or because it is making potentially distorting assumptions that themselves need to be explained. Being puzzled on one's own by reading is much harder than being stumped by someone's question. There is probably no limit to the inanity of ideas I can come up with if I work in isolation, and the healthy skepticism of others is a crucial corrective. That is why I advise students to talk about their ideas as much as they can, to anyone who will listen. Instead of becoming defensive about criticism, look at it as an opportunity to develop a better argument.

My first example is a model that came about from an offhand comment during a lunch break while I was still in graduate school. I was working through the literature on audience costs for a course assignment and was chatting with a faculty member (I cannot recall whom) in the lounge while waiting for the microwave to warm up my lunch. He had asked me what I was working on, and I was explaining the idea behind audience costs – that leaders who escalate a crisis are punished if they back down – when he interrupted me by asking: ‘why would they do that?’ As I was giving the usual ‘national honor and prestige’ answer, I began to realize that it involved an uncomfortable amount of hand-waving and that a good explanation would require these costs to arise endogenously in the model. In other words, there should be a reason for the audiences to be willing to impose costs on leaders for backing down. At this point, the microwave pinged, and the conversation shifted to something else. But the question bugged me.

The problem was that it was not actually at all clear to me why someone would punish a leader for avoiding war, especially if bluffing was an optimal strategy. I did not do much with this because I had to finish my dissertation, but a few years later I was discussing the importance of leaders with Hein Goemans and suddenly recalled the puzzle. We had a back-and-forth about this, and I searched the

(very small) formal literature on the topic only to find a couple of scholars asking the same question. I was not satisfied with the answers, so set out to model the problem myself.

My model stripped away all detail – like the presence of a foreign actor – that did not seem pertinent to the analysis of the interaction between a leader, a policy being implemented and the domestic audience.<sup>31</sup> I found a model developed by Dur to deal with the persistence of bad policies, and adapted it for my purpose, reasoning that a legitimate reason for punishing a leader (and thus imposing audience costs) would be the audience realizing that the policy implemented is bad and so preferring a more competent leader.<sup>32</sup> The nuance of the argument was that the imposition of costs had to happen with positive probability during the interaction, rather than being assumed as a hypothetical threat with the leader then taking action to avoid it. (If the imposition of costs remained a zero-probability event in the model, then the purported explanation of audience costs would amount to an assumption.)

As I developed my intuition, I realized that since the argument hinged on the asymmetry of information about the policy quality between the leader and the audience, the model might be useful in exploring other potential sources of relevant information, such as a political opposition and a possibly biased media.<sup>33</sup> Consequently, the model expanded to include these actors (along with the entailing auxiliary structural premises), and generated some surprising insights. Among them was the result that in the absence of a robust and unbiased free press, democracies were no more likely to generate audience costs for their leaders than autocracies. This was in contradiction to Fearon's working hypothesis in the original paper that claimed the opposite. Moreover, the argument helped explain why mixed regimes could be especially sensitive to policy failures.

My second example is a model that came about from a discussion with Christina

Schneider, who had been researching decision-making in international organizations (IOs). We were dog-walking and sharing what we knew about the literature on the topic when she mentioned the problem of compliance with IO decisions. When would actors abide by collective decisions without an exogenous system of enforcement? This quickly led to another question: how would actors agree on such collective decisions? Since most organizations involve voting, the answer seemed easy. But then we realized that the insights about voting come from models that assume that the outcome is enforceable and voters who disagree do not get to work to overturn decisions. Moreover, almost all such models assume that voting is sincere, which is not an issue in the setting they were developed in (secret votes) but that could be quite problematic where votes are public, as they are in many IOs. Indeed, many empirical studies implicitly rely on public votes being sincere when they use them to measure preference similarity of member states.

Thus, we ended up with a question: what makes voting in IOs meaningful in the sense that actors are likely to reveal their preferences with their vote and abide by the collective decision even if they disagree with it? We began formalizing the problem that very evening, and several months later had an answer.<sup>34</sup> The modeling exercise here also led us to some more fundamental issues such as conceptualizing of international cooperation not merely in terms of the free-rider problem, as the widespread reliance on repeated games with Prisoner's Dilemma preferences does, but also in terms of a conflict of interest between groups of like-minded actors with resources to pursue divergent policies. The introduction of this competitive element in the cooperation problem brought the original puzzle into sharper focus but also opened up a host of related issues for research.

In both examples, the impetus behind the model arose from the feeling that the existing arguments were not quite right because they relied on a premise that was itself in

need of explanation. This premise might have appeared for modeling convenience or because of the adoption of a model developed for an apparently related but in fact quite different context. In neither case was the empirical validity of the assumption relevant.

### ***Some Bad Ideas***

As I indicated above, thinking about published research could give one insight that begs to be formalized. For this to work, however, one needs to know enough about the substantive phenomenon being studied to understand which premises in the existing argument might be distorting the conclusion, and thus warrant change. Without this 'inductive' step, modifying existing models for the sake of 'relaxing assumptions' and 'making them more general' could turn out to be of interest only to a handful of modelers, or could prove a pointless exercise altogether. Do not formalize merely for the sake of formalization.

One unfortunate consequence of the misguided 'scientism' in the social sciences is the insistence that hypotheses be derived from a formal model. Somehow, this is supposed to imbue them with rigor and validity, never mind the fact that almost all such models are concocted after the fact and are absurdly trivial. I have argued at length about the supposed 'rigor' of formal models elsewhere,<sup>35</sup> so here I will limit myself to the following injunction: do not formalize just to give a formal gloss of your hypothesis (or to pretend that you are getting some precise point estimates).

As I will argue below, a good first step in building a model is to use an existing one and modify it as little as possible to adapt it to the problem under consideration. Thus, someone interested in crisis bargaining might start by looking at models of strikes or pre-trial negotiations: in all these settings the actors possess a power to hurt each other in the absence of an agreement, and in some there is also

uncertainty about who is going to prevail in a costly ‘fight’ if negotiations fail.

This sort of importation must be done very carefully, though, because there might still be very important differences between the contexts, and if one borrows the idea without accounting for them, the result could be worse than useless: it might in fact be harmful to subsequent research. Political scientists are especially prone to borrowing models from economists without due consideration of the different context or the fact that the economists themselves very often neglect crucial political considerations in their own models. It is very easy to end up with a very ornamental model and pages of equations that amount to no insight whatsoever. Sometimes all you need is a small change in a premise to adapt the model, but sometimes you might as well build your model from scratch. Do not attempt to shoehorn your intuition into an existing framework.

### HAVE PUZZLE, WILL MODEL

Most of my research time is spent thinking about puzzles and trying to make persuasive arguments that resolve them. When I decide to formalize something, I also tend to spend considerable time on what the model should look like. Understand that this is almost never the unidirectional process suggested by how the idea appears in print: idea → model → solution → interpretation. Instead, one should expect to make several attempts to formalize the idea, sometimes along with full or partial solutions to the models, and sometimes one might even have to modify the original puzzle in light of what the analysis uncovers. The process looks like this: simplify → model → solve → realize either model or question were not quite right → change appropriate premise → iterate. This goes on until one is satisfied about the match between the question and the model.<sup>36</sup> It is worth keeping a record of the failed attempts

to avoid having to repeat going down blind alleys and to assist with the write up and responses to referees (who often suggest things one has tried already).

When building the model, be aware that there is no guide that can tell you whether the model is going to be useful. Do not strive for ‘realism’. Keep it simple, and realize that all models, even the most ‘realistic’, are false. Utterly and irredeemably so. Strive instead for minimalism and elegance. Remember, you are going to be making an argument, so it pays to be clear, precise and concise, and to have as few moving parts as possible to convey your intuition. Resist the temptation to show off modeling chops with complex math. Any real mathematician is going to laugh at such a folly anyway. Unnecessary bells and whistles do not make the model ‘more realistic’; instead, they make it harder to solve and even harder to follow. Proliferating parameters and premises decreases the warrant to believe the robustness of the result: how can one be sure that it is not entirely dependent on one of these myriad assumptions or some unlikely combination of them?

Scholars are often tempted to make things unnecessarily complex (which is why impenetrable jargon is always *de rigueur* in seminars), mistaking incomprehensibility for profundity. The same is true for formal models: one can easily build a formidable-looking thicket of Greek letters and numbers without realizing that no flower of an idea survives in the forest of equations. Sometimes the authors themselves do not understand how these models work – having plowed to a solution in a mechanical fashion – let alone the audience, who might be duly impressed by the math fireworks but leave scratching their heads and quickly forgetting about the idea. One might hope that diligent scholars would pore over one’s brilliant but obfuscated work to tease out the intellectual gems concealed amid the baroque ornamentation, but most of us are neither Plato nor Aristotle, so what are the chances of that? Most likely, the work will perish in the scholarly wilderness

for lack of citation sustenance. If you want your ideas to make a difference, you need the audience to understand them, which means it is your responsibility to make them as clear as possible. Models should be as complex as they have to be, but no more.<sup>37</sup>

How can this be done? For starters, relating the model to something known can be helpful. This is why I talked about beginning with an existing model, especially one that the audience is likely to be familiar with. Building on previous work is not merely good scholarly practice (acknowledging the contributions of others should be the *sine qua non* of research), but also a valuable aid in fleshing out the argument – how similar is it to others, and what makes it different. It also helps communicate the ideas efficiently and effectively – the audience is more likely to grasp something that is not too far from what it already knows, and since it can then evaluate your argument better, it is more likely to be persuaded of its merits.

If an existing model is not readily adaptable, then you must build one from scratch. Begin by specifying who the actors are, what they want, what they think they know, what their constraints are and what they can do. There will be many tough choices to make here, and it is not at all obvious initially which are better. Should you limit yourself to two actors? If the interaction is dynamic, should it be just two periods or some sort of infinite-horizon game? If there is incomplete information, should you use two types or a continuum? If the latter, should you use an arbitrary distribution or something convenient analytically, like the uniform or normal distributions? Is the action space discrete or continuous? How many opportunities to act will the actors have, and in what order will they move?

The problem here is that when you are building a model from scratch, there is always the temptation to make the argument more general and the model more ‘realistic’. You might not know what assumptions are going to make it intractable and what assumptions

might make it trivial. So, there will be trial and error here. Build a model and try solving it. This will give you some intuition about its moving parts and how they interact. It will also give you some ideas about how to improve the model.

I cannot stress enough how important the process of building a satisfactory model is, how messy and iterative it can be, and how long it might take. Published papers make it look like the model sprang from the mind of the author fully formed and perfectly adapted to the task, like some sort of mathematical Athena from the head of Zeus. In my experience, nothing could be further from the truth. Coming up with a model that is an adequate representation of your argument is hard work, full of trial and error, over a very uncertain timeline. It has on occasion taken me months of trying different specifications and sets of premises to arrive at a model that is a reasonable approximation of my intuition while simultaneously being solvable.

This is where minimalism and elegance become crucial. By minimalism I mean trying to come up with the simplest non-trivial model you can think of that formalizes your intuition. Your goal is to expose the structure of your argument as cleanly as possible, so any unnecessary parameter or detail should be mercilessly pruned.

I have taken to writing a paragraph in my papers, right before the model specification section, where I enumerate the essential features the model should have in order to represent the puzzle the argument is going to address. This paragraph comes very naturally after the literature review, which has situated the puzzle in the relevant literature and shown why current research has not answered the question being posed. This review identifies gaps in existing arguments and points to premises that need to be incorporated into or omitted from the new argument. This paragraph also ‘sells’ the model to the audience by justifying its premises explicitly. Anything that cannot be justified in this paragraph should be removed from the model.



Elegance is an elusive concept. It is something that one recognizes when one sees it, but that cannot be defined very precisely. Minimalism certainly helps, but there is more to it. Does the model seem ‘natural’ for the question being posed (this is where that paragraph also helps)? Or does it incorporate some odd premises that artificially constrain the actors in their choices? All assumptions are false, but some are beyond silly and are likely to make your argument unacceptable despite its deductive rigor.<sup>38</sup> Always remember that the strength of your argument will rest not on the idea that your agents are optimizing but on what you have them optimizing over: their preferences and constraints.<sup>39</sup> Do the payoffs reflect reasonable preferences or do they appear contrived and complicated? Anything with more than a few parameters or with very specific functional forms begins to look suspicious to me. Is the notation intuitive or clunky? There is no standard notation in game theory, so here it is best to adopt the notation used in well-cited articles or textbooks such as *Game Theory* by Fudenberg and Tirole.<sup>40</sup> Here it is best to emulate the specification of prominent and well-cited models that you admire.

The build → solve → build again → solve again → build again → etc. sequence suggests that one could benefit from being smart about the analysis. Instead of trying the most general case, go through a few simpler variants first. Use numerical examples to get a handle on what might be possible in the model, and some intuition about the parameter space where it can happen. Simulations and graphs are an excellent way of exploring the model before you begin solving it analytically. Plot the payoffs and vary the strategies of the other players to see what form the best responses might take. You might notice a pattern. For example, some relationship between payoffs from two choices seems to persist no matter what values you assign to some parameter. Try to prove analytically that the optimal choice is independent of that parameter. You might see abrupt changes in the optimal solution. Try to

prove that it changes form at some threshold value. Once you derive the best responses, program them and then explore the comparative statics. You should use whatever programs you are comfortable with.<sup>41</sup>

When you start discovering analytical results, it is time to write them down in your draft paper. This is where the first lemmas and propositions will make an appearance. There might not be a lot connecting them yet, but the skeleton of the argument is being constructed as you learn from your model. I also like to typeset them in LaTeX immediately because the math looks beautiful, the proofs are easier to read and the text is readily useful. (It is also easier to make global notational changes.) I also write explanations of the intuition behind these results as if I am talking to an audience unfamiliar with the model. These often make it in some form into the final draft and are particularly useful to keep the argument running in my head.<sup>42</sup> Do not wait until the analysis is complete to write – write as you go along. You will end up with multiple drafts of various models and partial solutions as the record of your research endeavors, and you will have the basic draft of the formal exposition ready when you complete your analysis.

Throughout all of this, you must be ready to be taught by the model. Or, rather, you must understand the intuition behind the results you are getting, and you must be willing to either jettison the model if it does not represent your argument properly, or accept that your original intuition might have been incorrect or incomplete. In the end, learning from your model is the largest payoff from formalizing your argument.

## EXPOSITION: LEARNING FROM YOUR MODEL AND TELLING OTHERS

Now that the analysis is done and you have your main results, what next? If you followed the advice to write the intuition behind each

step in your argument, you have an excellent handle on how it works. Your goal now is to convey this to others and to persuade them.

Your final draft will not track either your research progress or the complexity of the argument itself. Instead, the paper should focus on how your results answer the question you posed (which may have been restated several times as your thinking has evolved while solving various models), and it should convey that connection efficiently and intuitively. The paper must lead the audience to your conclusions, not rely on it making its own inferences. This means exposing all necessary steps in the argument without getting bogged down in technical detail. It might be painful to relegate 50 pages of hard-won mathematical results to an appendix very few will ever read, but this is what you must do. The body of the paper should include just enough mathematical detail to carry the argument in plain prose.

If you cannot explain the behavior of your agents without reference to equations, you have a problem. You are telling a story, which means that your agents should have intuitive (given incentives and constraints) behaviors. Nobody will care about uninterpreted statements that refer to impenetrable mathematical conditions, no matter how correct they are. Nothing should remain ‘counterintuitive’ after you have presented your argument. Presenting the paper to colleagues is a very effective way of fleshing out the rough spots in the write up. What is obvious to you might be totally opaque to others. What you think is trivial might be crucially important to others. It is very difficult to put yourself in the position of an audience that has not spent any time on your research, so do not do it. Instead, go with a real audience for that. I never miss an opportunity to present ideas and have never turned down an invitation to do so, especially if the audience is not academic.

Word count limits prevent me from spending more time on advice about crafting the paper. Fortunately, there are excellent books about how to write elegantly and concisely.

Some of them are even specific to formal work.<sup>43</sup> Follow their advice. Read widely and emulate writers you admire. Like any skill, writing is made perfect with practice. And do not default to the dry, pedantic and, frankly, boring tone characteristic of academic papers.

Remember, your model is an argument, and persuasion hinges on how it is presented, on rhetoric. Strive for readable prose. Do not be afraid to be slightly imprecise when the alternative is a detour into technical detail. Use historical cases to illustrate your points (but do not pretend that they are some sort of ‘tests’ of your results). It is fine to be entertaining. It is you who must provide the justification and the interpretation of the model. It is you who must explain the argument. It is you who must hold your audience’s attention and persuade it. Leaving any of these steps to others is a guarantee that your modeling efforts will be for naught. And we would not want that, now, would we?

## Notes

- 1 I thank David Wiens for many useful discussions and William Clark for his comments on this chapter.
- 2 Jim Granato and Frank Scioli, ‘Puzzles, proverbs, and omega matrices: the scientific and social significance of Empirical Implications of Theoretical Models (EITM)’, *Perspectives on Politics*, 2, 2 (2004), pp. 313–23, p. 315; Rebecca B. Morton, *Methods & Models: A Guide to the Empirical Analysis of Formal Models in Political Science* (Cambridge University Press, Cambridge, 1999); Donald P. Green and Ian Shapiro, *Pathologies of Rational Choice Theory: A Critique of Applications in Political Science* (Yale University Press, New Haven, 1994). See James Johnson, ‘Models among the political theorists’, *American Journal of Political Science*, 58, 3 (2014), pp. 547–60 on what he calls ‘the standard rationale’ for models in political science.
- 3 Nancy Cartwright, *The Dappled World: A Study of the Boundaries of Science* (Cambridge University Press, Cambridge, 1999), p. 184.
- 4 Kevin A. Clarke and David M. Primo, *A Model Discipline: Political Science and the Logic of Representations* (Oxford University Press, Oxford,

- 2012), p. 11; Johnson, 'Models among the political theorists', p. 80.
- 5 James Johnson, 'Models-as-Fables: An Alternative to the Standard Rationale for Using Formal Models in Political Science' (Department of Political Science, University of Rochester, NY, 2017).
  - 6 Morton, *Methods & Models*, p. 102.
  - 7 Clarke and Primo, *A Model Discipline*; Nancy Cartwright, 'Models: parables v. fables', in R. Frigg and M.C. Hunter (eds), *Beyond Mimesis and Convention: Representation in Art and Science* (Springer, Dordrecht, 2010), pp. 19–32; J. Johnson, 'How not to criticize rational choice theory: pathologies of "common sense"', *Philosophy of the Social Sciences*, 26 (1996), pp. 77–91.
  - 8 One might wish to keep in mind the distinction between models (which is what I am discussing here) and a theory (which I will not discuss). Think of models as the building blocks of a theory: useful (or not) specifications of concepts and ideas that comprise a theoretical explanation. Theories can, and should be, subjected to empirical testing in a way that models cannot, and should not, be.
  - 9 Mark Fey and Kristopher W. Ramsay, 'Uncertainty and incentives in crisis bargaining: game-free analysis of international conflict', *American Journal of Political Science*, 55, 1 (2010), pp. 149–69.
  - 10 James Fearon, 'Rationalist explanations for war', *International Organization*, 49, Summer (1995), pp. 379–414. The trade-off arises because actors are unsure how much they must concede to each other to make peace mutually preferable to war. Conceding a lot makes the other actor more likely to accept a settlement but also forces one to forego some of the benefits of peace. Conceding too little, on the other hand, risks pushing the opponent into a fight. Generally, an actor that is uncertain about what the other one's minimal acceptable terms are would make an offer that carries some risk of rejection but that yields somewhat better terms if accepted.
  - 11 Robert Powell, 'Bargaining in the shadow of power', *Games and Economic Behavior*, 15, 2 (1996), pp. 255–89.
  - 12 Bahar Leventoglu and Ahmer Tarar, 'Does private information lead to delay or war in crisis bargaining?', *International Studies Quarterly*, 52, 3 (2008), pp. 533–53.
  - 13 Branislav Slantchev, 'On the proper use of game-theoretic models in conflict studies', *Peace Economics, Peace Science and Public Policy*, 23, 4 (2017), pp. 1–14.
  - 14 Thomas C. Schelling, *The Strategy of Conflict* (Harvard University Press, Cambridge, MA, 1980), ch. 9.
  - 15 For a modern treatment of this problem, see Sandeep Baliga and Thomas Sjöström, 'Arms races and negotiations', *Review of Economic Studies*, 71, 2 (2004), pp. 351–69.
  - 16 In this respect, the scholarly practice of publishing only what 'works' is quite a detriment to knowledge accumulation. Much can be learned from failed attempts to formalize an argument, and I wish people would be more upfront about the fact that the published model is probably the fifth iteration attempted.
  - 17 Ironically, Beethoven's markings with respect to dynamics and tempo were often misunderstood, and to this day there is controversy about how his work must be performed. This is nothing that more formalism could not have fixed.
  - 18 Deirdre N. McCloskey, *The Rhetoric of Economics* (2nd ed.) (Madison: The University of Wisconsin Press, 1998).
  - 19 David Weber, *At All Costs* (Baen, Riverdale, 2005), p. 664.
  - 20 The article preserves some of the original analysis as an extension to analyze endogenous tactical incentives.
  - 21 Hal R. Varian, 'How to build an economic model in your spare time', *American Economist*, 41, 2 (1997), pp. 3–10, p. 3.
  - 22 James D. Fearon, 'Signaling foreign policy interests: tying hands versus sinking costs', *Journal of Conflict Resolution*, 41, 1 (1997), pp. 68–90.
  - 23 Branislav L. Slantchev, 'Military coercion in interstate crises', *American Political Science Review*, 99, 4 (2005), pp. 533–47.
  - 24 Jeffrey S. Banks, 'Equilibrium behavior in crisis bargaining games', *American Journal of Political Science*, 34, 3 (1990), pp. 599–614.
  - 25 Strictly speaking, all that is necessary for the canonical results is for the status quo to be less costly than war. This is usually modeled as the negotiation outcome being costless.
  - 26 Robert Powell, 'Guns, butter, and anarchy', *American Political Science Review*, 87, 1 (1993), pp. 115–32.
  - 27 This line of reasoning was independently pursued by Andrew Coe, 'Costly Peace: A New Rationalist Explanation for War', Manuscript, Department of Political Science, Vanderbilt University, 2011, who identified three other causes of peace being potentially costlier than war.
  - 28 Branislav L. Slantchev, 'How initiators end their wars: the duration of warfare and the terms of peace', *American Journal of Political Science*, 48, 4 (2004), pp. 813–29.
  - 29 Technically, all that is needed for the results is that default is more likely after defeat. This makes debt service less costly in expectation if a war is

- fought than if peace obtains, which is enough to trigger war under some conditions.
- 30 Branislav L. Slantchev, 'Borrowed power: debt finance and the resort to arms', *American Political Science Review*, 106, 4 (2012), pp. 787–809.
  - 31 Branislav L. Slantchev, 'Politicians, the media, and domestic audience costs', *International Studies Quarterly*, 50, 2 (2006), pp. 445–77.
  - 32 Robert A. J. Dur, 'Why do policy makers stick to inefficient decisions?', *Public Choice*, 107, 3/4 (2001), pp. 221–34.
  - 33 Kenneth A. Schultz, 'Domestic opposition and signaling in international crises', *American Political Science Review*, 92, 4 (1998), pp. 829–44; Matthew Baum, 'Going private: public opinion, presidential rhetoric, and the domestic politics of audience costs in U.S. foreign policy crises', *Journal of Conflict Resolution*, 48, 5 (2004), pp. 603–31.
  - 34 Christina J. Schneider and Branislav L. Slantchev, 'Abiding by the vote: between-groups conflict in international collective action', *International Organization*, 67, 4 (2013), pp. 759–96.
  - 35 Branislav L. Slantchev, 'On the proper use of game-theoretic models in conflict studies', *Peace Economics, Peace Science and Public Policy*, 23, 4 (2017), pp. 1–14.
  - 36 This reality also makes a mockery of the idealized notion that one constructs a deductive model in some pristine analytical void and then 'tests' its conclusions. The process I am describing is a mix of induction and deduction, all informed by the researcher's knowledge of empirical patterns, other models, and cases.
  - 37 For a discussion of what model complexity is and isn't, see Catherine C. Langlois, 'Are complex game models empirically relevant?', *Conflict Management and Peace Science*, 35, 1 (2017), pp. 1–15. She also critiques Allan and Dupont's assertion of a 'tendency toward theoretical elegance to the detriment of empirical correctness' that supposedly plagues formal models. I leave it to the reader to decide what to make of an assertion that uses 'empirical correctness' as a criterion to evaluate models.
  - 38 An example from economics would be the notion that unemployment represents the workers' free choice of leisure without a job over working at previous pay, as the real business cycle theory would have one believe. See Robert Lucas and Leonard A. Rapping, 'Real wages, employment, and inflation', *Journal of Political Economy*, 77, 5 (1969), pp. 721–54.
  - 39 Robert Solow, 'The state of macroeconomics', *Journal of Economic Perspectives*, 22, 1 (2008), pp. 243–9.
  - 40 Drew Fudenberg and Jean Tirole, *Game Theory* (The MIT Press, Cambridge, MA, 1991).
  - 41 I use *Mathematica* for the analytical derivations (I prefer doing them by hand and checking the result), and I use *Gauss* for simulations. The only reason for these choices is that these are the programs I learned in graduate school. The numerical exploration approach is advocated in Catherine Langlois and Jean-Pierre Langlois, 'From numerical considerations to theoretical solutions: rational design of a debtor-creditor agreement', *Peace Economics, Peace Science and Public Policy*, 22, 4 (2016), pp. 403–12.
  - 42 Keeping notes for yourself is not as strange as it sounds. If you stop working on a model for a few weeks, by the time you come back you will have forgotten all the details. At this point, I usually have to re-derive everything from scratch, so it really helps to keep extensive notes and not rely on memory for anything. This is a practice I inherited from my days as a computer programmer, when staring at my own code a couple of months after I wrote it taught me to document it as if I was addressing a partially lobotomized monkey.
  - 43 I have found two very useful for general prose: Helen Sword, *Stylish Academic Writing* (Harvard University Press, Cambridge, MA, 2012); Francis-Noël Thomas and Mark Turner, *Clear and Simple as the Truth: Writing Classic Prose* (2nd ed.) (Princeton University Press, Princeton, 2011). For formal writing, one is indispensable: William Thomson, *A Guide for the Young Economist: Writing and Speaking Effectively about Economics* (The MIT Press, Cambridge, MA, 2001).



# Evidence-Driven Computational Modeling<sup>1</sup>

Ravi Bhavnani, Karsten Donnay  
and Mirko Reul

## AGENT-BASED COMPUTATIONAL MODELS

Computational modeling is a powerful, versatile tool for the analysis of complex social phenomena. Historically, scholars used computational modeling to investigate abstract causal relationships in artificial settings, highlighting simple but counter-intuitive dynamics. Seminal examples include work by Thomas Schelling on the drivers of segregation (Schelling, 1971), Robert Axelrod on the evolution of cooperation (Axelrod, 1984), Joshua Epstein and Robert Axtell on artificial societies (Epstein and Axtell, 1996), and Palmer et al. on artificial stock markets (Palmer et al., 1999). These early applications influenced subsequent research, including notable studies on the formation and dissolution of nation-states after the end of the Cold War (Cederman, 1997), the dynamics of ethnic violence and genocide (Bhavnani and Backer, 2000), and more recently civil violence in Baghdad and Jerusalem

(Bhavnani et al., 2014; Weidmann and Salehyan, 2013).

In contrast to *consolidative models*, which typically involve the development of ‘model’ systems to represent ‘real-world’ settings with measurable physical characteristics (for weather forecasts, see Gneiting and Raftery, 2005; Skamarock and Klemp, 2008), *exploratory* computational models stop short of formalizing the complexity of social systems (Bankes, 1993). Given the difficulty of fully observing, theorizing and validating processes in social and natural systems, our approach builds on work that is exploratory, not consolidative, in nature. One class of exploratory computational models used in the social sciences is agent-based computational modeling (ABM) (for an overview, see de Marchi and Page, 2014).

A key property of ABM is the specification of simple rules from which complex outcomes *emerge*. As such, an ABM may be specified as a non-linear function that relates combinations of inputs and parameters to

outcomes. ABMs are typically composed of agents, decision-making heuristics, an interaction topology and a non-agent environment (Epstein, 1999). Agents in an ABM can represent individuals (Bhavnani et al., 2008; Epstein, 2002), groups (Bhavnani et al., 2009; Kollman et al., 1992) or institutions (Cederman, 1997), to name a few possibilities. In this regard, the approach provides a high degree of flexibility or *granularity*, given the ability to integrate phenomena specified at different scales. ABMs are process-oriented and lend themselves well to studying dynamics, in contrast to approaches that tend to be more equilibrium-centered.

In most formulations of ABM, agents are endowed with a range of characteristics and decision-making heuristics. Individual agents may learn or adapt their behavior based on their own experiences, driven by heuristics or imitation, or change may be effected for a population of agents by means of evolutionary selection (Kollman et al., 1992; Laver, 2005; Mitchell, 1996). The interaction topology specifies how agents interact with each other and their environment, the latter being composed of physical features such as geography or topography (Axtell et al., 2002; Epstein, 2002) or various states of the world (Axelrod, 1984; Nowak and May, 1992; Tullock and Campbell, 1970). These elements constitute the key components of an ABM, which is run repeatedly to identify causal mechanisms, observe relationships, patterns and emergent outcomes, and explore counterfactual scenarios.

ABMs lend themselves well to the analysis of complex social phenomenon, in particular where ostensibly simple decisions have unexpected consequences (Epstein, 1999). Yet, while agent-based models have notable strengths, they are not immune to criticism (Richiardi et al., 2006). A notable weakness of ABM is the tendency to include too many factors and interactions, given the ease with which these may be specified. As a rule of thumb, a model becomes too complicated when comprehensive exploration

of the comparative statistics for each model parameter is infeasible (see Lustick et al., 2004). Under these circumstances, it is virtually impossible to determine what is driving model results. Yet another flaw is the lack of relevant theoretical and empirical anchors, which result in unrealistic or even arbitrary model specifications. These anchors are essential to address the *identification problem* – the notion that multiple plausible mechanisms may explain a given outcome (Fisher, 1966). This chapter provides practical advice for designing, implementing and using computational models that are evidence-driven and designed to address these shortcomings.

## EVIDENCE-DRIVEN COMPUTATIONAL MODELS

The evidence-driven modeling (EDM) framework rests on three methodological pillars: agent-based modeling (ABM), contextualization using geographical information systems (GIS) and empirical validation. EDM harnesses the strengths of ABM in capturing both social complexity (e.g., the heterogeneity of actor beliefs, preferences, attitudes and behaviors as well as the characteristics of specific institutional settings and local environments) and causal complexity (including questions about who interacts with whom, when, where and with what effects), while simultaneously achieving a high degree of real-world correspondence and resonance. The combination places EDM squarely at the intersection of theory and empirics.

Notable examples of EDM include studies about civil violence in Jerusalem and Baghdad (Bhavnani et al., 2014; Weidmann and Salehyan, 2013), neo-patrimonial networks (Geller and Moss, 2008), social inequality in pastoralist societies (Rogers et al., 2015), the rise and fall of the Anasazi people in what is now the Southwestern United States

(Axtell et al., 2002), social capital and civic culture in Italy (Bhavnani, 2003), legislative politics (Laver et al., 2011), party competition (Laver and Sergenti, 2011) and the occurrence of burglary (Malleon and Birkin, 2012). The diversity of research demonstrates the utility of combining computational models with rich empirical data – data that are spatially and temporally disaggregated – to analyze the links between micro-level dynamics and emergent, macro-level outcomes.

Where other, more aggregate analyses yield inconclusive results, EDM enables researchers to adjudicate between alternative explanations and reveal complex, conditional relationships and inter-dependencies that would otherwise be difficult to detect (Camerer, 2003; Kim et al., 2010).<sup>2</sup> In particular, EDM makes it possible to specify causal mechanisms in ways that are sufficiently intricate, conditional and, thus, ultimately realistic, while maintaining the ability to go beyond purely exploratory modeling by means of rigorous empirical testing. And in contrast to experimental approaches that control for contagion and spillover, EDM explicitly incorporates these ostensible threats to validity as part of the causal chain, for example by endogenizing the effects of geographical proximity and the heterogeneity of covariates across spatial units. The more specific benefits of EDM include the following:

- 1 *Model Topography*: The ability of EDM to harness GIS, in conjunction with empirical data, enables realistic topographies to be substituted for the abstract grids characteristically used in ABM. As such, the landscapes used in EDM more closely represent actual physical or social inter-dependencies, capturing complex, often endogenous relationships among adjacent units, rather than controlling for these relationships statistically or by means of experimental design.
- 2 *Agent Granularity*: EDM can simultaneously accommodate data of different spatial and temporal resolution, whereas other methodologies are often wedded to the use of specific, fixed

units. In contrast to ABM, these different units of analysis correspond to empirical observations and capture dynamics at meaningfully interlinked levels, e.g., individual decision-makers interacting with groups.

- 3 *Data Imputation*: EDM is typically used in data-rich contexts but also excels in data-poor contexts, where information on relevant indicators suffers from incompleteness, a lack of synchronization, mismatched units of observation, and differing levels of detail. Imputation in EDM works by seeding a model with potentially sparse empirical data, and then permitting model dynamics to evolve endogenously. The closer simulated outcomes are to empirical trends, the better the imputation. The estimation of different parameters across contexts, using the same model, is one way to increase model robustness.
- 4 *Identification*: As with ABM, EDM can be used to explore relationships between or adjudicate among competing micro-level explanations, relying on methods for data construction, such as participant observation, expert or field interviews. Insights from these methods help ground a model, ensuring that researchers 'get the story right' and tailor the model to the specificities of a given context.
- 5 *Counterfactual Analysis*: Once a model is calibrated and empirically validated, counterfactuals can be devised by adjusting values of certain parameters, including those capturing micro-level dynamics and the empirical context, or by introducing new parameters. The results, produced under an assortment of 'what-if' scenarios, offer an indication of what the world could look like if empirically observed trends were to change. In essence, this option enables experimentation through simulation. Short of true out-of-sample forecasts, counterfactual experiments make it possible to undertake evidence-driven forecasting.

In the remainder of this chapter, a step-by-step discussion guides the reader through the use of EDM in the *Modelling Early Risk Indicators to Anticipate Malnutrition (MERIAM)* project. We provide further detail on the building blocks for EDM, as well as on the choices and practical challenges of using the approach. Our discussion is

intended to serve as a point of departure for conducting research with EDM.

## EVIDENCE-DRIVEN MODELING OF MALNUTRITION

The MERIAM project illustrates how the EDM approach can be applied, from initial conception to final, policy-relevant application. MERIAM is a four-year project funded by the UK government, which brings together an inter-disciplinary team of experts across four consortium partners: Action Against Hunger, the Graduate Institute of International and Development Studies, Johns Hopkins University and the University of Maryland. MERIAM's primary aim is to develop, test and scale up models to improve the prediction and monitoring of undernutrition in countries that experience frequent climate- and conflict-related shocks.<sup>3</sup>

In 2017, the number of undernourished people was estimated at 821 million; this is closely associated with the spread of armed conflict (FAO et al., 2018). Regions across Nigeria, South Sudan, Somalia and Yemen face severe food insecurity, related in no small measure to their exposure to conflict as well as a host of other characteristics that increase vulnerability to famine. The gravity of these situations and high interest among stakeholders serve as inspiration for devising effective means of forecasting risks to better anticipate crises and guide appropriate responses.<sup>4</sup>

The research team at the Graduate Institute is tasked with the development of an EDM to analyze the effect of household-level decisions on nutrition-related outcomes (e.g., acute malnutrition and resilience), accounting for variation in household characteristics; local, contextual factors; and more macro- or aggregate-level covariates. How households adapt their behavior, changing or diversifying sources of household income in response to stressors and shocks, may serve to improve

or worsen a household's resilience to food insecurity over time. We develop and validate our EDM based on the case of Karamoja, Uganda, and subsequently expand our model framework to other cases in sub-Saharan Africa.

We use this project – and, in particular, the application of our approach to study malnutrition in Karamoja – as an example of applying cutting edge computational modeling techniques to a highly relevant policy issue. The 'entry-point' for the use of EDM is an abundance of theoretical knowledge on the issue, the complexity of interactions of the numerous factors that influence malnutrition outcomes at the household level and the need for a systematic, reliable and transparent forecasting technique. At present, practitioners and policymakers tasked with anticipating changes in the risk of acute malnutrition need to combine expert knowledge on malnutrition, including a deep understanding of its causes in particular cases, with statistical analysis of data from various sources on a regular basis. A prominent example in the context of sub-Saharan Africa is the Famine Early Warning Systems Network (FEWS NET), which classifies a country's risk of acute food insecurity, relying on expert discussions and analysis mainly of remote-sensing, market price and trade data. The outcome is an indicator for food insecurity on a five-point scale, ranging from 'Minimal' to 'Famine', with 'near'-term and 'medium'-term forecasting windows of up to seven months (see *IPC 2.0*<sup>5</sup>).

There are several potential weaknesses inherent to such an approach. First, the link between data and projected food insecurity lacks formalization (the scenario-building process is interpretive) and transparency (it is unclear how a particular prediction is made). In a related vein, expert discussions underlying the data analysis are undocumented for end-users, making comparisons of forecasts by different experts problematic should interpretations vary.<sup>6</sup> And finally, predictions of food insecurity by livelihood zone obscure the



relative weight of various risk factors and their effects at more disaggregated spatial units.

Our EDM approach attempts to address these weaknesses. First, we harness available expertise on food security and malnutrition, relying explicitly on expert surveys, to develop a theoretically grounded computational model. Second, we use existing, household-level data and household surveys conducted in the field to empirically contextualize and validate the model for a set of sub-national regions that vary in terms of their incidence and prevalence rates, livelihood zones, climate conditions and history of conflict. Third, we provide a tool that stakeholders can use to construct acute malnutrition scenarios across diverse contexts, exploring the relation between shocks and stressors, on the one hand, and more immediate and long-term outcomes on the other. In the section that follows, we provide an in-depth example of the EDM approach, beginning with why we believe this is an appropriate methodological choice. We then provide an overview of the model development process, data construction, model implementation, refinement and validation. Lastly, we discuss how the validated EDM can be used for scenario-based analyses and how its results can be presented to expert users and relevant stakeholders.

## A STEP-BY-STEP GUIDE TO THE EDM APPROACH

The MERIAM project seeks to identify how, in response to conflict and climate shocks, household-level decisions affect nutrition-related outcomes – effectively unpacking the ‘black box’ of household behavior. At the household level, our EDM analysis is motivated by a set of fundamental questions that link household characteristics and behavior to acute malnutrition outcomes:

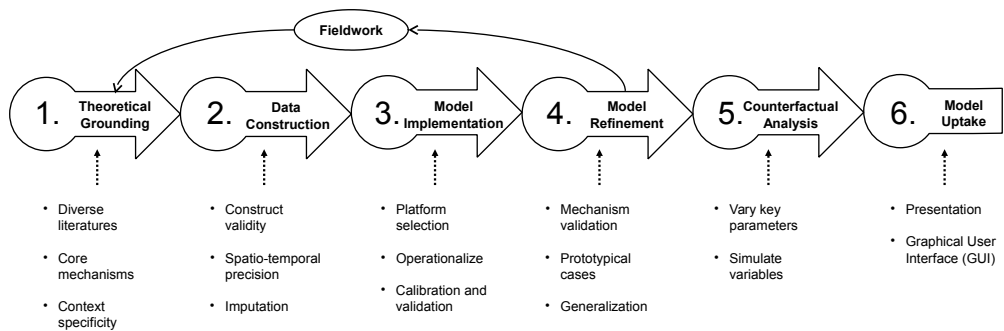
- Holding the context constant, why are some households affected by risk factors while others are not?

- To what extent do households within the same context react to risks in the same way?
- Is the same segment of a population recurrently affected, or is substantial flux observed?
- Do the risk factors for those affected remain the same year after year, or do they change over time?
- Does a given risk factor have the same effects across diverse contexts?

Our model uses *resilience*, the ability to cope with or adapt to various shocks and stressors, as a conceptual frame to investigate variation in nutritional outcomes in a manner that resonates with development stakeholders (e.g., Boukary et al., 2016; Food and Agriculture Organization of the United Nations, 2016; United States Agency for International Development, 2012; see Béné et al., 2015). In this particular domain, a resilient household – and in the aggregate, resilient communities, regions and countries – is better positioned to cope with the unfavorable effects of an exogenous shock, or has a greater ability to recover (say to pre-crisis intake levels of nutrition) in the aftermath of such a shock. Stakeholders may design interventions to boost endowments, moderate constraints, facilitate learning or strengthen systems for crisis management (Béné et al., 2015), all of which should contribute to greater resilience to nutritional crises.

The development of our EDM may be broken down into the six steps illustrated in Figure 4.1.

- 1 *Theoretical Grounding*: the EDM is grounded in existing theoretical and empirical knowledge of the subject matter.
- 2 *Data Construction*: relevant data to seed and validate the EDM are collected, analyzed and formatted.
- 3 *Model Implementation*: a preliminary version of the EDM is implemented as a computer simulation.
- 4 *Model Refinement and Cross-Case Validation*: the model is refined through expert interviews, fieldwork and out-of-sample testing.
- 5 *Counterfactual Analysis*: a valid EDM is extended by implementing counterfactual, ‘what-if’ experiments to explore how simulated trends are altered under different conditions.



**Figure 4.1 Model development process**

6 *Model Uptake*: suggestions to visualize and present EDM results are put forth in an effort to make the model accessible to relevant end-users and stakeholders.

In each of these steps, the researcher makes consequential decisions that affect the outcome of the EDM process. Yet by design, EDM makes these choices explicit and transparent. And while computer simulations are not as well understood among social scientists and policy makers, the basic intuition underpinning EDM is relatively simple, perhaps more so than a statistical model that addresses similar questions.

### **Theoretical Grounding**

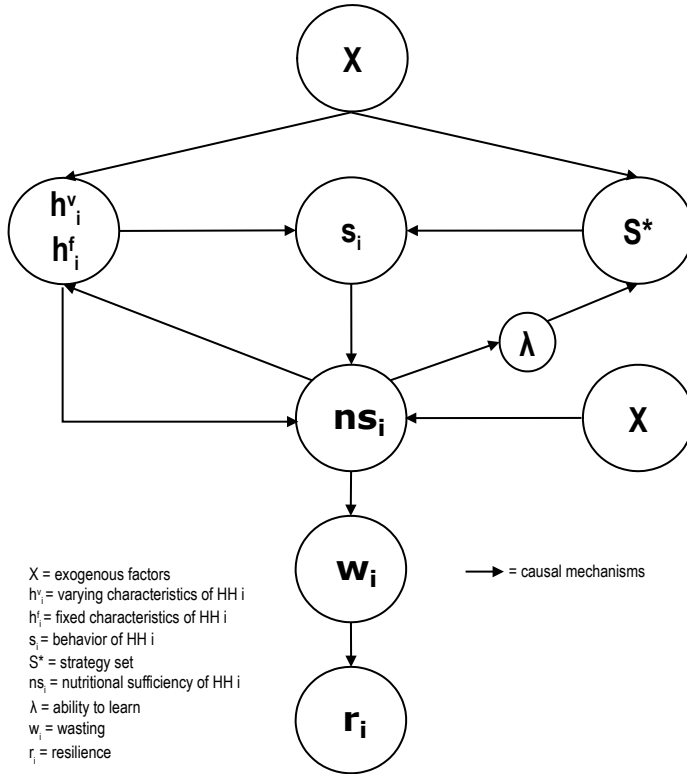
First, we surveyed a diverse body of research on malnutrition, including detailed qualitative case studies (e.g. Hatløy et al., 1998; Manners, 2014; Parker et al., 2009), comprehensive ‘broad-brush’ approaches that integrate a wide range of mechanisms to explain malnutrition (e.g., Young and Marshak, 2017) and statistical analyses (e.g., Ajieroh, 2009; Ehrhardt et al., 2006; Fotso, 2007). Like any computational approach, the internal validity of EDM depends, in no small measure, on prior, often qualitative work that describes social processes in their requisite complexity.

Second, we reviewed this work to map the relations between leading and underlying indicators, in an effort to identify the \*\*core

mechanisms\*\* that characterize malnutrition dynamics. The two defining categories into which these indicators fall are *shocks* and *stressors*. The first category of shocks includes the onset of a conflict, which usually has a sudden impact that is unanticipated by households. The second category of stressors accounts for the effects of longer-term or more gradual, recurring changes such as a lack of rainfall, which may vary in intensity, including its most extreme manifestation as drought.

Third, we examined *context-specific mechanisms*. In Karamoja, conflict has been endemic given the 20-year insurgency of the Lord’s Resistance Army (LRA), as well as more recent pastoralist conflicts that involved cattle raiding or rustling (DCAF-ISSAT, 2017; FEWS NET, 2005). This has had many negative effects on the Karamojong, including the loss of human lives, displacement, reduction in livestock, and the progressive spread of small arms used by herders for protection, indirectly contributing to an increase in violence (DCAF-ISSAT, 2017).

Our model evolved as we relaxed simplifying assumptions and improved our understanding of malnutrition in Karamoja. Moving beyond an initial specification of households endowed with an unbounded ability to adapt behavior, we defined households as boundedly rational actors (Arthur, 1994), focusing primarily on food provision. The latest version of the model is depicted in



**Figure 4.2 Flow diagram for MERIAM EDM**

Figure 4.2. In each iteration and for every household, we calculate nutrition levels (*ns*) based on previous actions, a set of health- and food-intake factors ( $h^{v,f}$ ), and exogenous constraints (*X*). Proportional to changes in *ns*, households adapt their behavior based on learning ( $\lambda$ ): if *ns* is stable and sufficient to feed household members, the household continues to behave the way it did before. But if radical changes occur, households adapt their behavior either (a) randomly, (b) by copying a locally optimal strategy from their neighbors, or (c) by combining existing strategies to create a new one (Holland, 1975; Kollman et al., 1992; Krakauer and Rockmore, 2015; Mitchell, 1996; Urbanowicz and Moore, 2009).

In the current formulation, sub-optimal behavior is the rule rather than the exception, and status-quo behavior is effectively reinforced if households attribute a worsening of

their situation to previous behavioral changes rather than exogenous factors (see ‘probe and adjust’ in Huttegger et al., 2014). The approach permits us to account for structural impediments to adaptation in a context like Karamoja, while still allowing for household-level change, e.g., from pastoral to agro-pastoral food production (e.g., Mercy Corps, 2016; see also Stites and Huisman, 2010). Note that the model, at this stage, may still be classified as an ABM. The next step is to construct the data necessary to enable the empirical contextualization of the model.

The process of theoretically grounded model development as is described here is prototypical, including the iterative refinement of model mechanisms and their operationalization as the modeler’s understanding develops. At this stage, choices are made based on the best available insights on the

case, bearing in mind the need for further refinement following fieldwork and empirical validation in the next stage.

### **Data Construction**

The EDM approach relies on data for empirical contextualization. The quality of the data – its accuracy, resolution and coverage – ultimately shapes our ability to seed and validate the model. A careful examination of data availability and quality is thus of paramount importance. In this section, we describe the main strengths and weaknesses of our household-level data for Karamoja, although our considerations may be generalized beyond the specifics of this case.

Our EDM is fundamentally about household characteristics and behavior, but it requires information exogenous to households as well, from district-level statistics about health facility capacities to more granular data at the grid or point level. Figure 4.3 presents an overview of the data we use to seed and validate the MERIAM EDM.

For our analysis of household characteristics and behavior in Karamoja, we utilize nutrition survey data provided by Action Against Hunger (2013). The dataset has two distinct advantages compared to other nutrition surveys. First, it contains behavioral variables at the household level. Second, the dataset is longitudinal: it consists of six survey rounds between August 2010 and May 2012, allowing us to validate and align the timescales of simulations against empirical outcomes repeatedly over this time period.

Our household-level data exhibits three principal weaknesses with respect to construct validity, spatio-temporal precision and completeness. First, some household-level variables do not measure the specific household attributes and behaviors we seek to model. For example, we use the variable ‘food source’ as a measure for how households obtain food. But only the ‘*most important*’ food source was measured in the survey,

which means that we cannot observe whether households use other means to obtain food.

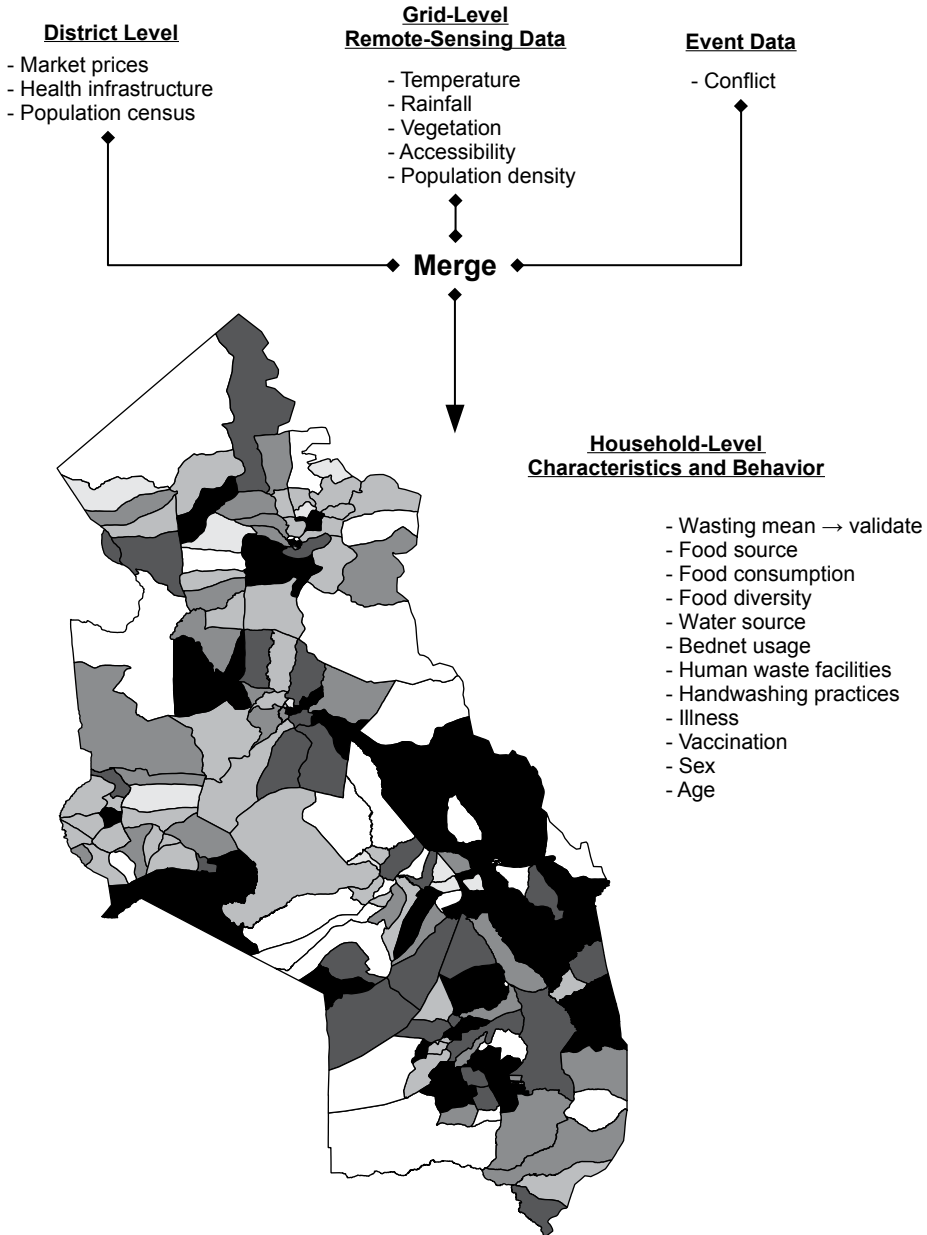
Second, the data are imprecise. They capture longitudinal trends, rather than following the same households over time. With panel data, we could seed and validate our EDM against the decisions and characteristics that each particular household makes over time. With trend data, this information exists at an aggregate level to the extent that we know, on average, households changed on the measured variables between samples.

In addition, our household data is spatially imprecise insofar as it is representative at the ADM1 district. Short of obtaining representative samples at lower levels of analysis, there are still ways to mitigate spatial imprecision using *imputation*. For Karamoja, census data is only available at sparse intervals – no census was conducted between 2002 and 2014, a period that saw a 2.4% population growth in Karamoja (Uganda Bureau of Statistics, 2017). An alternative is to use remote-sensing data to estimate population numbers at the grid level.

The challenges to data quality from construct validity, (lack of) precision and completeness recur across empirical settings and are by no means specific to the EDM approach. The ability to accommodate data at varying levels of granularity, though, is a strength of the approach. As such, we need not match empirical and simulated household characteristics at the same level of granularity, given that we measure other contextual factors related to conflict, climate and market prices (see Figure 4.3). Instead of joining data at the lowest common level granularity, each model component can be simultaneously specified/matched with empirical data at the maximal level of granularity permissible.

### **Model Implementation**

Several platforms are suitable for programming an EDM, with a trade-off between



**Figure 4.3 Data construction overview**

simplicity of use and performance, i.e., the ability to handle complex modeling setups. A popular and widely used solution is the NETLOGO platform (Wilensky, 1999), which is easy to learn but more restrictive in

its capabilities. At the other end of the spectrum are powerful libraries such as the Repast framework in Java (North et al., 2013) or MESA for Python (Masad and Kazil, 2015), both of which require greater customization.

The EDM discussed here uses a custom class-based implementation in Python. We note that the choice of programming language and framework do not affect model outcomes – only the ease of use and the ultimate runtime of the simulation.

In contrast, the *operationalization* of the model can have a profound impact. To operationalize model dynamics, it is necessary to specify the logical order of progression, sequence of actions and updates within the model – who does what, where and when and based on which information – paying close attention to attendant implications (Caron-Lormier et al., 2008). It follows that model heuristics can be universal or conditional, fixed or subject to change over time.

Consider, first, the problem of defining time progression within the model – the number of actions and updates that occur within a time step (e.g., an hour, day or year) to correctly reflect the timescales of the processes the model seeks to represent. A common solution is to make this correspondence explicit in the definition of a model time step, i.e., define time progression in terms of the fraction of possible actions or updates performed. For example, we consider a time step to have ended after updating the state of all households once. For the timescales of simulated and empirical outcomes to align, any time dependent parameters in the model that have empirical equivalents (e.g., the rate at which households adapt their strategies) have to be scaled such that their timescale aligns with that of the observed empirical process.

Second, EDM use geographical information for contextualization that ensures a high degree of correspondence between geography and the model topology. Exactly how space is operationalized within the model reflects an explicit choice to be made by the researcher. A common choice of implementation that reduces computational complexity is to discretize physical space. For example, in the model for Karamoja, household locations are defined on an underlying regular grid that is dynamically generated using

actual settlement locations and their associated densities. In order to account for both low and high population densities, we use data on population densities at the grid level such that the number of households in a grid approximates the population density in the corresponding area in Karamoja.

Choices related to model operationalization are by no means simple or straightforward. To avoid influencing simulation outcomes or unwittingly introducing errors and artifacts, competing operationalizations of the same model mechanisms should be tested to ensure that a specific operationalization is not driving simulation outcomes (see also Galán et al., 2009).

Analogous to testing in- and out-of-sample predictive power for statistical models, EDM are formally validated and calibrated to maximize the correspondence between simulation results and real-world outcomes. Figure 4.4 shows the full modeling cycle, from model operationalization and contextualization to enumeration and calibration. For the Karamoja case, we identify the degree to which households are able to adapt to changing conditions that, all else equal, best explain the observed patterns of malnutrition. The closer the calibrated model approximates empirical outcomes, the greater the validity of the model predictions. Yet, quantitative agreement is not the only important measure. The parameters that best predict empirical outcomes must also reflect plausible dynamics on the ground. Should this fail, further refinement and validation of the model are necessary.

The modeling cycle illustrated for Karamoja (Figure 4.4) serves as a template for identifying parameters that yield the closest correspondence to real-world outcomes, a process that constitutes the core of the evidence-driven approach: given a model specification that formalizes our theoretical understanding of a process and data to seed the model (possibly at varying levels of granularity), what is the model with maximal explanatory power for our outcome of

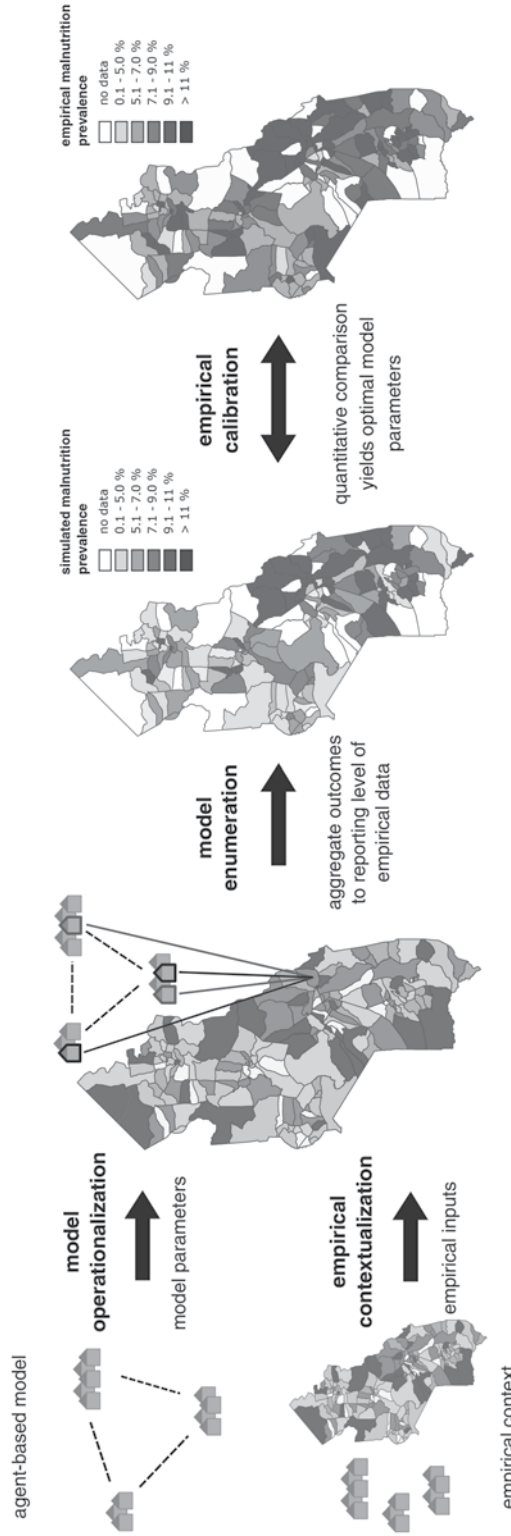


Figure 4.4 EDM modeling cycle

interest? The parameter constellation of the ‘best-fit’ model indicates which specific mechanisms in which constellation yields maximal agreement with empirical outcomes. Clearly, there is no guarantee that the constellation is empirically plausible, which makes the step of model refinement and validation an essential part of the process.

### ***Model Refinement and Cross-Case Validation***

While our model framework is derived from the existing literature, in addition to which we have undertaken a limited set of interviews with experts outside of the region, ground-truthing the model in Karamoja is an integral part of our EDM approach. Absent fieldwork, we run the risk of misrepresenting the dynamics at the core of our model, in particular, those that concern agent decision-making and behavior:

- Household decisions ( $ns, \lambda \rightarrow S^*$ ): Who are household decision-makers? To what extent are ‘households’ independent decision-makers? Can they make decisions in the way we envision them? Which household characteristics are adaptable? If household characteristics change, what is the time scale? If households learn new strategies, what strategies are learned and how?
- Nutritional sufficiency ( $ns, h^{iv}$ ): How do household decision-makers evaluate the nutritional intake of their children? What general knowledge/awareness can we assume? What factors explain why some households cope/adapt successfully, while others do not?
- Household and exogenous variables ( $h^{iv}, X$ ): Is the relative importance of household and systemic constraints in our model appropriately balanced? What amount of agency can we attribute to households?
- Resilience ( $r$ ): What types of resilience (coping vs. adaptation) can we expect in a context like Karamoja?

These and other questions pertain to core components of our model. They needed to be addressed with experts and government

officials working on the ground, and more importantly with Karamojong households whose nutritional situation we seek to understand.

The relationship between model development and field research can and should be treated as an iterative process. Multiple stages of model development can be interspersed with multiple rounds of field research. As such, the EDM approach is agnostic to the timing of fieldwork. A possible starting point is the development of a theoretically grounded model framework. A first round of field research can then be used to refine the model, identify relevant causal linkages and supply additional empirical input for the model. For MERIAM, we received some of our most valuable input through modeling exercises where we probed experts and households to make their ‘mental models’ explicit, through surveys and focus group discussions. Depending on one’s epistemological orientation and practical considerations, it is plausible to conceptualize a modeling framework inductively from preliminary field research.

Whereas ‘getting the story right’ in Karanoja was important, it is evident that an EDM initially developed and tested in one region need not be applicable to other contexts. Households may respond differently to the same exogenous shocks and stressors in ways that cannot be accounted for comprehensively in a single model specification. That said, re-building the entire model for each context analyzed is also unnecessary. Rather, a given model can be modified to incorporate contextual differences in a systematic manner, by identifying model *prototypes* that exhibit meaningful variation across key dimensions, and *validating and refining* the model for each prototype.

A necessary first step involves the identification of similar patterns across contexts, with respect to either (causal) drivers, mechanisms, behaviors or outcomes, as a means of building a set of computational model prototypes. For example, prototypes for malnutrition dynamics could be constructed to



account for similar patterns of incidence and prevalence in wasting across contexts. The selection of prototypical cases that exhibit consequential variation, both geographically and temporally, is essential to ensure that an EDM generalizes beyond the confines of a particular case. More specifically, a different case enables us to refine the model with a view toward maximizing external validity, while a similar case allows us to verify and strengthen the internal validity of the model in another context. For the MERIAM project, analyzing sub-national cases over finer-grained temporal units was preferable to selecting country-years as units of observation.

As a second step, case-specific grounding, data construction, and fieldwork are repeated for every model prototype to which the EDM will apply, building on prior work where possible, given that much of the data on exogenous factors is constructed uniformly across contexts (e.g., remote-sensing and conflict data). For our second prototypical case, West Pokot in Kenya, we assessed historical and cultural differences relative to Karamoja, as well as changes in salient causal mechanisms. Finally, we adapted the field survey used in the Karamoja case for the particularities of this context, making only the minimal changes required.

So while the classic trade-off between external and internal validity applies in no small measure, EDM can be systematically extended to produce valid results across contexts, while retaining internal validity for specific cases.

### ***Counterfactual Analysis***

Counterfactual analysis can be divided into two types of ‘what-if’ scenarios for EDM: those that relate to model parameters – determinants of model dynamics that have no direct empirical referent and were inferred from the model – and those that relate to model inputs specified by empirical data.

Counterfactual analysis for model parameters is equivalent to considering comparative

statistics. Here, one sets all model parameters to their optimal values, except for a particular parameter whose influence we seek to assess. For example, the effect of a household’s propensity to adapt could be analyzed, all the more tellingly if the effect is non-linear, i.e., if small changes in household behavior produce significant changes in nutrition outcomes.

Counterfactuals may also be conducted for model inputs. Instead of seeding all inputs with empirical data, one can use exogenously determined values for one or more inputs, treating them as equivalent to parameters. Examples include testing the impact of climatic and economic shocks, as well as the source, timing, location, type and scope of interventions (e.g., food imports, humanitarian assistance from international sources and education designed to shape household behavior) on household behavior and malnutrition. Stakeholders can then use these insights to understand the likely effects of different interventions under a variety of conditions across different contexts. While this type of counterfactual analysis is best constrained to the period for which the model was optimized – in other words, within sample – out-of-sample counterfactual analysis, including forecasting, is feasible when one clearly specifies how parameters and empirical inputs might change in the future.

It follows that EDM are well suited to providing data-driven, scenario-based analyses, with the caveat that underlying assumptions are transparently communicated. The EDM developed as part of MERIAM forms the basis for a tool to make scenario-based forecasts of malnutrition and explore the efficacy of various interventions in response to climate- or conflict-related shocks. A concern, then, is to develop an effective means to communicate the methodology beyond a purely academic or expert audience.

### ***Model Uptake***

The EDM approach requires a combination of technical knowledge and relevant domain

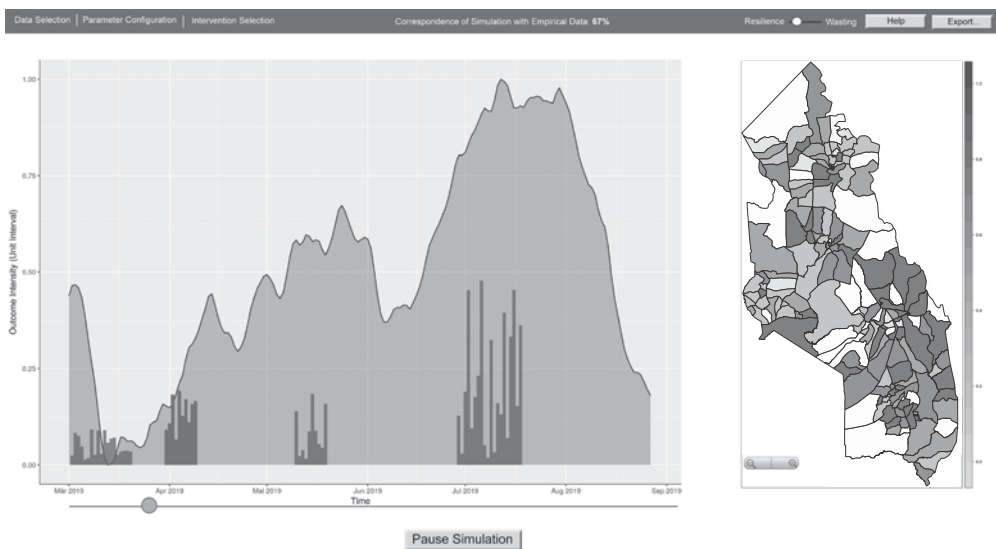
expertise. As such, the specification, contextualization, validation and refinement of an EDM is typically undertaken by academic researchers or trained experts. On the other hand, if done correctly, the kinds of scenario-based analysis for which EDM is suited may be of immediate interest to practitioners, subject experts or policy makers largely unfamiliar with the methodology. The question then concerns how the EDM methodology can be pitched to stakeholders beyond a mere visualization of results, to provide a ‘feel’ for the model, its specificity and generalizability, and its use as a tool for making evidence-driven policy decisions.

With an expert and practitioner audience in mind, we plan to complement an academic research paper with a policy brief written for a general, non-specialist audience. Rather than using technical jargon, a brief would explain the steps involved in model construction, much in the way that this chapter does, highlighting key policy-relevant insights. The brief would clearly communicate the limits and uncertainties associated with the EDM scenario-based

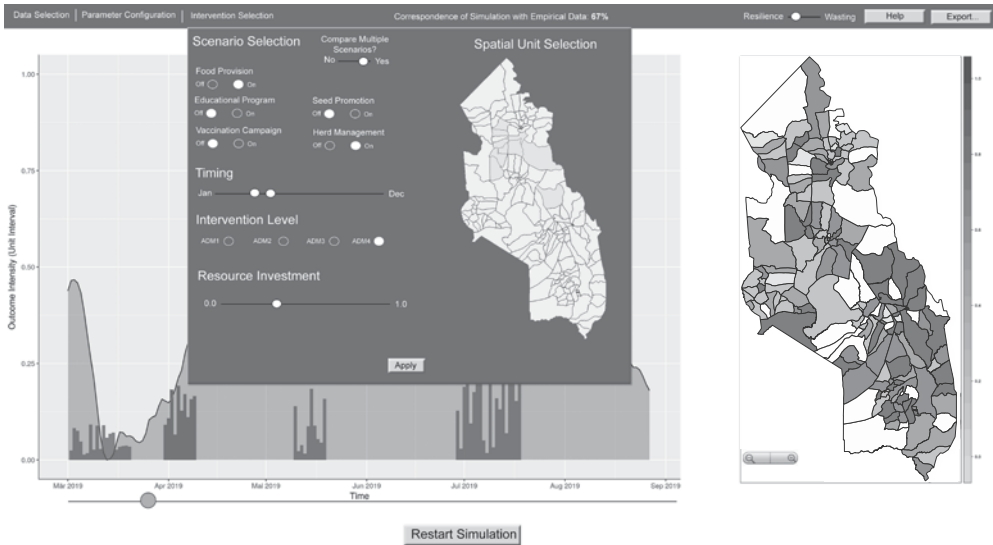
forecasts, considering the effect of specific interventions on malnutrition outcomes discussed above.

While a brief will certainly help communicate policy-relevant insights, it falls short of providing a true ‘feel’ for the EDM. The only way to achieve this is by developing a tool with an interactive graphical user interface (GUI). Such a tool would preserve the full complexity of the EDM while allowing non-expert users to easily engage with the model and translate output. The tool may also require an expert to revise a particular model specification for a new case, after which the GUI will perform its intended function.

Figures 4.5 and 4.6 depict a mock-up of the Simulating Acute Malnutrition Toolkit (SAMT). At the center of the GUI is a trend-based forecast and a map of the region being analyzed. Users can switch between outcomes (e.g., resilience or wasting prevalence) and select the time point for which outcomes on the map would be displayed. The heat map shows normalized levels of the selected outcome (here, resilience) at a fine-grained level of analysis for the selected time point, with the slider below the time



**Figure 4.5** Mock-up of SAMT (main window)



**Figure 4.6 Mock-up of SAMT (intervention configuration)**

series graph. Bars in the time series indicate the intensity of shocks and aid interventions, respectively. The degree to which the simulated results correspond to empirical data is specified at the top of the interface.

GUIs and software tools for decision support are generally starting to gain traction. These include a large array of domain-unspecific tools for data handling and visualization, as well as more specific policy support tools in diverse domains such as epidemiology (den Broeck et al., 2011) and public safety (Chooramum et al., 2016). Making EDM accessible to others requires this kind of explicit engagement with stakeholders, allowing them to develop a better intuition for the approach.

## CONCLUSION

Evidence-driven computational modeling effectively harnesses the strengths of ABM, while achieving a high degree of real-world correspondence and resonance. As our

discussion of the MERIAM project demonstrates, the EDM approach incorporates contextual knowledge and theoretical insight, captures complex spatio-temporal inter-dependencies, explicitly accounts for endogenous relationships, uses realistic topographies and harnesses data at varying levels of measurement. The combination places EDM at the intersection of theory and empirical work. For the MERIAM project, we harness the power of EDM to make scenario-based forecasts and undertake counterfactual analyses, developing a tool for policy makers tasked with addressing the high-stakes problem of malnutrition. The development of the MERIAM EDM has been elaborate, costly and time consuming, given that many of the standard elements of research design in political science – theory building, case selection, data collection and fieldwork – comprise the approach. We believe the contribution to evidence-driven decision-making is well worth the effort, and trust that the procedures and best practices outlined in this chapter will result in the development and use of EDM across diverse domains.

## Notes

- 1 The authors thank Alessandra Romani from the Graduate Institute, Geneva for helpful comments on draft versions of this chapter. This document is an output from a project funded by UK Aid from the UK Department for International Development (DFID). The views expressed do not necessarily reflect the UK government's official policies.
- 2 Take, for example, contact theory. Researchers have found empirical evidence to support the notion that increased inter-group contact leads both to higher and lower levels of violence. EDM have been used to explore the conditions that give rise to these divergent outcomes. For a detailed discussion, see Bhavnani et al. (2014).
- 3 Visit <https://www.actionagainsthunger.org/meriam> for more information on the project.
- 4 This description, drawn from the MERIAM project proposal, serves as the overarching motivation for the larger project as well as our specific contribution to the same.
- 5 <http://fews.net/IPC>.
- 6 See Samimi et al. (2012) for an earlier critique, and <http://fews.net/our-work/our-work/scenario-development> for details on the FEWS NET scenario-building process we seek to complement with our approach.

## REFERENCES

- Action Against Hunger. 2013. 'Nutrition Surveillance Data Analysis: Karamoja, Uganda, December 2009 – May 2012.' Technical Report, [https://www.actionagainsthunger.org/sites/default/files/publications/Nutrition\\_Surveillance\\_Data\\_Analysis\\_Uganda\\_08.2013.pdf](https://www.actionagainsthunger.org/sites/default/files/publications/Nutrition_Surveillance_Data_Analysis_Uganda_08.2013.pdf) (Retrieved July 15, 2019).
- Ajieroh, Victor. 2009. 'A Quantitative Analysis of Determinants of Child and Maternal Malnutrition in Nigeria.' *Nigeria Strategy Support Program* (NSSP10): 1–69.
- Arthur, W. Brian. 1994. *Increasing Returns and Path Dependence in the Economy*. Ann Arbor, MI: University of Michigan Press.
- Axelrod, Robert. 1984. *The Evolution of Cooperation*. New York: Basic Books.
- Axtell, R. L., J. M. Epstein, J. S. Dean, G. J. Gumerman, A. C. Swedlund, J. Harburger, S. Chakravarty, R. Hammond, J. Parker and M. Parker. 2002. 'Population Growth and Collapse in a Multiagent Model of the Kayenta Anasazi in Long House Valley.' *Proceedings of the National Academy of Sciences* 99(Supplement 3): 7275–7279.
- Bankes, Steve. 1993. 'Exploratory Modeling for Policy Analysis.' *Operations Research* 41(3): 435–449.
- Béné, Christophe, Derek Headey, Lawrence Haddad and Klaus von Grebmer. 2015. 'Is Resilience a Useful Concept in the Context of Food Security and Nutrition Programmes? Some Conceptual and Practical Considerations.' *Food Security* 8(1): 123–138.
- Bhavnani, Ravi. 2003. 'Adaptive Agents, Political Institutions and Civic Traditions in Modern Italy.' *Journal of Artificial Societies and Social Simulation* 6(4).
- Bhavnani, Ravi, and David Backer. 2000. 'Localized Ethnic Conflict and Genocide: Accounting for Differences in Rwanda and Burundi.' *Journal of Conflict Resolution* 44(3): 283–306.
- Bhavnani, Ravi, David Backer and Rick R. Riolo. 2008. 'Simulating Closed Regimes with Agent Based Models.' *Complexity* 14(1): 36–44.
- Bhavnani, Ravi, Karsten Donnay, Dan Miodownik, Maayan Mor and Dirk Helbing. 2014. 'Group Segregation and Urban Violence.' *American Journal of Political Science* 58(1): 226–245.
- Bhavnani, Ravi, Michael G. Findley and James H. Kuklinski. 2009. 'Rumor Dynamics in Ethnic Violence.' *The Journal of Politics* 71(3): 876–892.
- Boukary, Aboubakr Gambo, Adama Diaw and Tobias Wünscher. 2016. 'Factors Affecting Rural Households' Resilience to Food Insecurity in Niger.' *Sustainability* 8(3): 181.
- Camerer, Colin F. 2003. *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton, NJ: Princeton University Press.
- Caron-Lormier, Geoffrey, Roger W. Humphry, David A. Bohan, Cathy Hawes and Pernille Thorbek. 2008. 'Asynchronous and Synchronous Updating in Individual-Based Models.' *Ecological Modelling* 212(3–4): 522–527.
- Cederman, Lars-Erik. 1997. *Emergent Actors in World Politics: How States and Nations Develop and Dissolve*. Princeton, NJ: Princeton University Press.
- Chooramum, Nitish, Peter J. Lawrence and E. R. Galea. 2016. 'Urban Scale Evacuation Simulation Using buildingEXODUS.' *Interflam* 2: 1645–1656.

- DCAF-ISSAT. 2017. 'Cattle Rustling and Insecurity in Africa: A Comparative Perspective.' Technical Report, <https://issat.dcaf.ch/Learn/SSR-in-Practice2/Countries-Regions/Madagascar/Reinforcing-African-Union-SSR-Unit-support-to-national-SSR-processes-Madagascar> (Retrieved July 15, 2019).
- de Marchi, Scott, and Scott E. Page. 2014. 'Agent-Based Models.' *Annual Review of Political Science* 17(1): 1–20.
- den Broeck, Wouter Van, Corrado Gioannini, Bruno Gonçalves, Marco Quaggiotto, Vittoria Colizza and Alessandro Vespignani. 2011. 'Implementation Results and Discussion Conclusions Availability and Requirements Declarations References Software Open Access Open Peer Review the GLEaMviz Computational Tool, a Publicly Available Software to Explore Realistic Epidemic Spreading Scenarios at the Global Scale.' *BMC Infectious Diseases* 11: 37.
- Ehrhardt, Stephan, Gerd D. Burchard, Carsten Mantel, Jakob P. Cramer, Sarah Kaiser, Martina Kub, Rowland N. Otchwemah, Ulrich Bienzle and Frank P. Mockenhaupt. 2006. 'Malaria, Anemia, and Malnutrition in African Children – Defining Intervention Priorities.' *The Journal of Infectious Diseases* 194(1): 108–114.
- Epstein, Joshua M. 1999. 'Agent-Based Computational Models and Generative Social Science.' *Complexity* 4(5): 41–60.
- Epstein, J. M. 2002. 'Modeling Civil Violence: An Agent-Based Computational Approach.' *Proceedings of the National Academy of Sciences* 99(Supplement 3): 7243–7250.
- Epstein, Joshua M., and Robert Axtell. 1996. *Growing Artificial Societies: Social Science from the Bottom Up*. Washington D.C.: Brookings Institution Press.
- FAO, IFAD, UNICEF, WFP and WHO. 2018. 'The State of Food Security and Nutrition in the World 2018: Building Climate Resilience for Food Security and Nutrition.' Technical Report, <http://www.fao.org/3/I9553EN/i9553en.pdf> (Retrieved July 15, 2019).
- FEWS NET. 2005. 'Conflict Early Warning and Mitigation of Resource Based Conflicts in the Greater Horn of Africa: Conflict Baseline Study Report Conducted in the Karamajong Cluster of Kenya and Uganda.' Technical Report, Famine Early Warning Systems Network (FEWS NET)/GHA.
- Fisher, Franklin M. 1966. *The Identification Problem in Econometrics*. New York: McGraw-Hill.
- Food and Agriculture Organization of the United Nations. 2016. 'RIMA-II: Resilience Index Measurement and Analysis – II: Analysing Resilience for Better Targeting and Action.' Technical Report, <http://www.fao.org/3/a-i5665e.pdf> (Retrieved July 15, 2019).
- Fotso, Jean-Christophe. 2007. 'Urban-Rural Differentials in Child Malnutrition: Trends and Socioeconomic Correlates in Sub-Saharan Africa.' *Health & Place* 13(1): 205–223.
- Galán, José Manuel, Luis R. Izquierdo, Segismundo S. Izquierdo, José Ignacio Santos, Ricardo del Olmo, Adolfo López-Paredes and Bruce Edmonds. 2009. 'Errors and Artefacts in Agent-Based Modelling.' *Journal of Artificial Societies and Social Simulation* 12(1): 1.
- Geller, Armando, and Scott Moss. 2008. 'Growing QAWM: An Evidence-Driven Declarative Model of Afghan Power Structures.' *Advances in Complex Systems* 11(2): 321–335.
- Gneiting, Tilmann, and Adrian E. Raftery. 2005. 'Weather Forecasting with Ensemble Methods.' *Science* 310(5746): 248–249.
- Hatløy, A., L. E. Torheim and A. Oshaug. 1998. 'Food Variety – a Good Indicator of Nutritional Adequacy of the Diet? A Case Study from an Urban Area in Mali, West Africa.' *European Journal of Clinical Nutrition* 52(12): 891–898.
- Holland, John. 1975. *Adaptation in Natural and Artificial Systems*. Ann Arbor, MI: University of Michigan Press.
- Huttegger, Simon M., Brian Skyrms and Kevin J. S. Zollman. 2014. 'Probe and Adjust in Information Transfer Games.' *Erkenntnis* 79(S4): 835–853.
- Kim, Sung-Youn, Charles S. Taber and Milton Lodge. 2010. 'A Computational Model of the Citizen as Motivated Reasoner: Modeling the Dynamics of the 2000 Presidential Election.' *Political Behavior* 32(1): 1–28.
- Kollman, Ken, John H. Miller and Scott E. Page. 1992. 'Adaptive Parties in Spatial Elections.' *American Political Science Review* 86(4): 929–937.
- Krakauer, David, and Daniel Rockmore. 2015. 'The Mathematics of Adaptation (Or the Ten Avatars of Vishnu).' In *The Princeton Companion to Applied Mathematics*, ed. Nicholas J.

- Higham. Princeton, NJ: Princeton University Press, 591–597.
- Laver, Michael. 2005. 'Policy and the Dynamics of Political Competition.' *American Political Science Review* 99(2): 263–281.
- Laver, Michael, and Ernest Sergenti. 2011. *Party Competition: An Agent-Based Model*. Princeton, NJ: Princeton University Press.
- Laver, Michael, Scott de Marchi and Hande Mutlu. 2011. 'Negotiation in Legislatures Over Government Formation.' *Public Choice* 147(3–4): 285–304.
- Lustick, Ian S., Dan Miodownik and Roy J. Eidelson. 2004. 'Secessionism in Multicultural States: Does Sharing Power Prevent or Encourage It?' *American Political Science Review* 98(2): 209–229.
- Malleson, Nick, and Mark Birkin. 2012. 'Analysis of Crime Patterns Through the Integration of an Agent-Based Model and a Population Microsimulation.' *Computers, Environment and Urban Systems* 36(6): 551–561.
- Manners, Kristy. 2014. 'Nutrition Causal Analysis: Isiolo County, Kenya.' Technical Report, Action Against Hunger, [https://www.actionagainsthunger.org/sites/default/files/publications/ACF\\_NCA\\_Kenya\\_Report\\_Feb2014.pdf](https://www.actionagainsthunger.org/sites/default/files/publications/ACF_NCA_Kenya_Report_Feb2014.pdf) (Retrieved July 15, 2019).
- Masad, David, and Jacqueline Kazil. 2015. 'Mesa: An Agent-Based Modeling Framework.' <https://github.com/projectmesa/mesa>
- Mercy Corps. 2016. 'Karamoja Strategic Resilience Assessment.' Technical Report. <https://www.mercycorps.org/sites/default/files/UgandaSTRESSKaramojaFinalRep.pdf> (Retrieved July 15, 2019).
- Mitchell, Melanie. 1996. *An Introduction to Genetic Algorithms*. Cambridge, MA: MIT Press.
- North, Michael J., Nicholson T. Collier, Jonathan Ozik, Eric R. Tatara, Charles M. Macal, Mark Bragen and Pam Sydelko. 2013. 'Complex Adaptive Systems Modeling with Repast Symphony.' doi:10.1186/2194-3206-1-3.
- Nowak, Martin A., and Robert M May. 1992. 'Evolutionary Games and Spatial Chaos.' *Nature* 359(6398): 826–829.
- Palmer, Richard G., William Brian Arthur, John H. Holland and Blake LeBaron. 1999. 'An Artificial Stock Market.' *Artificial Life and Robotics* 3(1): 27–31.
- Parker, Dawn C., Kathryn H. Jacobsen and Maction K. Komwa. 2009. 'A Qualitative Study of the Impact of HIV/AIDS on Agricultural Households in Southeastern Uganda.' *International Journal of Environmental Research and Public Health* 6(8): 2113–2138.
- Richiardi, Matteo, Roberto Leombruni, Nicole J. Saam and Michele Sonnessa. 2006. 'A Common Protocol for Agent-Based Social Simulation.' *Journal of Artificial Societies and Social Simulation* 9(1): 15.
- Rogers, J. Daniel, Claudio Cioffi-Revilla and Samantha Jo Linford. 2015. 'The Sustainability of Wealth Among Nomads: An Agent-Based Approach.' In *Mathematics and Archaeology*, eds Juan A. Barcelo and Igor Bogdanovic. Boca Raton, FL: CRC Press, 431–438.
- Samimi, C., A. H. Fink and H. Paeth. 2012. 'The 2007 Flood in the Sahel: Causes, Characteristics and Its Presentation in the Media and FEWS NET.' *Natural Hazards and Earth System Sciences* 12(2): 313–325.
- Schelling, Thomas C. 1971. 'Dynamic Models of Segregation.' *Journal of Mathematical Sociology* 1(2): 143–186.
- Skamarock, William C., and Joseph B. Klemp. 2008. 'A Time-Split Nonhydrostatic Atmospheric Model for Weather Research and Forecasting Applications.' *Journal of Computational Physics* 227(7): 3465–3485.
- Stites, Elizabeth, and Carrie S. Huisman. 2010. 'Adaptation and Resilience: Responses to Changing Dynamics in Northern Karamoja, Uganda.' Briefing Paper, Feinstein International Center, Tufts University, Boston..
- Tullock, Gordon, and Colin D. Campbell. 1970. 'Computer Simulation of a Small Voting System.' *The Economic Journal* 80(317): 97–104.
- Uganda Bureau of Statistics. 2017. 'National Population and Housing Census 2014 – Analytical Report.' [https://www.ubos.org/onlinefiles/uploads/ubos/2014CensusProfiles/National\\_Analytical\\_Report\\_nphc%202014.pdf](https://www.ubos.org/onlinefiles/uploads/ubos/2014CensusProfiles/National_Analytical_Report_nphc%202014.pdf) (Retrieved July 15, 2019).
- United States Agency for International Development, ed. 2012. *Building Resilience to Recurrent Crisis: USAID Policy and Program Guidance*.
- Urbanowicz, Ryan J., and Jason H. Moore. 2009. 'Learning Classifier Systems: A Complete Introduction, Review, and Roadmap.'

- Journal of Artificial Evolution and Applications*, 1–25. doi.10.1155/2009/736398.
- Weidmann, Nils B., and Idean Salehyan. 2013. 'Violence and Ethnic Segregation: A Computational Model Applied to Baghdad.' *International Studies Quarterly* 57(1): 52–64.
- Wilensky, Uri. 1999. 'NetLogo.' <http://ccl.northwestern.edu/netlogo/>.
- Young, Helen, and Anastasia Marshak. 2017. 'Persistent Global Acute Malnutrition.' Technical Report, Feinstein International Center, Tufts University, Boston.



# Taking Data Seriously in the Design of Data Science Projects

Mitchell Goist and Burt L. Monroe

## INTRODUCTION

In its broadest sense, every empirical project is a ‘data science’ project, so every chapter in this *Handbook* contains relevant advice for data science projects. Conversely, some might define ‘data science’ more narrowly, associating with particular approaches to inference from data (e.g., *machine learning*, *deep learning*, *Bayesian inference*), or with particular modes of data collection (e.g., *web scraping*) or with the set of methodologies and technologies that have emerged for drawing inferences from relatively new sorts of ‘nonrectangular’ data, ranging from data with complex structure (e.g., spatial or network data) to data with no obvious structure as social science data at all (e.g., text or image data). Most of these topics have one or more chapters devoted to them here as well. That is to say, we’re all data scientists now.

That larger context acknowledged, our focus here is on the data in ‘data science’,

especially in computation-intensive or data-intensive projects where data is<sup>1</sup> unusually complex or ‘big’ in some way and requires more than conventional attention to how it is handled and processed. We try to address common issues and challenges that arise as data moves through the workflow of data science projects in the social sciences, and direct the reader to possible solutions and useful concepts and tools.

## ***Workflow of a Data Science Project***

There are several frameworks for characterizing the ‘workflow’ or ‘pipeline’ of a data science project. The most ubiquitous are Mason and Wiggins’ (2010) semi-romantic characterization of data science as OSEMN (‘awesome’, if you squint) – Obtain → Scrub → Explore → Model → iNterpret – and the industrial data warehousing characterization of pipelines composed of individual, ETL



modules – Extract → Transform → Load (Kimball and Caserta, 2004). There are many others of varying levels of complexity, with various pieces of pipe named, conceptualized or ordered differently. These share some common elements and a general idea of a process that ingests some raw data, transforms the data in various ways and then provides some meaningful output at the end.

In the present context, we'll delineate four stages in a social scientific data science pipeline:

- Data Collection: Human Behavior → Raw Data
- Data Wrangling: Raw Data → Clean/Structured Data
- Data Preparation: Clean/Structured Data → Analysis-Ready Data
- Data Analysis: Analysis-Ready Data → Inference/ Interpretation

We note that these are the stages in a *replication* pipeline. Typically, each (and always some) of those stages can be instantiated in code, such that if given the same inputs any researcher using the code would get the same (or sufficiently similar) outputs.

We also note that each stage may have a more complex structure underneath. Any one step may be a series of consecutive transformations.<sup>2</sup> There may be multiple inputs that have to be merged in some way. There may be multiple types of analyses that require multiple outputs. There are almost always processes that benefit from a 'split-apply-combine' or 'map-reduce' conceptualization, under which data is split, each split has an identical sequence of transformations applied to it and the results are recombined (Wickham, 2011). We return to this concept below.

And, of course, such a pipeline does not describe the actual *workflow* of a social scientist conducting research and constructing the pipeline in the first place. When conducting the original analysis, creating the pipeline, each stage is embedded in a social science workflow that might be something like:

- **Anticipate:** What outputs are needed as inputs to the next stage? What could go wrong?

- **Theorize:** What do I expect?
- **Design:** How do I design this stage to provide valid outputs (at acceptable cost in acceptable time)?
- **Collect, Wrangle, Prepare, or Analyze...**
- **Store:** How do I save the output for use in the next stage or the future?
- **Validate:** Is the output what I think it is? Is it right?
- **Document:** What exactly did I do? What would I, or anyone, need to know, or have, to do it again?

In particular, theory and research design are the focus of other chapters in this *Handbook*, so we touch on these only lightly here. Finally, we note that given the extensive coverage of many approaches to the Analysis stage elsewhere in this *Handbook*, as well as many aspects of, and approaches to, the Collection stage (including experimental design, survey design and web scraping), we focus primarily on a somewhat idiosyncratic collection of remaining pieces of a data pipeline that the social scientific data scientist may need or wish to consider.

### A Basic Data Science Toolkit

The core of your data science toolkit<sup>3</sup> will likely be a *high-level programming language*, and there are two such languages that currently dominate data science: Python and R.<sup>4</sup> Python is generally perceived as better at the data manipulation processes that occur in the Collection and Wrangling stages of the pipeline, better matched with a machine learning approach – especially a deep learning/neural net approach – to the Preparation and Analysis stages, and better at scale. R is generally perceived as better at the Analysis stage, especially with a statistical inference approach, and better when visualization plays an important role in validation, analysis or documentation. That said, each has dramatically improved on perceived deficiencies in recent years, and there is not as much that one can do that the other cannot. Python is the more used in industry data

science; R is the more used in academic data science, at least in the social sciences. It's not unusual for a pipeline to involve both (ours typically do), and we discuss below some ways of communicating between them.<sup>5</sup>

It is also good practice to develop your code in an *IDE (Integrated Development Environment)*, such as Spyder for Python (and R) or RStudio for R. These integrate well with notebook systems such as R Notebooks and Jupyter, an important tool for 'literate programming', which we discuss further below under the topic of open science and reproducibility. Spyder and Jupyter are included in the open-source Python (and R) distribution, Anaconda, which comes with most data science oriented packages (including the 'PyData stack' of at least numpy, scipy, pandas and matplotlib) installed alongside a user-friendly Python package and environment manager, conda.

Most scientific computing environments, and essentially all university research computing clusters, require you to interact with a *Linux/Unix shell*, most commonly bash. It is easy enough to learn the bare minimum necessary to move around directories, manage files, start software and submit scripts for later processing, but we encourage some investment in learning a bit about data manipulation from a unix shell. You will be amazed at how handy it is for a data scientist to know unix commands like `head`, `tail`, `cut`, `split`, `count`, `sort`, `uniq`, `tr`, `ag/grep`, `sed`, `awk` and `curl/wget`.<sup>6</sup> These can be used for very efficient retrieval, cleaning, manipulation and management tasks in a surprising variety of settings. It is also likely that the easiest way to integrate diverse pieces of a pipeline containing modules in Python, R and other languages is with a shell script that evokes these modules in sequence.

You should also incorporate *version control* into your workflow, most likely using *git* and a git repository like github. The most immediate benefit is saving your work. Not only can you insure against inevitable computer crashes, but the ability to revert back

to older versions or to create tentative project 'forks' helps insure against your own mistakes. Git really shines when you need to work with a partner or team, as it smoothly handles conflicting changes made along different project branches. Moreover, github and similar offer convenient ways to code in the open, aiding reuse and replicability.<sup>7</sup>

The core data science programming tool that in practice seems least likely to be learned in passing by social scientists is *SQL (Structured Query Language)*. SQL is the language for interacting with conventional relational databases, by far the most likely database system you will encounter or yourself set up. SQL's core functionality is querying, extracting, filtering, aggregating, reshaping and merging data stored as tables in a relational database into output tabular data. Furthermore slight SQL variants are used by other database systems (e.g., Spark SQL or Apache Drill) and SQL concepts inform more familiar data wrangling and management formalisms in R (e.g., `dplyr`, `data.table` or `cdata`) and Python (e.g., `pandas`). We return to these concepts below.

## DATA

Our pipeline above imagines three major intermediate states that data can be in: *raw* (or *dirty* or *unstructured*), *clean* (or *structured*) and *analysis-ready*. Of course, whether data is raw or clean, structured or unstructured, analysis-ready or not is largely determined by context. Different analyses may call for different arrangements of the same data. For example, an effort to create political event data ('Group A protested against Government B on date C in place D') from news articles might require text input in full sentences with original capitalization (to help identify 'named entities') and punctuation (to help identify the subject and object of complete sentences). An effort to model the topics of those same articles might require

per article counts of words that have been stripped of order and punctuation and ‘normalized’ (e.g., ‘case-folded’ to lower case, or ‘stemmed’ or ‘lemmatized’ to root form).

Moreover, the clean data of one process may be the raw data of another. The Nexis database of news articles would be considered by information scientists to be highly clean and structured, able to respond to many simultaneous complex queries defined by search terms, outlet, date and so on, as well as incorporating and indexing new data each day. But from the point of view of an event data or topic modeling project, the output of such a query is something between raw data (What’s the text of the article vs. metadata like title, source, date? Are there duplicates or multiple versions?) and not-data (Am I actually allowed to download all the article content?).

### **Data in Memory**

Within any of the computational processing steps of the pipeline, the data being collected, wrangled, prepared or analyzed are held, at least temporarily, in a computer’s volatile working or cache memory. Sometimes all of the relevant data is in the memory of one computer simultaneously. Sometimes it is in pieces that pass through memory sequentially. (Even transferring a file from disk to disk involves pieces of the data passing temporarily through memory.) Sometimes it is in pieces that are distributed across the memory of multiple processors or computers. There are at least three levels at which a data science project might deal with data in memory: as bytes and primitive data types, as abstract data types and data structures or as specialized data objects including ‘tables’ and other analysis-ready data.

At a low level, data is what your cellphone provider means by data: sequences of eight-digit binary numbers, called *bytes*, that take up ‘space’ in memory, in storage or in bandwidth. That is, any digital information is,

by this definition, data. The ‘digital revolution’ has been enabled by the greater number of human activities that are born digital (e.g., social media posts, web searches), the improved technology for sensors and other technologies to convert nondigital activity to digital traces (e.g., optical character recognition, GPS devices) and our greater capacity to gather, store, share and mashup digital information.

The ways in which bytes can be directly interpreted correspond to the *primitive data types* of a programming language, defining what sorts of data a variable of a given type can hold and what operations can be applied to it. Loosely speaking, there are three primitive data types that essentially all systems have: integer, floating point (‘numeric’ in R) and Boolean (TRUE/FALSE, binary, ‘logical’ in R). Knowing how these work in your current computing setting can help you avoid some common problems. So, for example, in R, `sqrt(2)^2 == 2` returns FALSE, because floating point data (‘numeric’ in R) can represent real numbers like  $\sqrt{2}$  only to a certain level of precision. A vector of integer ones (like `x <- rep(1L, 100)`) takes about half as many bytes to hold in memory as a vector of floating point one-point-zeroes (like `x <- rep(1, 100)`), but exactly the same as a vector of Boolean ones (like `x <- rep(TRUE, 100)`), because Boolean data is just integer data interpreted differently (0 means FALSE, everything else means TRUE).

More important, these bear only loose correspondence to the way social scientists have traditionally thought of data as measures. The most basic social scientific understanding of data is commonly characterized by the Stevens (1946) *levels of measurement* hierarchical typology. Stevens asserted that scientific measures are on one of four types of ‘scales’ – *nominal* (often called *categorical* or *qualitative*; two observations can be equal or not equal), *ordinal* (two observations can be compared and one deemed greater than the other), *interval* (two observations can be compared – subtracted – for meaningful

difference) and *ratio* (two observations can be compared – divided – for difference in magnitude relative to a meaningful zero).<sup>8</sup> The integer data type can be mapped to the ordinal, but floating point does not distinguish between interval and ratio measures, and Boolean can represent nominal data directly in a single variable only when there are exactly two categories.

To represent multi-category data (in one variable), we need a new *composite data type*. We need to enumerate the categories, store the integers representing each observation's category *and* store the mapping from integers to category labels. In R this is the 'factor' data type. Python itself does not have a categorical data type, but the library pandas does ('category').

The notion of labels brings us to textual (in most languages *character* and/or *string*) data, which is also generally represented in a composite data type. Strings are particularly common in data science, but can be tricky. They are intended to be read by humans, but they must be structured and stored as sequences of bytes, using one of many possible *encodings*, systems that map specific characters to specific byte sequences. A common error results from software – you may have noticed this in a browser or Excel, if not in your scripts – encountering UTF-8 encoded (a Unicode encoding) characters and interpreting (*decoding*) them as if they were, say, 'Latin-1', 'ISO-8859-1' or 'Windows-1252' encoded. These character sets overlap for the 256 characters encoded with a single byte, so an English or Western European language text will look mostly right. But characters that require two bytes to encode in UTF-8 will be interpreted as two characters. So a Spanish word may mysteriously contain 'Ã±' instead of 'ñ' and a text enclosed in 'curly quotes' may instead conclude with 'â€' (the telltale signs are the random accented 'A' and 'a' characters). Those working with data from social media or cell-phones might be mystified by the four-byte '\U0001F44D' or b"\xf0\x9f\x91\

\x8d' if their font or environment doesn't include ☹.<sup>9</sup>

Composite data types lie on a fuzzy continuum between primitive data types and *abstract data types* or *data structures*. These are often base objects that can be used and manipulated in a given programming language, and include things like lists, tuples, sets, queues, stacks, linked lists, trees, graphs, heaps or hash tables. Conceptually, data structures are often paired with 'algorithms' (in book and course titles; e.g., Klein, 2016), and many industry coding interviews contain a 'data structures and algorithms' section), since efficiency of algorithmic tasks such as sorting or searching depends on the data structure.

To a social scientist, structured data often refers to data that is ready for analysis. For the vast majority of techniques this means *rectangular* data which can, hypothetically at least, be displayed in a spreadsheet-like format with rows corresponding to observations and columns corresponding to variables or features. These are typically referred to as *tables* or *data frames*.<sup>10</sup>

A table can have features of different types, but it is often the case that 'big data' is big because there are many features with identical primitive data types, levels of measurement and scale. That is, the cells of the table are all integer counts (e.g., of a given word), all Boolean (e.g., presence/absence of a relationship link), all on the same scale (e.g., ratings from one to five stars), or similar. In this case, these can be thought of – and generally benefit computationally – from being held and manipulated in *matrix* or (two-dimensional) *array* data structures (e.g., in Python a numpy array rather than a pandas DataFrame).

Further, it is almost always the case that such data are mostly zeroes (or missing), in which case there are often substantial computational benefits to storage in a *sparse matrix* format that stores only information about each nonzero.<sup>11</sup> There are some common pitfalls when handling data in sparse matrices, however. One of these is accidentally forcing

the conversion of the sparse matrix into a dense one by, for example, adding a floating point 1.0 to every element and then taking the log of every value. Now you still have a lot of zeroes, but you're storing them all. The other common pitfall has to do with whether the sparse matrix is *row-oriented* or *column-oriented*. Consider the example of a typical observation–feature matrix (like documents represented by rows, and words represented by columns). The row-oriented format will store values for each document together; the column-oriented format will store values for each word together. As a result, calculations on each document (like length) will be very fast in the row-oriented format and very slow in the column-oriented format, and vice versa for calculations on each word (like average frequency).

Ultimately, many types of analysis require data arranged in accordance with a social structure of interest, such as space, time, network or hierarchy. Some of these are specialized sorts of matrices (e.g., adjacency matrices representing networks, or images represented by grids of pixel intensities), but some are genuinely different (e.g., GIS vector structures that describe the shape and location of geometric objects in geographic space). This funneling into specialized data formats and tools is another source of friction in data science projects. Social media data or political event data or GPS tracking data all contain spatial, temporal and network structures. But we have only limited ways to represent and analyze the network information in a conventional geographic information analysis tool like ArcView, and limited ways to represent or analyze the spatial structures in a specialized network analysis tool like Gephi. In part because of this tool-dependence, we tend to technically specialize around these data structures, limiting our ability to conceive of the data along other lines. (Peuquet (2002) has an elegant discussion of this phenomenon in the context of incorporating time into the static spatial representations at the core of GIS systems.)

This is one motivation for a ‘team science’ approach, a concept we return to in our concluding remarks.

### **Data in Files and Streams**

Data science gets its reputation for being dominated by data cleaning largely because of data in files. Our ability to gather, store (or *persist*) and copy data far outstrips our ability to make sense of it. Your research project's needs do not typically inform how that data is stored or delivered to you; moreover, we need to persist and transfer our own data in a way that allows for replicability and future unknown use. We can't begin to address the full variety of possibilities here – for example, Wikipedia provides an idiosyncratic list of more than 2,000 distinct file formats<sup>12</sup> – so we concentrate here on a few of the more commonly encountered formats and a few general concepts.

Perhaps the most foundational concept for the design of data pipelines is *serialization*, the process of converting data objects in memory into a sequence of bytes (i.e., ‘serially’) that can be transmitted or stored and reconstructed (*deserialized*) later, potentially in a different computing environment. If you're going to construct data pipelines where the output of one stage is written to disk before being ingested by a subsequent stage in a (potentially) different computing environment, or even if you wish simply to save intermediate or final data for future unknown purpose, serialization formats offer the greatest portability and consistency. If you are ingesting data from an open data service and/or data API,<sup>13</sup> well-constructed ones will provide the data in one or more serialization formats. Serialization formats vary in their efficiency, their portability, their transparency and their universality.

Among the most familiar, because they are common formats in which data is delivered, are *JSON* (JavaScript Object Notation) and *XML* (eXtensible Markup Language). *JSON*

and XML are *human-readable text-based* serialization formats, both of them arranged around a hierarchical tree-like structure and human-readable ‘keys’. This is inefficient from a storage point of view, but makes these formats portable, searchable, editable and more or less *future-proof*.

XML is the older and less fashionable of the two, but is still fairly common in ‘open data’ settings. It can be constrained by a *schema* and easily queried, which has led to many specialized variants of XML for specific purposes. Examples include DOCX and XLSX (Microsoft Office files), ODF (Open Document Format, for Open Office files), XHTML (a formalized variant of the web page format HTML), KML (Keyhole Markup Language, a format for specifying geographic data for display in Google Earth and similar) and RDF/XML (a serialization of RDF, Resource Description Framework, the format for internet resources, including metadata and semantic web/‘internet of things’ objects). XML can get complex enough that its human readability can become questionable.

JSON is the newer of the two, and increasingly the one used as a data serialization format, particularly with data APIs. The Twitter APIs, for example, return search results in JSON. There are differences in what you can do with JSON and XML, but JSON is generally more readable, more compact and less memory intensive to parse. Although originally designed for JavaScript, it is now so ubiquitous that most programming languages, including R and Python, have dedicated JSON parsers that can be used to navigate JSON files. JSON also has some standardized variants, such as JSON-LD (‘linked data’), an emerging format for enabling interoperability of open data, and JSON Lines, a compelling record-per-line format (allowing, for example, parallel processing) that has been adopted by some APIs (e.g., plotly) and is the serialization format recommended by Python web scraping module Scrapy.<sup>14</sup>

At the cost of human readability and some future-proofness, binary serialization formats offer more compact storage, more efficient read/write operations and the ability to store more general classes of data objects. Python users will be familiar with the *pickle* and *cPickle* modules which provide serialization for Python objects. R users seem much less likely to be familiar with the *RDS* (‘R Data Serialization’) format for serialization of R objects. These are ideal for efficient passing and persisting of data between stages of a data pipeline, as long as human readability isn’t necessary for debugging or some similar purpose. (R users seem more likely to save collections of objects, or entire workspace images, in ‘Rdata’ files. This is useful for maintaining continuity across different R sessions, but considerably less useful for passing data objects across segments of a data pipeline, particularly from R to Python or another language.)

JSON has a binary variant called *BSON* (‘binary JSON’), used by the popular document-oriented data store MongoDB as a storage format, and can also be used in the relational database PostgreSQL. BSON offers more compact storage than JSON as well as the ability to store some additional data types. There are also multiple serialization formats designed to work with distributed big data environments such as Hadoop and Spark, with Apache Avro (developed as part of Hadoop), Apache Parquet (developed at Twitter and Cloudera) and Protocol Buffers (developed at Google) among the most popular.

Of course, much more common than any of these in standard social science practice are *flat files*, text files representing tabular, spreadsheet-like data. Each line of the file represents a row of the table (or maybe a header with variable names), with values for each column separated by a designated *delimiter* character. This is most often a comma, and such files are of course called *CSV* (Comma-Separated Values, or occasionally Character-Separated Values) files. CSV is both more familiar and more compact than

JSON or XML, taking about half the space. (A flat file can be stored in almost identical space as CSV in JSON Lines, however.) CSV is more or less human-readable, has standard read and write commands in Python, R and most languages, and can typically be opened directly in spreadsheet software like Excel or econometrics software like Stata. CSV is, however, as much bug as it is feature. The encoding assumed may be wrong, data types might be misinterpreted, strings might be truncated at a maximum length or any number of other errors might be introduced, changing your data when you only meant to look at it.

CSV is considered a serialization format, but an inconsistently defined one with many ‘dialects’ defining different encodings, string delimiters, escape characters and so on. Among other potential problems, raw string data, especially running text and numerical data represented as strings, contains ubiquitous commas or quotation marks that can cause havoc when the file is read in under different assumptions. The use of a tab separator rather than a comma – often given the extension TSV – or another character (ideally one that cannot occur in your data) can help make these problems less likely, but there is still no way within a delimited file itself to specify what data types the strings in the file are meant to represent. A format like JSON or JSON Lines avoids these problems altogether.

Arguably, the most frustrating file format the data scientist will regularly encounter is *PDF* (‘Portable Document Format’), known infamously as the format ‘where data goes to die’.<sup>15</sup> You will encounter PDFs *you* can read or print, but which are just scanned images of the original text. These require manual keying or OCR (‘optical character recognition’). You will encounter PDFs with OCRed text riddled with errors (99% accuracy at the character level can still leave a mistake every other line). Even if the text is searchable and there are no character level mistakes, PDF is a *display format* that arranges the characters

on the page to display correctly, but not generally in a semantically logical order. This can be particularly maddening when the page includes multiple columns, tables or characters from right-to-left languages such as Arabic or Hebrew. The best case scenario is often that a tool like xpdf, poppler, tabulizer, pdftools (R) or PyPDF2 (Python) can properly identify blocks of consecutive text, which still require some bespoke wrangling.

There are many further complications that can come with details associated with the generation and interpretation of a file. We have already mentioned encodings of text files, but there are even more complicated *codecs* (‘coding/decoding’) issues that arise with the handling of media data such as image, audio and video. You need to get these details right to make sense of such data, and to be sure others can make sense of yours. Moreover, such files tend to be large in their natural state, so they often undergo *lossy compression* when stored, and especially when streamed over a network. Many of the most familiar formats (e.g., JPEG for images, MPEG for video, MP3 for audio) contain only an approximation of the original data. This may be fine for watching a video on your phone, but problematic for training a model to learn from that video.

### **Data Stores and Databases**

To information science and database specialists, data is structured if it is in ‘normal form’ (or *normalized*) for a *relational database*.<sup>16</sup> Roughly, this means that every type of object has its own object × attribute table, and there are no duplications in data or dependencies in relations among those tables. So, for example, a relational database of individual campaign contributions to political candidates would have separate tables for (at least) contributions, donors and candidates. The donor table would have one row per donor, with, say, (donor id, donor name, address, zip code, registration) as columns. The candidate

table would have one row per candidate, with, say, (candidate id, candidate name, party) as columns. Finally, the contribution table would have one row per donation, with, say, (contribution id, donor id, candidate id, amount, office, date). The schema of the database would define data types of fields and the relevant relationships between the tables, namely that each contribution is associated with one donor and one candidate. To calculate, for example, the total contributions by zip code, one would first need to join/merge the relevant information from the donor table with the contributions table, creating an unnormalized table of the relevant columns of the contribution table with the matching zip code added as a new column, and then aggregating contributions by zip code. SQL is the standard language for executing queries of this sort.

The advantage of this is largely in limiting errors when data is entered or changed. If a donor changes address, we need only change a single record. The primary disadvantage is that our analysis step almost always needs the data unnormalized and these join operations become time consuming and unwieldy as our data grows. So, while it is certain that you will need to understand SQL concepts, if not directly use SQL, to query someone else's database, it's often not terribly useful for a social scientific data scientist to incur the overhead of setting up and maintaining a relational database and server.

Alternatively, we encourage you to look into nonrelational and *postrelational* database or data store options for management of unnormalized data at scale. First we note that our data now can easily get too big to store on a single node, which has led to the development of *distributed data stores* that spread and replicate data across different nodes of a cluster. The most well-known perhaps is HDFS, the Hadoop File System. Many postrelational databases can be deployed on Hadoop and similar data stores. There are also search and query engines that can be used to access data on such stores. Popular

pairings include Elasticsearch and HDFS, and BigQuery designed to work with Google Storage.

Second, the fairly dizzying array of options for postrelational data management is otherwise best distinguished by the data model it contains. One of the most popular, and one that has penetrated social science, is the already mentioned MongoDB, a *document-oriented* database that uses JSON as a model. Another important alternative model is 'wide column store', which stores data based on fields/features (columns) rather than by records/observations (rows). Much of Google infrastructure is run on proprietary column store BigTable, and there are several open-source spinoffs, with Cassandra currently the most popular.

If you need to store large 'unstructured' binary objects – image, audio and video are common examples – there is often little to be gained over simply storing them as files. A separate database can hold file pointers and any metadata, which is usually all you could execute a query on.<sup>17</sup> One option to consider, however, is an *object store*. This is similar to a familiar file system, except objects can have arbitrary and searchable metadata attached to them in the store. These are most commonly used in cloud storage systems like Amazon's S3 service and Microsoft Azure's Blob Storage. Another possible aid in the storage and exchange of complex collections of arbitrary binary data is *HDF5* (Hierarchical Data Format). HDF5 is actually a file format that can effectively contain an elaborate self-describing data store within a single file. It is not for serialization as such, but HDF5 files can be straightforwardly read into and written out from Python.

## DATA COLLECTION AND JINGLE-JANGLE FALLACIES

One of the main promises of data science is the almost endless array of data that is now collected and available: digital records of all



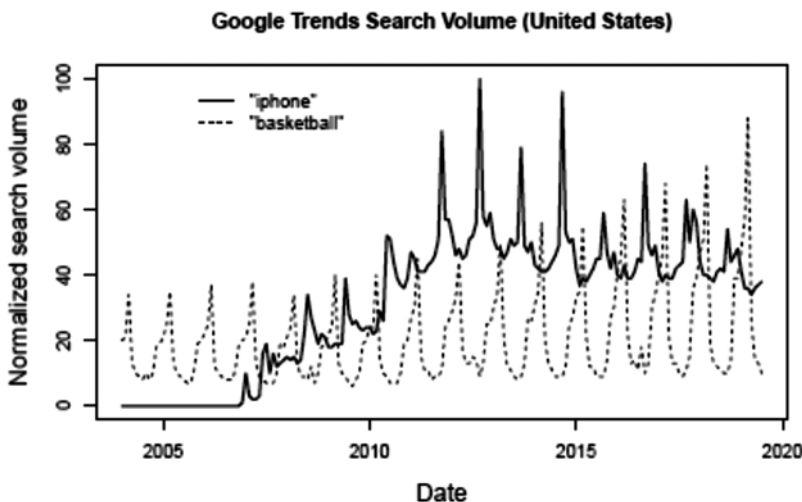
shapes and sizes, from governments and companies, from individuals through social media and distributed sensors, through innovative experimental designs, through innovative uses of human-in-the-loop computing and citizen science platforms, and in forms – natural language, images, audio, video, etc. – that have only recently begun to be exploited in the social sciences. There is almost endless room for creativity (see Salganik, 2017 for discussion of dozens of examples).

By far the most exploited class of new data is *data exhaust*, the digital traces of contemporary life from social media, cellphones, shopping, online searches and so on, which appear ripe for interpretation as *unobtrusive measures* (Webb et al., 1966) of human and social behavior. The key characteristics of data exhaust are that it is more or less always being passively generated, it is often proprietary and generated by proprietary processes and it is typically too big to use or quality check in its entirety.

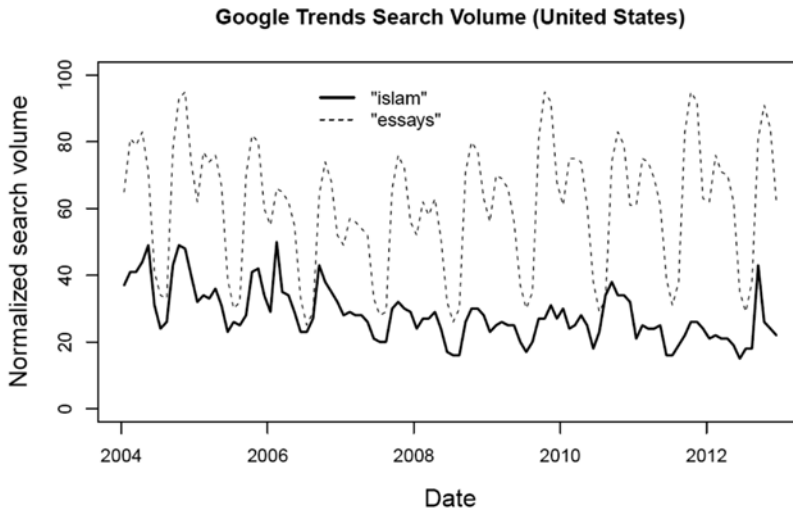
In this section, we discuss some subtle issues when such data are taken at face value as more or less direct measures of

underlying behaviors. Measurement theorists have warned us for many decades to beware the *jingle fallacy* – that two things with the same name are the same thing (Thorndike, 1904) – and the *jangle fallacy* – that two things with different names are not the same thing (Kelley, 1927). We seem particularly inclined to these fallacies in the context of data exhaust.

Consider, for example, the use of Google search data as an indicator of general public interest in a subject, and in turn, particular drivers of that interest. It seems almost tautological that searches for ‘word’ should indicate interest in the subject of ‘word’, and generally speaking this seems to have face validity. For example, Figure 5.1 shows plausible dynamics in searches for ‘iphone’ and ‘basketball’. This type of approach has been used most famously, and infamously, in the ‘Google Flu’ project (Lazer et al., 2014), in which searches for terms like ‘flu symptoms’ were (initially) shown to predict traditional, but slower, influenza outbreak data reported by doctors through US Centers for Disease Control. Political scientists have used search



**Figure 5.1** Searches for ‘basketball’ have the expected seasonality and a plausible pattern of growth. Searches for ‘iphone’ do not appear until the iPhone is invented, display expected spikes around model release dates and suggest a plausible decline in interest from a peak reached with the iPhone 5 and 6



**Figure 5.2** The prominent dynamic in searches for ‘islam’ is the same as that in searches for ‘essays’. US internet search for some topics is dominated by the academic calendar and the demands placed on high school and college students in common courses

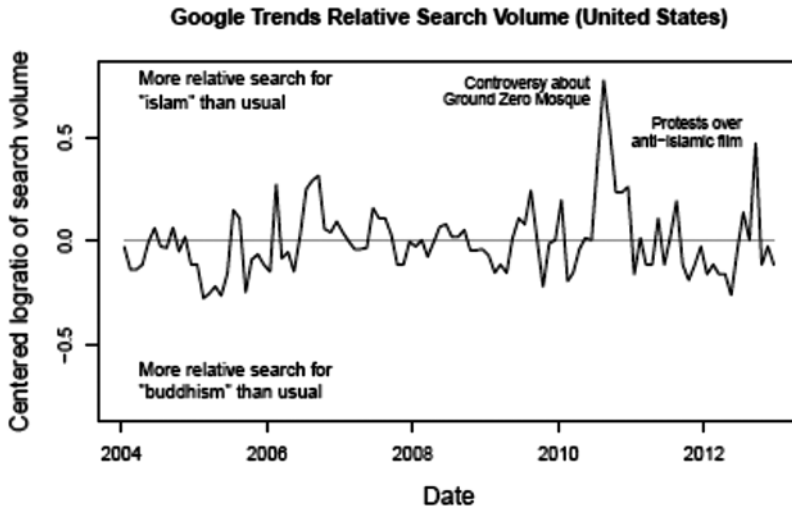
data as well, e.g., in searches using racially charged language as an indicator of spatial variation in racial animus during the Barack Obama presidential campaign (Stephens-Davidowitz, 2014).

Consider, however, Figure 5.2, which shows search volume for ‘islam’, occasionally used as an indicator of salience of Islam-related issues among the general public (e.g., Smith, 2013). The annual seasonality is mysterious<sup>18</sup> until you see a similar pattern in a term like ‘essays’. It appears that internet search volume in the United States for topics that might be covered in common courses such as western civilization or world history – especially before 2012, as depicted – is dominated by college and high school students.<sup>19</sup>

Similar phenomena have been found in the Google Ngrams data built from Google Books. One of the most telling is the apparent exponential growth since 1900 in use of the term ‘Figure’ relative to the term ‘figure’ (Pechenick et al., 2015). This reflects, of course, the growth in the percentage of university library holdings – the source material

for Google Books – that are composed of scientific journals. Someone researching the history of internet abbreviations might be alarmed that the terms ‘lol’ and ‘wtf’ reached their peak usage around ... 1660. This is a consequence of OCR mistakes being particularly egregious in older texts, and these mistakes being amplified by there being a trivial number of texts with 17th-century dates in the first place.

Perhaps even more pernicious with commercially created data exhaust such as this is *algorithmic confounding*. The relationship between the phenomenon of interest and its digital traces are a function of proprietary and changing algorithms, and those are driven by commercial motivations rather than utility for your research. Lazer et al. (2014) discuss, for example, *red-team/blue-team dynamics*. The blue team identifies changes in an algorithm that provide commercial value, such as click-through or purchases. The red team defends the algorithm from manipulation and malicious attack. A plausible blue-team culprit in Google Flu’s decline was the introduction of search auto-complete suggestions, which



**Figure 5.3** Comparing to a reference term captures more relevant shifts in attention

made search more helpful for users but likely changed the natural tendency of ‘flu’ to be extended to ‘flu symptoms’, or ‘bird’ to ‘bird flu’. Both red and blue teams, in any case, have an incentive to obfuscate changes and their impacts. As a result, the outside researcher cannot count on digital exhaust to carry consistent conceptual meaning over time.

There are a couple of strategies one can use to mitigate these problems. The first is to calibrate your big/cheap/noisy/biased data with small/expensive/precise/accurate data that you believe in, if possible. A parallel in traditional design is ‘double sampling’, where you might, for example, estimate the age of trees of a given species in a forest by (a) estimating their height by eye from the ground and then (b) cutting a small number of those down and counting the rings, using those to determine the relationship between your height guesses and their actual age (Thompson, 2012). In the context of surveys, Salganik (2017) calls this ‘amplified asking’, in which a small subset of individuals in a large imperfect data source are surveyed about the exact concepts of interest. These are then linked statistically to convert the large data into estimates of the concept for everyone in the larger source.

As with Google Flu, however, algorithmic confounding can make it difficult to confidently generalize your calibrations beyond the original time or context.

The second is to think of such measures with this question in mind: *relative to what?* Is there a sensible ‘control term’? For example, Figure 5.3 shows the relative Google Trends data for ‘islam’ vs. ‘buddhism’. This removes the shared content (world religions, etc.), and reveals spikes that seem to be associated with the summer 2010 stir up about Sharia law and the ‘Ground Zero Mosque’, and the September 2012 worldwide protests over the anti-Islam film *The Real Life of Muhammad*. Also, one need not have discovered the ‘essays’ phenomenon to identify ‘buddhism’ as a plausible control term.

Similarly, it is often useful to track ‘placebo’ terms, which aren’t expected to display any pattern at all. The reader is invited to search on the Google Books Ngram Viewer<sup>20</sup> ‘the’ since 1650. This provides us with warnings about the early noisiness, as well as a mysterious (to us) decline that began about 1850 and accelerated around 1970. This can be even more useful in your own bespoke data collections, as placebos can be diagnostic

of things like scraper and server errors that might otherwise cause drops in terms of interest to be interpreted substantively.

Another practice that can ‘jingle’ in our ear misleadingly is the use of specific terms to identify a relevant subset of impossibly large data for one’s analysis. Consider, for example, the dataset created by Mohammad et al. (2016). This has been used as a gold standard test bed for the NLP task of ‘stance classification’, identifying whether a text (in this case a tweet) is ‘favorable toward’, ‘against’, or neither, with respect to a target concept (in this case, ‘atheism’, ‘climate change is a concern’, ‘feminist movement’, ‘Hillary Clinton’, and ‘legalization of abortion’). A tweet got into this dataset in the first place by ending with one of a manually selected set of hashtags thought to be about the target and favorable (e.g., #GOHILLARY), against (e.g., #HillNo), or ambiguous (e.g., #hillary2016). For the ‘climate change is a concern’ target they were unable to identify a hashtag they considered clearly favorable (#climatechange was considered neutral), but nonetheless ended up with a dataset so skewed the *other* way (fewer than 5% were hand-labeled ‘against’) that the classification task was effectively impossible. At a minimum, the chosen hashtags did not have the filtering properties the researchers thought they had. Recent work on keyword expansion (King et al., 2017; Linder, 2017) provides techniques for understanding and tuning the precision and recall of your search terms.

## DATA WRANGLING

The received wisdom in data science is that most of your time will be spent in the data wrangling phase. It seems to also be widely held that data wrangling is entirely *ad hoc*. There are, however, theoretical frameworks, common wrangling tasks and relevant tools that can help make this process more principled. The most well-constructed theoretical

frameworks for data science come embedded in particular software, with prominent current examples including tidyverse/dplyr (R), data.table (R), cdata (R), pandas (Python) and Trifacta Wrangler. These each come with quasi-religious adherents and opponents, but we have found each of them useful both for practical projects and for framing our thinking on wrangling.

First, the big picture. What’s the goal? What would clean data look like for you at the end of the pipeline? It’s almost definitely some sort of table or matrix, or a set of them. We often find we are aiming for something of a hybrid, often a matrix (think document-term or donor-candidate), usually sparse, with an additional data frame of relevant data or metadata for the rows and one for the columns, consisting at least of labels.

Next, think of your input as tables. For all our discussion of new forms of data as ‘non-rectangular’, it is always possible and usually useful to think of even raw inputs of a data wrangling pipeline as tables. The main necessary abstraction is allowing that the ‘cells’ of these tables may contain complex objects, including other tables. So, near the beginning of a pipeline, your ‘table’ may, for example, be of dimension  $n \times 2$ , with one column a list of  $n$  raw data objects (e.g., documents, images, audio files, websites, adjacency matrices, shapefiles, JSON files) or even *pointers* to such objects (filenames, urls), and the second column a list of *metadata* associated with the objects (e.g., labels, source, date). Or you might have an unprocessed natural language text, which might be thought of as a single column of observations, each of which is a document, or page, or line, or word, or character – whatever is most useful. At the extreme, your input is a  $1 \times 1$  ‘table’ whose one cell contains all of your raw data in arbitrary form.

Now you’re just converting one set of tables into another. There are only so many things you can do to tables. The most straightforward things are actions that change at most one dimension – rows or columns – at

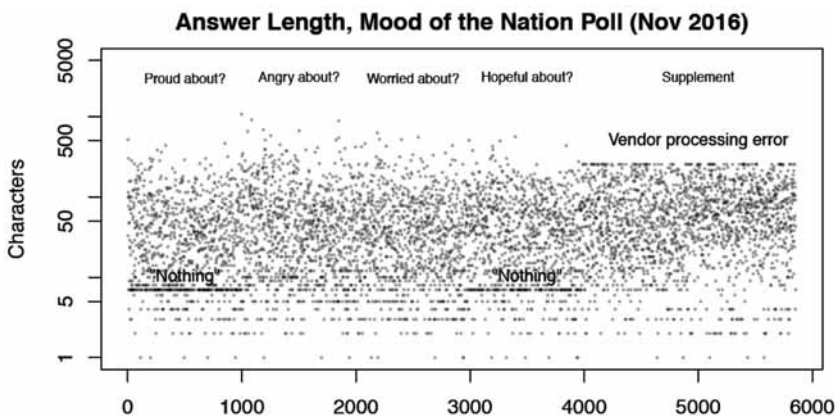
a time. You can delete some rows or columns, making your table shorter or thinner. You can use patterns to split some rows or split some columns, making it taller or wider. You can extract information from some rows or columns to create new ones. You can combine multiple rows or columns to create new ones. You can use patterns to change the contents of a row or column.

What do we mean by ‘use patterns’? In most instances, we mean the use of *regular expressions*, a sort of ur-language with slightly different dialects in different programming languages including R and Python, that is used to specify patterns in strings that you want to find, extract, or replace. Regular expressions push you into puzzle-solving mode, which appeals to some more than others. Computer scientist Jamie Zawinski is credited with an infamous quip: ‘Some people, when confronted with a problem, think “I know, I’ll use regular expressions.” Now they have two problems.’ Regular expressions can be frustratingly unwilling to do what you mean rather than what you say, but there are numerous graphical tools available online to visualize what your regular expression

is actually doing. We can also recommend Wrangler, which dynamically predicts and suggests patterns based on examples you show it (Ratterbury et al., 2017).

One set of exceptions to the ‘usually regular expressions’ advice is when the text in question is in a nested hierarchical format such as JSON, XML or HTML. It is often difficult and occasionally impossible to extract certain patterns from these with regular expressions. There are dedicated and well-built JSON, XML and HTML parsing libraries in both Python and R that make this task much easier.

It is also advisable at various stages to look for patterns that shouldn’t be in your data, and may be the result of errors. This process goes by many names, including *data profiling* and *data editing*. Some basic techniques include looking at the distribution of variables (or derivatives like length of strings) overall and sorted or plotted against other variables like time. Figure 5.4 shows an example we recently faced. Puts et al. (2015) discuss statistical and graphical methods for detecting and correcting errors in large scale datasets.



**Figure 5.4 Profiling of open-ended responses in the Mood of the Nation Poll (Wave 4), run by Penn State’s McCourtney Institute for Democracy, revealed three suspicious patterns, one of which was in fact the result of an error at the vendor (processing in Stata truncated the responses at 244 characters). (The lines at length seven are not errors. In every wave of the poll, the modal answer to ‘What recently in politics/the news made you proud?’ and ‘Looking forward to the next 12 months, what makes you hopeful?’ is ‘Nothing’.)**

Some wrangling changes the shape of your table in more than one dimension. Common variants of this include *pivoting* information from columns to rows, making your table shorter and wider, or *unpivoting* information from rows to columns, making your table taller and thinner. This is one of the more conceptually confusing aspects of data wrangling. We recommend the recent work by the cdata team on *coordinatized data* and *fluid data*, which offers among other things encouragement to worry about this less than we tend to do (Zumel, 2018).

Some wrangling involves using an existing set of tables to create a new table. This is almost always a *grouped summary* that involves some *split-apply-combine* logic, terminology generally attributed to Wickham (2011). In such a task we have, for example, a table with one or more variables we can ‘group by’ and one or more statistics we want to calculate for each group. Conceptually, we split the table into groups, apply our aggregate statistics function to the group, and combine the answers into a new table with one row per group. This is a core operation in SQL, pandas, data.table, dplyr and cdata, and you will see it repeatedly.<sup>21</sup>

Finally, you can combine two (or more) tables into one, merging records that match on some key variable(s). With clean data this is straightforward. With any ambiguity this is one of the most difficult and consequential problems the data scientist will regularly face. The most common social scientific term for the process and problem is *record linkage* (is ‘Angela J. Nunez’ on the voter registration list the same person as ‘A. Nuñez’ in this survey?) but there are variants of the problem called other things, such as *entity disambiguation* or *duplicate detection*. Getoor and Machanavajjhala (2012) provide a good overview of the sources and consequences of this problem. The state of the art (e.g., Enamorado et al., 2019) involves *locality sensitive hashing* and related concepts at the edge of computer science, statistics and social science methodology.

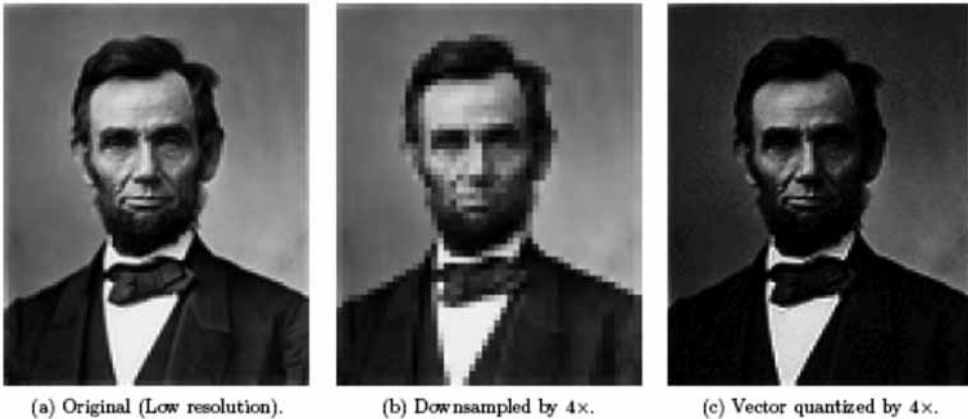
## DATA PREPARATION

While we may now have clean data, structured the way we need it structured, this still may not be exactly the data we want to analyze. There are a number of ways in which we might alter, reduce or expand our data, and a number of reasons we might want to do so. Some of these improve computational properties of models. Some of these improve the inferential or learning properties of models, including regularization, robustness and generalizability. Some of these improve the interpretability or domain relevance of our models. Typically, there are tradeoffs between these objectives.

### *Combining Observations*

Let’s consider the possibility of reducing observations by combining them, creating a new combined unit with a new aggregate statistic calculated from the component observations. In data conceived of as ‘signals’, such as audio and image, this is a form of *downsampling*. This could serve a variety of purposes. It might be a way to ease computation or a way to smooth our data across noisy individual observations. In the case of audio, image or video signals, it might filter out high frequency audio beyond human hearing or images of higher resolution than the phones being used to view them. In more familiar settings where the original observations are people, such aggregation might serve to protect privacy of individuals who could be identified from the disaggregated data.

We’ll use the famous photo of Abraham Lincoln in Figure 5.5a as a running example throughout this section. The raw data depicted in 5.5a is an intentionally low resolution ‘png’ (bitmap) version of the photo in ‘grayscale’. It is a matrix of pixels, 120 rows and 96 columns (96 × 120 in the convention of image descriptions) with each pixel



**Figure 5.5** Downsampling aggregates pixels (observations) into new pixels, reducing the resolution of the image. Vector quantization aggregates the intensities (features) into fewer possible values, reducing the depth of the image

defined by a gray level from 0 (black) to 255 (white).

As can be seen in Figure 5b, downsampling – in this case, combining groups of four pixels into one – can be thought of as lowering the *resolution* of our data. This is one form of *lossy compression*. It is compression, because we have used 1/4 as many numbers (2,880 pixels, down from 11,520), taking advantage of the grid/neighborhood structure of the pixels to approximate the original data. It is lossy, because we can't recover the original data from the compressed version. We now do not need to squint to see the distinct pixels.<sup>22</sup> Aggregation of nested units – as with geographic data nesting people in census tracts, counties in states, etc. – is similarly a sort of lossy compression, but with the added complication that aggregation units are generally not the same size, the same shape, or organized on a grid.

### **Adding Observations**

Conversely, let's consider the possibility of expanding our dataset. In the context of machine learning, *data augmentation* involves adding fake data to your training set.

How could this possibly be either legitimate or helpful? Consider our Lincoln image as itself one piece of a data in a face recognition training set. What if we rotate it clockwise a quarter turn? Wouldn't we want our image classifier to recognize this as Lincoln also? So, we add some rotated, translated and rescaled variants of our training images to help the model learn to recognize such changes.

To make your model robust to noise, you may also wish to introduce some to your training data. This form of data augmentation is now used in *adversarial learning*, where the idea is for the adversarial part of the model to try to trick the learner, making the learner harder to fool (Goodfellow et al., 2016).

An implicitly common, but often unacknowledged, type of regularization can be achieved by adding *pseudo-observations* to your data. A particularly common variant is 'adding one to all my counts before I take the log', often referred to in machine learning as *additive smoothing*. There is an explicitly Bayesian justification for how this works. Many regularization techniques have a direct analogue and interpretation as Bayesian shrinkage priors. The most

well-known are L2-regularization (if applied to the coefficients of a linear regression, this is *ridge regression*) and L1-regularization (if applied to the coefficients of a linear regression, this is the *LASSO*), which are equivalent to the use of a mean zero normal or Laplace distribution prior, respectively. In turn, *conjugate Bayesian priors* have a direct interpretation as pseudo-observations added to the dataset. So, equivalently, adding pseudo-observations is equivalent to (some) Bayesian prior. Monroe et al. (2008) demonstrate the use of this technique. A prominent implicit example from machine learning is the impactful word2vec model for word embeddings. Training involves a technique called *negative sampling*, which adds unobserved word non-cooccurrences to its training data as if observed (Mikolov et al., 2013).

## Altering Features

It is important to recognize that the features in our raw data, even if clean, may not be the best way to *represent* our data. There may be transformations of the data that make it easier for the model to learn, or for us to make sense of what it learns. Of course, even apparently trivial alterations can have a massive impact on what features and observations are similar to each other (Denny and Spirling, 2018).

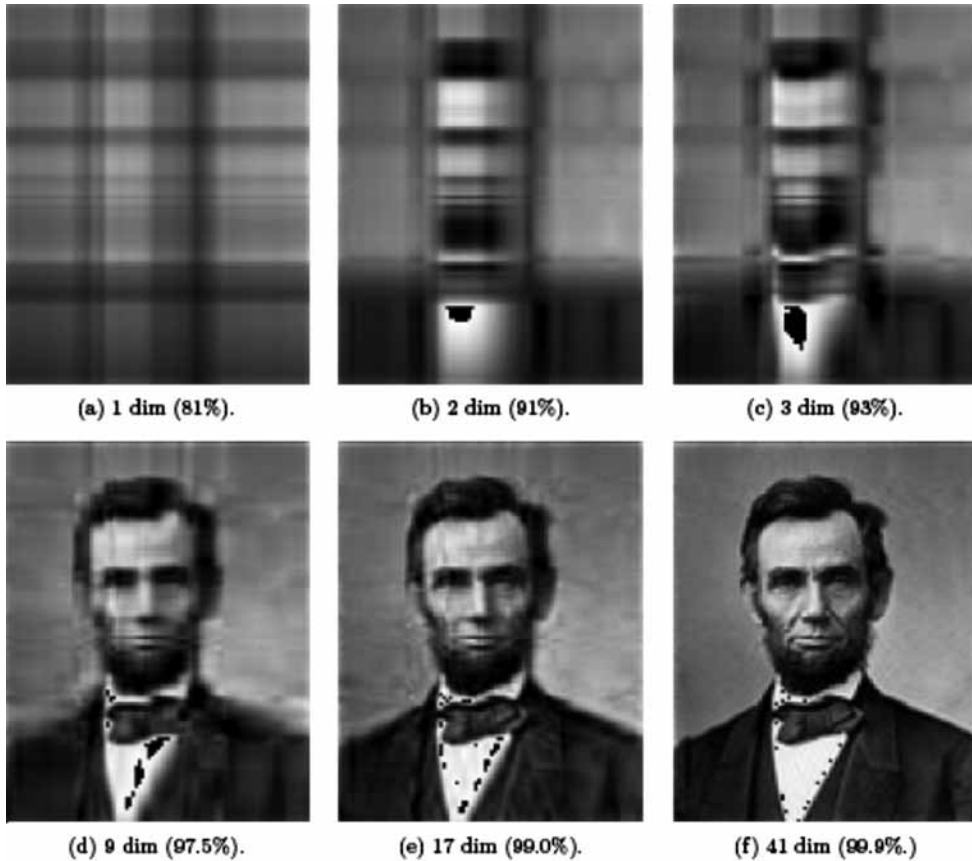
We may, for example, wish to reduce the dimensionality of our feature set, to find a smaller set of features that still provide a reasonable representation of our data. One approach might be *vector quantization*, in which we cluster our observations, summarize each cluster by its mean and then replace each observation with this cluster prototype. If we vector quantize our Lincoln image by a factor of four, we need to cluster the 256-pixel intensities into 64 representatives. The reduced image that results is shown in Figure 5c. The resolution is the same as the original, but now the *depth* has suffered, losing important detail.

An alternative is to decompose the data into a new set of features. One objective of this decomposition might be to reduce redundant information in the features. There are several possibilities for decomposing this matrix, including Singular Value Decomposition (SVD), Principal Component Analysis (PCA), Independent Component Analysis (ICA) and Nonnegative Matrix Factorization (NMF).<sup>23</sup> One characteristic of SVD (and PCA) is that the order of the new features reflects their importance in capturing variation in the original matrix, so we can take the first  $k$  new features to get the best  $k$ -dimensional recreation of the picture, in this sense. Figure 5.6 shows the quality of the representation at different levels of  $k$ .

If your data has some structure that implies connectedness among data points – such as sequential structure, as with time series, audio or natural language, or two-dimensional spatial structure, as with geographic or image data – you may benefit from the application of *convolution* to your data or in the structure of your model. The specific functional form of convolution applied or learned – the function that *is convolved with* your data – is referred to as a *kernel* (or in some application areas a *filter*). The math of convolution can be intimidating, but the idea is simple and you have undoubtedly encountered it before, whether you knew it or not, most likely as some form of *smoothing*.

A kernel is just a function; a smoothing kernel is just a probability distribution function, often a normal (Gaussian) distribution. An easy example is as in Figure 5.7, where we have a set of  $N$  integer-valued observations, in this case the 11,520 grayscale values of 0 to 255 in the original Lincoln image. Loosely speaking, to convolve a kernel with such data, you replace every data point with a  $1/N$  sized copy of the kernel and add them up. Essentially each observed data point melts into neighboring regions of the space, providing a smooth version of a histogram, or *kernel density* plot. You have probably seen the same idea applied for *kernel smoothing* of





**Figure 5.6** Effect of dimension reduction through singular value decomposition (SVD)

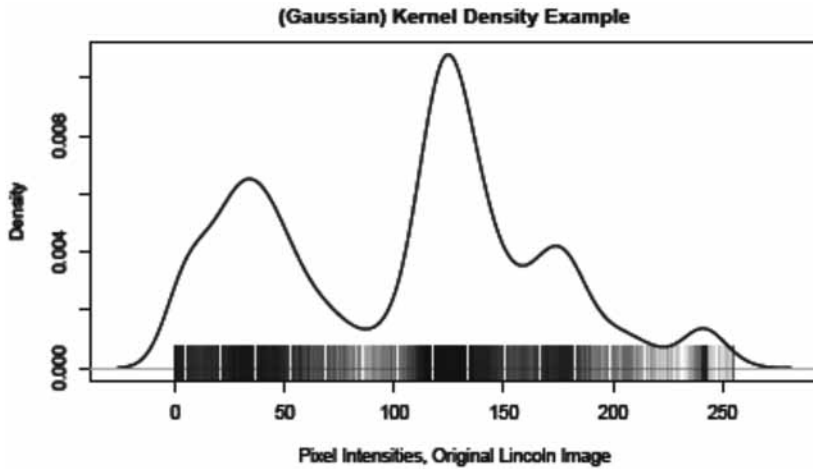
*Note:* Captions indicate the number of dimensions retained (from an original with 96 dimensions) and the percentage of variance in the original's pixel intensities that is captured.

geographic point data. Each discrete observation is melted across its (two-dimensional) spatial neighbors, providing a smoothed version of the data.

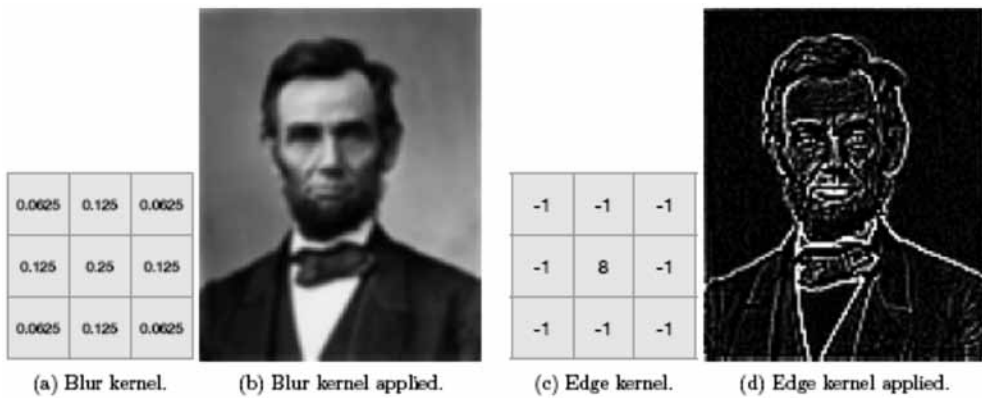
Convolution is much more general than smoothing, however – a point again most easily demonstrated in image data. We can define a simple *image kernel* or *filter* as a  $3 \times 3$  matrix of numbers like that in Figure 5.8a. We will take this kernel and slide it like a window over each  $3 \times 3$  block of pixels in the original Lincoln image (Figure 5.5a), for each replacing the central pixel's value with a sum of the nine pixel values in its window, weighted as defined by the kernel. This

particular kernel is similar to our Gaussian from before, smoothing neighboring values toward each other and creating a 'blur' effect as seen in Figure 5.8b.

Now consider the kernel in Figure 5.8c. This kernel brightens pixels that are brighter than surrounding pixels, where there is a sudden change in values. This makes it an edge detector, as shown in Figure 5.8d. Convolutional neural nets work, in part, by learning multiple such filters to apply to data like pixel values to infer, for example, different combinations of edges in different directions sufficient to distinguish faces or particular objects.



**Figure 5.7** A kernel density estimation of the Lincoln images pixel intensities convolves a Gaussian curve (normal distribution) over the observed distribution of values



**Figure 5.8** Effect of convolutional image kernels

**RESPONSIBLE DATA SCIENCE PROJECTS**

Data science is a new field that combines elements from many existing disciplines. One downside to this novelty is that scientific and ethical standards have not been clearly defined by the community. And, of course, there are broader public concerns about corporate and government practices in the collection of data, the protection of data and the social

implications of machine learning, artificial intelligence and other data science techniques. In this section, we discuss three considerations that arise when trying to design responsible data science projects: reproducibility, data privacy and algorithmic bias.

**Reproducibility**

Concern about the reproducibility of findings has become a defining feature of social

science and, indeed, science in general in recent years. This has produced laudable progress, accelerated by policies of journals and funding agencies, in the development of standards and mechanisms for things like study preregistration and public archiving of replication data and code. Against that background, the data-intensive projects pursued by data scientists present unique hurdles to transparent and reproducible research.

One might consider the bare minimum to be posting the data necessary to replicate your analysis on dataverse or a similar publicly accessible archive, but this can be impossible in many cases. First, the data may be too large for permanent storage with sufficiently low latency network bandwidth to be technically or financially feasible. Even more frequently, the data has inherent value to a commercial owner, and sharing it would violate copyright or other legal constraints, licensing agreements or terms of service. One reasonable way of working around this is by sharing *dehydrated data*, i.e., providing pointers and scripts that can be used by the replicating user to recreate and collect the full data themselves, given sufficient time and access. A very nice example of this is the Documenting the Now project (Clark et al., 2019), which archives collections of tweets around important social events (such as Black Lives Matter protests or the Women's Marches) and shares them as sets of dehydrated tweet ids. They also provide a variety of tools, easily incorporated into replication pipelines, for rehydrating these data consistent with Twitter's terms of service and other policies.

There are also emerging academic–industry partnerships designed to provide academic researchers with controlled access to, and scientifically acceptable mechanisms for both publishing and replicating with, proprietary data. Most notable is the Social Science One project facilitated by Harvard, the SSRC and eight foundations and currently partnering most prominently with Facebook (King and Persily, n.d.).

As we have seen, data science projects, almost by definition, do not consist of one tabular dataset and one analysis script. Instead, there is often extensive data wrangling and feature engineering performed just to arrive at a point where analysis is possible. These are consequential parts of a data science project and they should be subjected to transparency and replication standards as well. It is ideal that you provide potential replicators not just the analysis-ready data from the end of the pipeline, but data in its intermediate forms back as close as possible to the raw data ingested at the beginning of the pipeline, along with as much as possible of the pipeline instantiated in code.

From here, the data science community and literature are not entirely settled on what approach best meets the ideals of transparency and reproducibility. At one extreme, the ideal is one-button replication. In this view, everything should be provided sufficient for a user to very simply execute the full pipeline on the original data and receive an identical result. Taken to its logical limit, this can best be guaranteed by providing production-level software encased in a virtual copy of the original computing environment via something like a *Docker container*.

At the other extreme, the idea is to expose the bones of the pipeline and the decisions that went into it, with the objective that a user can develop intuition and adapt the code to work with their data or to check the robustness of pipeline choices. Here the demands are for heavy documentation, perhaps even of bits of code and data ultimately not used in the final pipeline, and highly flexible, even interactive, code. Taken to the logical extreme, this can best be achieved by instantiating your code and documentation in literate programming tools like Jupyter or R notebooks.

Some emerging paradigms and accompanying tools strike a balance between these two motivations under the umbrella of *workflow management* systems. Our current favorite is *Snakemake*, which provides a

very readable Python-like language and tool for managing and executing data pipelines. It can incorporate both R and Python scripts, can recreate virtual computing environments via conda for Python or packrat for R and can work with data and other resources on a variety of cloud platforms and through a variety of channels. So a user can execute a more or less exact replication, but also has ready access to the underlying scripts and how they fit together in the pipeline, and can change them as desired. Another emerging workflow management system, which holds considerable promise for tracking *data provenance* directly within data itself, is *Pachyderm*, which has been conceived as ‘git for data’.

### **Data Privacy**

When using large socially generated data sources, projects will eventually brush up against privacy concerns that push in the opposite direction of transparency and open data impulses. We may sincerely wish to protect the privacy of individuals in our data, but naive ‘anonymization’ – just removing names and other direct personal identifiers – isn’t enough. We can remove names from a survey but a combination of demographic and other features may be sufficient to tie a person uniquely to their answers. We can remove names from a GPS tracking dataset, but easily deanonymize most people by the location where their tracks tend to end at night and start in the morning.

The dominant contemporary framework for thinking about statistical data privacy is *differential privacy*. It is a highly technical literature, closely related to cryptography, but the high-level intuition is straightforward: an individual should face (almost) no risk from participating in a statistical database. A differentially private algorithm applied to a dataset seeks to provide a statistical guarantee to any individual appearing in the dataset that the answer from the algorithm will allow an adversary to learn essentially no more

about that individual than they could learn if the individual were not in the dataset. The simplest differentially private mechanisms add noise of sufficient magnitude and type to meet the needed guarantee. Among the more interesting approaches is using the actual data to generate *synthetic data* that has maximal utility in terms of capturing the statistical properties of the original, subject to the differential privacy constraint.

And, of course, data release can have unintended consequences for privacy over and above even the strict requirements of differential privacy. A striking recent example was the release by Strava of their ‘global heat map’. Strava is a fitness social networking site, where users upload tracks from their phones or other GPS devices of runs, bike rides, etc. The heatmap showed all of Strava’s users’ GPS tracks in aggregated form. Given that tracks merge into one another and had no times attached, it would be very difficult to learn much about the paths of any individual. But observers were quick to locate regular ovals and similar tracks where there shouldn’t be any, such as in a remote part of Mali. These were the traces of US soldiers going on regular jogs around undisclosed military bases. That is, by using a fitness tracking app, soldiers unwittingly released sensitive military information and, arguably, compromised their own safety.

### **Algorithmic Bias**

Many people have the unfortunate impression that algorithms produce an ‘objective’ understanding of the world around us, distinct from our own human biases. This is, of course, untrue. Models are designed to find patterns in data, and, in a biased world, those patterns will reflect that bias and often worse. Say we want to train an algorithm to classify resumés into ‘would hire’ and ‘wouldn’t hire’ categories. If we train it on past ‘did hire’ and ‘didn’t hire’ data, and hiring practices in the past displayed racial bias, the algorithm will

seek out features such as names or zip codes that not only encode that bias, but use it to make recommendations that repeat or even amplify that bias.

There is much work being conducted on debiasing algorithms, but it's an elusive concept (Caliskan et al., 2017). For example, Bolukbasi et al. (2016) find that word embedding models encode gendered usage of words, placing the vector for the word 'woman' closer to that of 'homemaker' and placing the vector for 'man' closer to 'programmer'. This is useful as a measurement of the underlying societal bias. This is obnoxious if it underlies a black box algorithm recommending career advice websites to our children. It is similarly obnoxious that such models place 'man' nearer to 'doctor' and 'woman' nearer to 'nurse', particularly given that we just framed words as job titles. But framed as verbs perhaps we should see a difference, given that either a man or a woman can doctor a test result, but only a woman can nurse a baby.

This is another reason why interpretability of models is so important. These subtleties are difficult enough to unparse when they're isolated like this. They are more or less impossible to identify when the biases in the data are unexamined or the impact of biased data on our algorithmic task is black-boxed.

## CONCLUSION

The sheer breadth of subjects discussed here hopefully makes clear that the prospective data scientist needs to be comfortable with *team science*. The scope of data science projects, and the number and variety of specialized tasks a typical project requires, very often necessitates a collaborative interdisciplinary research team. Interdisciplinary collaboration is a learned skill that requires patience. Other disciplines have different assumed knowledge, different conference and publishing cultures, different scientific

questions, different approaches to learning from data and jargon just different enough to create abundant opportunities for confusion. Our best advice is that when you inevitably reach a point of mutual confusion with your potential new interdisciplinary colleague, stay in the room. All the good stuff is after that.

We also want to re-emphasize that calling the process 'data science' does not exempt us from social scientific and ethical standards (Monroe et al., 2015). While the practice of data science is often associated with early stage Facebook's infamous internal ethos 'move fast and break things', we less snap-pily encourage the social scientific data scientist to move purposefully, learn things and play nice.

## Notes

- 1 The senior author insists that 'data' is, like information, an abstract noun that *is*, and like sugar or sand, a collective noun that *is*. Our agenda are too full here to elaborate at length.
- 2 These may be literally instantiated in code with a *pipe operator*, like `|` in a Unix shell or `%>%` in R (through `magrittr`).
- 3 This, and all of our specific advice regarding 'tools', will no doubt eventually read as dated. This reflects a snapshot as of this writing in 2019.
- 4 Arguably, we should perhaps also include Java and (Java-based) Scala. Many 'big data' tools – e.g., Hadoop, Hive, Flink, Elasticsearch, Spark ... more than half of the projects under the Apache umbrella – are written in Java and/or Scala, as are many established machine learning, NLP and similar libraries – e.g., Weka, Mahout, CoreNLP, Deeplearning4j. The learning curve is steeper, however, and the advantages of Java reveal themselves mainly in the context of modular software engineering for large multi-person/multi-team industry settings. Viewed as tools for social scientific data science, most of these can be interfaced from Python and increasingly R. (There are other possibilities, of course – Julia, MATLAB/Octave, Haskell, etc. – but none of these yet has a footprint in data science, much less social science, to rival Python or R.)
- 5 Scholars working on tasks that require intensive numerical computations may benefit from learning a *compiled language* like C++, Cython or

- Rust. Many common R and Python libraries, as well as R and Python themselves, call such code ‘under the hood’ for speed purposes. This is useful in that context, but the learning curve is considerably steeper.
- 6 For example, the one-liner `curl http://www.gutenberg.org/files/28885/28885-h/28885-h.htm | tr '[:punct:]' ' ' | tr 'A-Z' 'a-z' | tr -s ' ' | tr ' \n' | sort | uniq -c | sort -rn` retrieves the text of *Alice in Wonderland* (in a few seconds) and outputs a sorted table of word frequencies from it (instantly).
  - 7 There is an additional practical benefit, as increasingly one’s github account serves as part of a job portfolio.
  - 8 Although this is a common characterization in textbooks, it is widely understood in measurement theory that this typology is insufficient. For example, it does not distinguish between ordinal and count data, and has no category for percentage (or ‘counted fractions’ or ‘gradation of membership’) data (Mosteller and Tukey, 1977).
  - 9 A similar-looking problem occurs with scrapers ignoring *html entities* leaving text littered with things like ‘&nbsp;’ or ‘&quot;’ (the telltale signs are the opening ampersand and concluding semicolon). But this is unrelated to the encoding, and best dealt with at the scraping stage with an HTML specific parser like the Python library BeautifulSoup.
  - 10 Loosely speaking, this is a ‘data.frame’ in base R, a ‘tibble’ in the R tidyverse, a ‘data.table’ in R’s data.table package and a ‘DataFrame’ in Python’s pandas, although some of these can hold more general structures. In different contexts, these types of structures are also known as ‘attribute-value representations’, ‘flat data’, ‘object-predicate tables’, ‘Aristotelian data’ and more.
  - 11 There are several ways to do this. For example, Python’s scipy offers seven distinct sparse matrix formats, each with particular strengths and weaknesses.
  - 12 [https://en.wikipedia.org/wiki/List\\_of\\_file\\_formats](https://en.wikipedia.org/wiki/List_of_file_formats).
  - 13 API stands for ‘Application Programming Interface’. It is a general concept in computer programming, indicating a set of methods for communication between two distinct software components. But in the data science context ‘API’ almost always refers to a web-based service for accessing underlying data. Twitter, for example, provides several APIs that can be called by apps or scripts to retrieve (limited) data about tweets using certain search terms or recent history of particular users.
  - 14 A third human-readable text-based format, YAML (too cutely by half, ‘YAML Ain’t Markup Language’) is arguably even more readable and compact than JSON, as well as being more flexible, but is not yet very widely used as a general data exchange format.
  - 15 A folk quote attributed to dozens of different sources.
  - 16 Technically, ‘third normal form’, which is a bit much for us here.
  - 17 It does help tremendously if filenames serve as unique identifiers, as well as sort in a way that is meaningful. This means if you number files, use enough leading zeroes to avoid ‘file1.img’ → ‘file10.img’ → ‘file2.img’; if you date files use ‘YYYY-MM-DD’ or similar format.
  - 18 Students often guess this is related to Ramadan or the Hajj, but those are on the 354-day cycle of the Islamic calendar, and out of phase with the dynamics observed here.
  - 19 Terms highly correlated in weekly search volume with ‘islam’ include: ‘judaism’ ( $\rho = +.80$ ), ‘christianity’ (+.80), ‘greek art’ (+.79), ‘buddhism’ (+.79), ‘hinduism’ (+.78), ‘essays’ (+.78), ‘roman art’ (+.78), ‘voltaire’ (+.78), ‘chaucer’ (+.78), ‘articles about’ (+.77), ‘uses of the’ (+.77), ‘aristotle’ (+.77), ‘summary of the’ (+.77), ‘interpretation of’ (+.77), ‘chemical elements’ (+.76), ‘international relations’ (+.76).
  - 20 <https://books.google.com/ngrams>.
  - 21 This is also closely related to the functional programming notions of *map* and *reduce* as well as the MapReduce logic underlying the distributed processing of Hadoop.
  - 22 This pixel aggregation is a naive way to downsample an image. There are other ways to downsample images that minimize these sorts of visual artifacts.
  - 23 If we were taking this seriously as image analysis, we would be more likely to use a Fourier Transform or similar.

## REFERENCES

- Bolukbasi, Tolga, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama and Adam T Kalai. 2016. ‘Man is to computer programmer as woman is to homemaker? Debiasing word embeddings.’ In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabell Guyon, and Roman Garnett (Eds) *Advances in Neural Information Processing Systems 29, Proceedings of the 30th Annual Conference on Neural Information Processing Systems*, Barcelona, Spain. pp. 4349–4357.
- Caliskan, Aylin, Joanna J. Bryson and Arvind Narayanan. 2017. ‘Semantics derived

- automatically from language corpora contain human-like biases.' *Science* 356(6334): 183–186.
- Clark, Meredith, Bergis Jules and Trevor Muñoz. 2019. 'Documenting the now.' <https://www.docnow.io>.
- Denny, Matthew J. and Arthur Spirling. 2018. 'Text preprocessing for unsupervised learning: why it matters, when it misleads, and what to do about it.' *Political Analysis* 26(2):168–189.
- Enamorado, Ted, Benjamin Fifield and Kosuke Imai. 2019. 'Using a probabilistic model to assist merging of large-scale administrative records.' *American Political Science Review* 113(2):353–371.
- Getoor, Lise and Ashwin Machanavajjhala. 2012. 'Entity resolution: theory, practice & open challenges.' *Proceedings of the VLDB Endowment* 5(12):2018–2019.
- Goodfellow, Ian, Yoshua Bengio and Aaron Courville. 2016. *Deep Learning*. Cambridge, MA: MIT Press.
- Kelley, Truman L. 1927. *Interpretation of Educational Measurements*. Yonkers-on-Hudson, NY: World Book Co.
- Kimball, Ralph and Joe Caserta. 2004. *The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data*. Indianapolis: Wiley.
- King, Gary and Nathaniel Persily. (n.d.). A New Model for Industry–Academic Partnerships. *PS: Political Science & Politics*, 1–7. doi:10.1017/S1049096519001021
- King, Gary, Patrick Lam and Margaret Roberts. 2017. 'Computer-assisted keyword and document set discovery from unstructured text.' *American Journal of Political Science* 61(4):971–988.
- Klein, Shmuel Tomi. 2016. *Basic Concepts in Data Structures*. Cambridge: Cambridge University Press.
- Lazer, David, Ryan Kennedy, Gary King and Alessandro Vespignani. 2014. 'The parable of Google Flu: Traps in big data analysis.' *Science* 343(6176):1203–1205.
- Linder, Fridolin J. 2017. 'Improved data collection from online sources using query expansion and active learning.' <http://dx.doi.org/10.2139/ssrn.3026393>.
- Mason, Hilary and Chris Wiggins. 2010. 'A taxonomy of data science.' <http://www.dataists.com/2010/09/a-taxonomy-of-data-science/>.
- Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado and Jeff Dean. 2013. 'Distributed representations of words and phrases and their compositionality.' In Christopher J. C. Burges, Léon Bottou, Max Welling, Zoubin Ghahramani, and Kilian Q. Weinberger (Eds.) *Advances in Neural Information Processing Systems 26, Proceedings of the 27th Annual Conference on Neural Information Processing Systems*, Lake Tahoe, Nevada. pp. 3136–3144.
- Mohammad, Saif, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu and Colin Cherry. 2016. 'A dataset for detecting stance in tweets.' In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Portorož, Slovenia, May 2016. pp. 3945–3952.
- Monroe, Burt L., Michael P. Colaresi and Kevin M. Quinn. 2008. 'Fightin' words: lexical feature selection and evaluation for identifying the content of political conflict.' *Political Analysis* 16(4):372–403.
- Monroe, Burt L., Jennifer Pan, Margaret E. Roberts, Maya Sen and Betsy Sinclair. 2015. 'No! Formal theory, causal inference, and big data are not contradictory trends in political science.' *PS: Political Science & Politics* 48(1):71–74.
- Mosteller, Frederick and John Wilder Tukey. 1977. *Data Analysis and Regression: A Second Course in Statistics*. Boston: Addison-Wesley.
- Pechenick, Eitan Adam, Christopher M. Danforth and Peter Sheridan Dodds. 2015. 'Characterizing the Google Books corpus: strong limits to inferences of socio-cultural and linguistic evolution.' *PLoS ONE* 10:e0137041.
- Peuquet, Donna J. 2002. *Representations of Space and Time*. New York: The Guilford Press.
- Puts, Marco, Piet Daas and Ton de Waal. 2015. 'Finding errors in big data.' *Significance* 12(3):26–29.
- Ratterbury, Tye, Joe Hellerstein, Jeffrey Heer, Sean Kandel and Connor Carreras. 2017. *Principles of Data Wrangling*. Sebastopol, California: O'Reilly Media.
- Salganik, Matthew J. 2017. *Bit by Bit: Social Research in the Digital Age*. Princeton, NJ: Princeton University Press.

- Smith, Christopher. 2013. 'Anti-Islamic sentiment and media framing during the 9/11 decade.' *Journal of Religion & Society* 15:1–15.
- Stephens-Davidowitz, Seth. 2014. 'The cost of racial animus on a black candidate: evidence using Google search data.' *Journal of Public Economics* 118:26–40.
- Stevens, S. S. 1946. 'On the theory of scales of measurement.' *Science* 103(2684): 677–680.
- Thompson, Steven K. 2012. *Sampling*. 3rd edition. Hoboken, NJ: John Wiley & Sons.
- Thorndike, Edward L. 1904. *An Introduction to the Theory of Mental and Social Measurements*. New York: The Science Press.
- Webb, Eugene J., Donald T. Campbell, Richard D. Schwartz and Lee Sechrest. 1966. *Unobtrusive Measures: Nonreactive Research in the Social Sciences*. Chicago: Rand McNally.
- Wickham, Hadley. 2011. 'The split-apply-combine strategy for data analysis.' *Journal of Statistical Software* 40(1):1–29.
- Zumel, Nina. 2018. 'Fluid data reshaping with "cdata": beyond pivot and unpivot.' <https://doi.org/10.5281/zenodo.1173299>.





# Designing Qualitative Research Projects: Notes on Theory Building, Case Selection and Field Research

Ezequiel González-Ocantos

## INTRODUCTION

Why did World War I break out in 1914? Why did states create the International Criminal Court, and what explains the choice of certain design characteristics over others? Why did Mexico democratize in 2000 after 70 years of one-party rule? Why did Britain, Germany and Sweden adopt different types of welfare states? Why did Argentina and Peru prosecute and punish former dictators following their most recent episode of democratization whereas Brazil did not? Qualitative researchers in political science and international relations are usually interested in explaining outcomes such as these. Instead of seeking to estimate the average effect of a causal factor or independent variable in a population, our projects focus on puzzling within-case transformations, or variation in events and institutions across a limited number of cases. Qualitative empirical research thus proceeds to unpack the historical, social and political processes that cause

these effects. To be sure, the puzzle-driven nature of qualitative scholarship means that our projects tend to be highly contextual and, more often than not, research questions feature proper names. The goal, however, always remains to contribute to the general understanding of the conditions favoring classes of events or outcomes, such as inter-state conflict, sovereignty-constraining forms of international cooperation, democratization, redistributive public policies or progressive judicial activism.<sup>1</sup>

Outcome-oriented qualitative research leverages within-case evidence to probe the theoretical models proposed to answer these questions. While cross-case comparisons are often important features of qualitative designs, the main source of inferential power comes from the in-depth tracing of processes in each case (Goertz and Mahoney, 2012). Instead of devising methods to systematically obtain comparable measures of a host of variables across a large number of cases, and using statistics to estimate correlations or the

average effect of a variable, we rely on multiple, context-specific data sources to uncover pathways linking independent and dependent variables (Brady et al., 2004). For qualitative researchers, to explain is to show ‘how’. And this usually requires diving into a case to carefully document the existence of an uninterrupted chain of events or actions and reactions that together bring about the outcome of interest. We thus trade the appeal of generalizations based on finding regular conjunctions of causes and effects in a population, to gain relative certainty that a plausible connection exists between cause and effect in a limited number of instances (George and Bennett, 2005; Beach and Pedersen, 2013; Bennett and Checkel, 2015). As we shall see below, models guide the choice of evidence and the selection of relevant rival explanations for consideration in each case, with the goal of reconstructing these causal chains in full.

How do we go about designing projects to understand outcomes in specific cases, trace processes and thus meet the aforementioned explanatory standards? In this chapter I draw from my experience to answer this question. I argue that it is essential to begin the research by proposing a theoretical model of the sequence or pathway linking cause and effect in positive cases. This sequential account should identify the *steps in the causal chain* and *mechanisms* that explain the flow of causal energy that brings about the outcome of interest. I also suggest that doing so involves providing a stylized description of a ‘field’ of action on the basis of three parameters: the *actors* involved, with their preferences, logics of action and resources; the *spaces of action* where these actors act; and the consequences of the *interactions* between them.<sup>2</sup> After conceptualizing these three parameters, and illustrating them with examples, I contend that this type of theorization can help researchers discipline three additional components of the research process. First, I discuss how sequential models serve as roadmaps during fieldwork. Second, I suggest that sequential models facilitate the

identification of relevant rival explanations. Third, I show how they can assist the case selection strategy. Toward the end, the chapter makes it clear that while well-specified theoretical priors are crucial for the success of qualitative research projects, induction and the dialogue between theory and data also play a critical role.

## CAUSAL PATHWAYS

The method of within-case analysis used by outcome-oriented qualitative researchers is often referred to as ‘process tracing’. According to George and Bennett (2005: 206) process tracing ‘attempts to identify the intervening causal process – the causal chain and causal mechanism – between an independent variable (or variables) and the outcome of the dependent variable’. In Mahoney’s (2012) formulation, process tracing is about documenting sequences: ‘process tracing can help a researcher establish that: (1) a specific event or process took place, (2) a different event or process occurred *after* the initial event or process, and (3) the former was a cause of the latter’ (571; emphasis in original). As Collier (2011: 823) suggests, this involves the ‘systematic examination of diagnostic evidence selected and analyzed in light of research questions and hypotheses’. In this sense, Bennett (2015) and Beach and Pedersen (2013) have shown that process tracing is governed by a form of Bayesian probabilistic reasoning which uses the tools of logic and deep knowledge of context to evaluate the inferential power of individual ‘causal process observations’ (Brady et al., 2004: 252). The method ‘emphasizes that the probative value of evidence relative to competing explanations is more important than the number of pieces of evidence’ (Bennett and Checkel, 2015: 16).

These definitions highlight three essential features of process tracing *qua* method.

First, the goal of process tracing is to evaluate hypotheses. These are not hypotheses about average independent effects or about the constant conjunction of cause and effect in a population. Instead, process tracing involves the evaluation of hypotheses about the presence or absence, in a specific case, of the constituent steps of a stylized causal pathway. Process tracing therefore results in a narrative structured by theory, variables, competing explanations and concomitant observable implications. Second, process tracing primarily involves the quest for causal processes or connections. It is grounded in the assumption that although these processes may be part of the unobservable ontological 'deep', they are still traceable because they leave fingerprints in reality. In fact, explanations that do not account for processes, and therefore do not transcend correlational statements, are incomplete and unsatisfactory (Bunge, 1997; Waldner, 2007). Finally, good description is essential for effective process tracing (Mahoney, 2012). Researchers must strive to describe the component parts of the hypothesized causal pathway in their cases. On their own, these descriptive inferences are not 'causal', but when looked at as a system and ordered in a temporal sequence, and when logical links that comport with theory can be established between them, they allow researchers to put together convincing stories about how causal energy flows from independent variables to outcomes.

The causal pathways that are the object of process tracing have two components: steps in the causal chain and mechanisms (Waldner, 2014). Steps refer to the theoretically relevant events that need to be obtained in order to move from the causal condition to the outcome of interest, or to observe a change in the dependent variable. The different events included in the model must be able to provide a logical or internally sufficient account of the states of affairs, decisions or actions required to produce the transformation under study. For example, if we want to explain a democratic transition,

these could include deteriorating economic conditions, followed by a split in the ruling coalition, followed by mass protests, and so on. Mechanisms, by contrast, refer to what 'makes the system in question tick' (Bunge, 2004: 182). In social science this usually points to the reasons why actors behave in certain ways, the 'ultimately unobservable physical, social, or psychological processes through which agents with causal capacities operate' (George and Bennett, 2005: 137). Mechanisms account for the progression of a case along the chain of steps, and specify why actors or other entities transmit 'either a physical force or information that influences the behavior of other agents or entities' further down the line (Waldner, 2012: 76). As Hedström (2008: 320) puts it, in mechanistic explanations 'the focus is not on relationships between variables, but on actors, their relationships, and the intended and unintended outcomes of their actions'. In the above example, a mechanism could be the calculus that, in light of changing economic conditions, induces soft-liners to cause a split in the ruling coalition, or that leads trade union leaders to stage protests once they observe a divided government.

Researchers are advised to 'draw' causal pathways using flowcharts, with boxes representing the events, and arrows representing the mechanisms that connect them and indicate the direction of causality.<sup>3</sup> In my view, this is a necessary first step in any outcome-oriented qualitative project. While methodologists have identified different types of process tracing, including theory *building* and theory *testing* process tracing (Beach and Pedersen, 2013), a project cannot start without a clearly defined set of theoretical priors. Even when building theory, the researcher should arrive in the field with a preliminary map of the events, actions and reactions that she thinks could account for the connection between the independent variable and the outcome. Without the guidance of a proto-theory of the causal pathway at play that contains relatively well-specified steps and

mechanisms, it will be impossible to know what evidence to look for, distinguishing the accessory from the essential, and to weight the explanatory value of that evidence *vis-à-vis* alternative accounts.

Having to come up with a theory or model that specifies the causal pathway linking the independent variable and the outcome of interest sounds intimidating. To make matters worse, very little has been written in terms of practical advice on ‘how to theorize’ (e.g. Kohli et al., 1995; Lake and Powell, 1999; Gerring, 2001; Coppedge, 2012). In what follows I will try to demystify this crucial component of the research process by providing a series of guidelines that have helped me in the past.

## MODELS AS FIELDS

A ‘theory is a mental model designed to make sense of reality: to describe and explain or predict what we observe’ (Coppedge, 2012: 49). In order to be useful and portable heuristic devices, theories and models need to simplify reality and be pitched at a high level of abstraction. But we never theorize or model out of thin air. This is especially true in outcome-oriented qualitative research, where theories or models are usually mid-range, highly bounded and narrow in scope. We always have some level of familiarity with the countries or institutions we want to explain, and this knowledge inevitably informs theoretical assumptions about the nature of the causal pathways at play. In fact, one can think of theories or models as a transparent exposition of these priors, or what we think is going on in our universe of cases. We then probe these priors with new evidence from the cases. This does not mean, however, that our theories or models will be necessarily idiosyncratic, or cases ‘minus the proper names’. Any mid-range theory is also informed by, and anchored in, more general assumptions about how the world works,

derived from other theories, documented regularities in human action or cognition and common sense. This kind of integration makes theorizing less daunting, facilitates knowledge accumulation and adds intelligibility to our arguments within a broader scientific community (Coppedge, 2012: chapter 3). Crucially, it also renders our models testable in cases that do not constitute the direct object of the research, but are part of the relevant universe.

I like to think of the models that inform qualitative research projects as stylized descriptions of the *field* of action in which our outcomes of interest are generated. I borrow the term ‘field’ loosely from Pierre Bourdieu (e.g. Bourdieu and Wacquant, 1992: 94–110). Bourdieu uses the term to refer to distinctive social domains that emerge throughout history, such as the artistic field, the academic field or the political field. It is useful to think of a field as a space of action. These spaces have certain morphological and institutional features. For example, in the artistic field you have museums, auction houses, foundations, academies, art studios and so on. Spaces also house a variety of actors, such as artists, critics, art collectors and museum curators, which take different positions in the field and play different roles in the game that determines artistic success. The agents that inhabit a field have varying resource endowments, ‘whose possession commands access to specific profits that are at stake in the field’ (ibid: 97). Artists can paint, foundations have money and critics have authority. Importantly, not all agents have enough resources to secure goals, and most importantly, not all of them have the right kind of resources to succeed. In fact, fields favor the possession of some forms of capital over others. In the case of artists, the types of skills that can be converted into success very much depend on the taste of rich collectors and influential critics. Depending on the state of the field at any given point in time, the master of still life may or may not become a name. There are also formal and informal rules that govern behavior inside

the space. The field thus ‘guides the strategies whereby occupants of those positions seek, individually or collectively, to safeguard or improve their position in this field’ (ibid: 101). In this sense, the field structures behavior, but is also a site of ‘struggles aimed at preserving or transforming’ the rules (ibid: 101). For instance, marginalized artists may try to court critics to transform the dominant sense of taste. Finally, the boundaries that delimit the field exclude other spaces and actors, such as the millions of tourists that pack museums day in and day out but have little impact on the dynamics of artistic success.

Theorizing is therefore the process whereby we specify the boundaries and characteristics of the field in which the outcome we seek to explain is generated. Models indicate what is relevant to understand a particular phenomenon and what is not. The political world is not obviously compartmentalized into fields, so we need to bring order to it by thinking through, and justifying, our assumptions about what our domain of interest looks like and how it works. To do so, it is important to specify three key parameters of the field of action. These parameters suggest who and what matters to explain the outcome. First, the model needs to identify the *actors*, as well as their preferences, logics of actions and resources. Some actors will be considered relevant, whereas others will be considered irrelevant in the process that engineers the outcome. Similarly, some resources are crucial for them to become competent and effective players, while others are completely superfluous and therefore orthogonal to the explanatory effort. Second, the model needs a clear conceptualization of the *spaces* in which these actors act. This involves making a choice about which are the venues where the relevant developments unfold. It also entails describing the main characteristics of those venues, including the scripts, rules, vocabularies and resources relevant to navigate through such spaces, and the institutional features that constrain or enable behavior. And finally,

the model should include an account of the *interactions* between the actors, the effect of these exchanges on other actors and the space of action, and how interactions permit the flow of causal energy. Interactions could be of a strategic nature, meaning that actors with stable preferences make choices in light of what they anticipate others will do (Lake and Powell, 1999). They could also be constitutive, for example, when some actors behave in ways that fundamentally transform the preferences, beliefs and capabilities of others. Collectively, these three parameters – actors, spaces and interactions – form the basis of any explanation for why and how we get from an antecedent causal factor to the outcome of interest.

Table 6.1 summarizes the key components of the ‘field’ approach, and lists a series of questions to guide the specification exercise. In what follows I illustrate how I used this guide in my own research (González-Ocantos, 2016). I will then discuss the implications of my modeling choices for other components of the research design, including data collection, identification of alternative explanations and case selection.

## SPECIFYING THE FIELD

In *Shifting Legal Visions: Judicial Change and Human Rights Trials in Latin America* (González-Ocantos, 2016), I investigate the determinants of judicial behavior during the wave of human rights trials that swept Latin America beginning in the 1990s. Why did Argentina and Peru experience waves of human rights prosecutions after democratization, whereas Mexico did not? To answer this question I trace the processes whereby litigants and other civil society actors diffused knowledge of international human rights law as well as new professional role conceptions among local judges, thus enabling the successful prosecution of human rights criminals – including former dictators.

**Table 6.1 Specifying the field**

<i>Components of the field</i>	<i>Guiding questions</i>
<i>ACTORS</i>	Who are the main actors? Which actors play a supporting role, and which ones are irrelevant? What do these actors want? For example, status, money, power, prestige, social approval? How do these actors think? For example, are they interest-maximizers? Or are they rule followers, thinking in terms of professional norms or some other type of shared standard? Where do these norms/worldviews come from? How are they produced and reproduced? What kind of cognitive constraints do these ways of thinking impose on the actors? What resources do these actors have, and which ones do they lack?
<i>SPACE OF ACTION</i>	Where does the relevant action unfold? For example, is the space of action found in formal organizational or institutional structures, or in more informal venues? Is it limited to the nation state or does the space include international venues? Is the space temporally bounded or does it extend into the distant past? What features of the space of action matter for the process in question? For example, are there any formal institutional constraints that block or discourage certain courses of action? Are there any important power imbalances? How do actors become effective players in this space of action? Are any resources or forms of capital needed to act in certain ways or to visualize productive modes of behavior?
<i>INTERACTIONS</i>	How do the actors come to encounter each other in the space of action? Do these interactions happen naturally or do they have to be actively pursued by some of the actors? What does each actor do to the others in the space of action? For example, do actors issue threats, mount pressure, transfer money, communicate new ideas? What are the mechanisms via which these actions produce reactions? What are the reactive properties of the things transmitted through action, and what are the conditions under which they trigger the desired reactions? Which types of interactions fail to produce reactions that are consequential for engineering the outcome of interest? How is the chain of actions and reactions connected to the outcome of interest?

Theoretically, the book adopts a sociological institutionalist perspective on judicial behavior that is centered on the concept of ‘legal preferences’. I argue that legal preferences constitute a relatively stable and deeply engrained bureaucratic lens through which judges perceive the cases they are asked to adjudicate and determines which solutions they deem appropriate or possible. In the case of human rights prosecutions in Latin America, victims and their lawyers had to undermine the hegemony of formalistic legal preferences, which predisposed judicial actors to dismiss lawsuits against the perpetrators and left judges and prosecutors ill-equipped to investigate these complex crimes. Their formalistic instincts had to be replaced with a new bundle of legal preferences anchored in the values of international human rights law. This is because international human rights law makes possible innovative readings of

constitutions and criminal codes and opens up the necessary legal space for successful prosecutions. The theory specifies the conditions under which external shocks produced by organized litigation campaigns lead to such ‘shifts in legal visions’, triggering structural transformations in the jurisprudential criteria applied by the courts. The book thus explores how the reproduction and transformation of ideas about the law and received standards of legal praxis shape the role judges are willing and able to play in politically salient debates, conditioning the exercise of judicial power.

Political outcomes are always the product of people doing things. Even if the story is deeply structural, there must be a discussion of the *actors* involved, and why they are driven to act in certain ways. In my case, the actors who determine whether or not there is justice for the victims of state repression are the judges and prosecutors. For example,

they are the ones who must decide whether to uphold amnesty laws or declare them unconstitutional and proceed with the investigations. Judges and prosecutors are also in charge of devising creative strategies to collect evidence. This is an extremely difficult task because the crimes in question were usually committed decades before and covered up using the state apparatus (e.g. forced disappearances). And once the evidence is collected, judges and prosecutors have to evaluate whether or not it is enough to convict the defendants. All of this may sound obvious, but many accounts of transitional justice in Latin America ignore judicial actors, or relegate them to a secondary or passive role.

In addition to specifying what the protagonists need to do in order to produce the outcome, it is also important to explain our assumptions about how these actors think and what they want. In the case of judges and prosecutors, I contend that they inhabit an environment that is highly regulated by norms of professional conduct, which are the result of socialization processes inherent to the legal field.<sup>4</sup> Socialization promotes dispositions, ways of reasoning and patterns of behavior that are scripted by the language of the law, and results in the institutionalization of what I call 'legal preferences'. Legal preferences encompass views about the reach and pliability of formal judicial prerogatives, and what constitute legitimate sources of law (domestic, international, doctrinal), acceptable forms of legal argumentation or reasonable standards of proof. These preferences thus engender a logic of behavioral appropriateness that leads judges to believe that certain legal solutions ought to be favored and defended, regardless, for example, of what politicians may prefer or instruct them to do. In addition, because legal preferences promote adherence to routine decision-making templates, they structure the judicial imagination and constrain the types of decisions judges are likely to reach. Unlike other scholars of judicial politics, I therefore do not think of judicial actors as

interest-maximizers who behave strategically to advance policy or career goals, but rather as followers of professional behavioral standards. This characterization is relevant because it means that judges need to be persuaded that there are plausible legal arguments in favor of a specific case outcome before they can become champions of particular causes or interests. Political pressures in one direction or another do not necessarily determine how they rule. Moreover, radical changes in decision-making routines and the adoption of novel legal criteria that enable jurisprudential innovations are unlikely to happen in the absence of re-socialization processes that uproot existing legal preferences and diffuse alternative ones.

Latin American judiciaries were historically dominated by formalistic legal preferences, which, in cases of state repression, biased judicial actors in favor of upholding amnesties and statutes of limitations, and left them ill-equipped to gather the necessary evidence and evaluate its probative value in ways that benefited the victims of repression. Importantly, these legal preferences made most judges ignorant or deeply skeptical of international human rights law, which is the body of law that makes prosecutions possible. This characteristic of the main actors in the story and their milieu has critical implications for identifying other key actors, as well as the forms of capital they need to amass to induce judges to produce waves of human rights trials.

Unlike other accounts, mine does not emphasize social pressure in the form of protests or activists' presence in the media, the relative power of the military post-transition, the levels of support for transitional justice among politicians or the international processes whereby human rights legal norms are created and diffused. I put judges at the center of the story; they are the ones who hold the keys to the outcome. Importantly, because I assume that judges behave as rule followers and not interest-maximizers, the argument is that politicians or passionate protests cannot

easily whip them into handing down prison sentences or acquittals out of fear of losing their jobs or seeing a decline in popularity. Moreover, since the standards which judges follow are deeply engrained, international human rights norms do not easily trickle down into domestic jurisprudence. Learning doesn't happen by osmosis; actors outside the judiciary must actively promote it.

If the above assumptions about how judges think are correct, the success of criminal prosecutions depends on the extent to which formalistic legal preferences are replaced by new ones, specifically, legal preferences based on the values of international human rights law. This allows me to identify the other key actor in the story, namely victims of state repression and the organizations they create to fight for justice. Victim organizations prevail when they are able to bring cases to court, and, crucially, engage in re-socialization efforts, diffusing new legal knowledge and innovative standards of adjudication that enable the success of those lawsuits. And to do so, they must amass very specific kinds of resources that allow them to infiltrate the judiciary, speak the language of judicial actors and craft skillful and ambitious interventions to change how judges and prosecutors see their obligations in cases of state repression. Staging mass protests and obtaining other forms of political capital doesn't do the trick; instead, victims need to recruit professionalized legal teams and allies in the academic world capable of transforming formalistic legal preferences.

The model therefore highlights the relevance of formal and informal processes at the level of the nation state associated with litigation efforts by victims and their lawyers. The *space of action* is not the street, presidential palaces, military barracks or the international forums where human rights norms develop, but national courts and their surroundings. It is here that the battle for justice is ultimately won. Furthermore, my characterization of the institutions and logic of action of the legal field leads me to identify the forms of capital

that are needed to act effectively in pursuit of certain objectives. Victims must be able to intervene in very specific ways in order to achieve their aims. In this sense, I emphasize the importance of contacts between victims and judicial actors that enable the spread of new ways of thinking about the law.

After specifying a space for action embedded in the judiciary and inhabited primarily by judges, prosecutors-and litigants, the next step is to describe the relevant *interactions* conducive to the outcome of interest. I argue that when victim organizations are equipped with professionalized legal teams, they are able to understand the obstacle posed by the incompatibility of arguments based on international human rights law and dominant formalistic legal preferences. By taking seriously how judges and prosecutors think about these cases as professionals of the law, litigants are able to design effective tactics of legal contention that produce necessary changes in legal preferences. The main tactic consists of orchestrating pedagogical interventions.<sup>5</sup> These include the organization of seminars on human rights law for judicial personnel and the circulation of academic material in court. Litigants thus teach members of the judicial community complex and often ignored juridical doctrines, and in the process disrupt bureaucratic inertia, manufacturing a legal framework that favors indictments and convictions. The ideas circulated in these venues have a technical component that creates a roadmap for new patterns of judicial behavior, and also a normative component that legitimizes unusual courses of action.

Pedagogical interventions activate a re-socialization mechanism that transforms legal preferences and jurisprudence. I rely on existing theories of education and socialization to specify the characteristics that make these interventions effective. I suggest that if pedagogical interventions are designed in certain ways they will trigger psychological reactions that facilitate persuasion. I also describe contextual factors that make success more likely. For example, I contend that the timing



of pedagogical interventions vis-à-vis similar efforts by those who seek to promote impunity, or vis-à-vis court-packing initiatives by anti-transitional justice coalitions, matters a great deal. If victim organizations strike first, they usually encounter a wider pool of persuadable judges and can engineer more robust ideational transformations. Robust ideational transformations guarantee that trials are more far-reaching and ambitious.

The model specified to answer my research question suggests that waves of human rights trials are not so much the result of the political will or capacity of democratic politicians to hold dictators accountable, but occur when judges are equipped with the legal preferences needed to deal with these unusual and complicated cases. Judicial actors must learn about international human rights law to be able to deliver justice, and victim organizations are the ones who supply this information. The causal chain thus begins with victim organizations forming professionalized legal teams, which in turn enable them to understand the problem posed by formalistic legal preferences and the need to change them. It continues with these organizations deploying specific tactics of contention, that is, pedagogical interventions that activate a re-socialization mechanism. The third step is the ensuing transformation in legal preferences, which equips and commits judges to adopt innovative patterns of behavior compatible with transitional justice. Importantly, the new sense of professional mission that is instilled by the new legal preferences encourages judges to resist military or political pressures contrary to prosecutions, and continue working for the victims. All of these modeling choices are based on prior knowledge of the cases (e.g. Latin American legal cultures and their incompatibility with the international legal case for prosecutions), as well as on well-established theories and empirical findings (e.g. assumptions about the psychological underpinnings of socialization processes).

In the following sections, I discuss how spending time on theory development along

the lines of the field approach can be of great use further down the line.

## IMPLICATIONS FOR DATA COLLECTION

Thinking in terms of fields, and the processes that unfold within them, disciplines key components of the research project. In particular, these models serve as roadmaps during fieldwork and structure the data collection process by making it easier to derive observable implications.

Ideally, our fieldwork should follow a sequence that mirrors the sequence of steps specified in the causal chain. There are two reasons for this. First, our models are only validated when all steps in the causal chain are adequately documented (George and Bennett, 2005: 207; Beach and Pedersen, 2013: 39; Waldner, 2014: 132; González-Ocantos and LaPorte, forthcoming). Insofar as our goal is to provide a logical or internally sufficient account of the outcome of interest, the inability to document some of the steps in the model will force us to go back to the drawing board to rethink the causal chain. And because models assume that each step describes a reaction or state of affairs that is brought about by a preceding step, it makes sense to try to document them in the order we specified in the model.

For example, when I conducted my research on transitional justice, I first documented how victim organizations assessed the obstacles they faced at the start of the process, followed by an analysis of the tactics they deployed as a result of their diagnosis of the nature of the challenge ahead. Once I had gathered enough evidence suggesting that they indeed understood the problem posed by formalistic legal preferences, and reconstructed the informal spaces subsequently created for pedagogical interventions, I proceeded to interview the judges to probe legal preference change. Had I not been able to

establish the first two steps in the sequence, I would have had to rethink the model before contacting judicial personnel.

Second, this sequential approach to data gathering makes fieldwork more productive. Each step in the causal chain essentially involves actors doing things to other actors. So if we first obtain information about the actors instigating the process and their behavior, we immediately have a better sense of what information to look for regarding the targets of those actions. In my case, knowing the details of pedagogical interventions allowed me to identify relevant judicial personnel, and improved my ability to question them about those experiences and the effects, if any, that the interventions had on legal preferences.

Another way in which the field approach contributes to more effective data collection is by making it easier to derive observable implications. Before arriving in the field site, researchers must produce a comprehensive list of the kind of empirical evidence that would be compatible with the model. These are the fingerprints that the process leaves in the available record if our story is correct. The assumptions and choices we make at the model specification stage are an obvious point of departure: 'if this assumption is correct, I should observe X, Y and Z'. Moreover, listing observable implications for each step of the chain makes the task more manageable.

For instance, my assumptions about the relevant space of action led me to think in advance about the characteristics of the informal venues where pedagogical interventions take place. As a result, I had a sense of what to look for when rummaging through the archives of victim organizations: funding bids to support the interventions, reports to funding agencies describing how the funds had been used, lists of participants, syllabi and course packages distributed among participants, and so on. Another challenge was to document legal preference change and its consequences, in line with my assumptions about the non-strategic nature of judicial

behavior. I posited that judges and prosecutors must recognize shifts in their knowledge and standards of adjudication as a result of the pedagogical interventions, and assign greater importance to these changes than to their prior legal knowledge or other forms of influence. Specifically, changes in legal ideas should be reflected upon and admitted, and prior formalist preferences should be evaluated negatively. I deemed this a hard test for the argument because I expected judicial actors to be quite reticent to admit prior ignorance of international law or the informal influence of the victims. The exact nature of the evidence was of course determined by access and country-level characteristics. Whereas in Argentina I relied mainly on in-depth interviews, in Peru I supplemented interviews with a survey of prosecutors that I was able to conduct with the help of the authorities. Similarly, the sheer volume of court cases in Argentina allowed me to build a dataset of rulings, and use simple statistics to test the impact of instances of overt political pressure on the judges on their decision-making patterns. This produced evidence incompatible with strategic models of judicial behavior.

## **IMPLICATIONS FOR IDENTIFYING ALTERNATIVE EXPLANATIONS**

A robust modeling exercise at the start of the research project also facilitates the identification of relevant rival explanations. Process tracing methodologists advise us to 'cast the net widely for alternative explanations' (Bennett and Checkel, 2015: 23). This is crucial for the success of the qualitative enterprise because the quality of the data is not determined by the sheer amount of evidence, but by the extent to which different observations are compatible with our story and incompatible with its rivals. But casting a wide net can be quite challenging, and is usually a source of anxiety. An obvious resource

is the existing literature on the topic. After all, when we design the project we always do it with reference to this literature in an effort to carve out a space for our work, showcasing its originality and import vis-à-vis existing accounts. While the literature will probably not reveal an exhaustive list of rival explanations, it will certainly give us a sense of the most influential ones.

Modeling in terms of fields is equally helpful. As Lake and Powell (1999: 14) put it, '[a]ll theories attempt to simplify a complex reality, and, as a result, they all reflect judgements made by theorists as to what to put into their analysis and what to leave out'. When we specify the contours of the field and its inhabitants we make choices about who and what matters, and who and what does not. We contend that some venues may be hidden from view but actually exist and are crucial for understanding the outcome, that some logics of action take primacy over others, that some forms of capital are more productive than others or that some tactics work better than others. It follows that what we consciously leave out or downgrade in importance is a good candidate for a rival explanation. This does not mean that all of the assumptions that produce exclusions of this sort can or need to be tested, but our work will be more convincing if we minimize the number of assumptions that goes untested.

In my case, the model produced two important rivals. First, I assumed that judges hold the keys to unleash waves of human rights trials. Moreover, I argued that they behave as rule followers rather than interest-maximizers that think strategically to determine how to behave. This led me to position myself against influential models of transitional justice, which emphasize, for example, the centrality of the balance of power between military and civilian authorities post-transition, or the preferences of elected politicians vis-à-vis human rights prosecutions, in determining the outcome of the trials. In these accounts, judges are epiphenomenal actors, and respond to signals from politicians and military officers, handing

down convictions or stalling the investigations accordingly. By contrast, my depiction of the space of action and the cast of relevant actors suggested something very different, and forced me to document the crucial role of factors and processes downplayed by other work.

Second, I circumscribed the space of action to the nation state – specifically, to judicial institutions and the informal venues litigants create for pedagogical interventions. Whereas the human rights doctrines that litigants need to teach are the product of international processes of norm development, my claim was that the relevant action in cases of transitional justice takes place at the domestic level. International norm development is insufficient because it does not explain why or how national judges come to learn and accept standards of adjudication so foreign to their legal culture. Accounts that emphasize the direct influence on judicial behavior of events that take place abroad (e.g. international litigation efforts, international campaigns against impunity, etc.), affirming global human rights standards and shaming local authorities who remain reluctant to assist the victims, were therefore competing explanations to which I had to attend.

## IMPLICATIONS FOR CASE SELECTION

Qualitative researchers often choose cases for non-methodological reasons, including our personal obsessions and linguistic skills. We are also drawn to cases because of their historical significance, or because they have been sparsely researched. For instance, I am originally from Argentina and could not conceive of a research design that did not include this case. Transitional justice has been a defining issue of the country's politics during my lifetime, and always preoccupied me. In my view, there is nothing wrong with applying such criteria. As Guillermo O'Donnell put it, 'my questions still come from broad political and moral concerns [...] I've tried to

deal with the kinds of real-world problems that deeply bother me when I'm shaving' (Munck and Snyder, 2007: 297).

Having said this, justifying the case selection with methodological criteria is of critical importance. For example, a good case selection strategy may allow for a rigorous probe of the scope of the argument by examining it under a variety of theoretically relevant conditions. This in turn enables a defense of the portability of the model beyond the main empirical referents. Similarly, if one is lucky enough to be able to craft tightly controlled comparisons, it may be possible to rule out alternative explanations by design. The literature already offers excellent advice on case selection, so my goal here is admittedly quite modest.<sup>6</sup> I simply want to point out ways in which thinking of our models as stylized descriptions of a field may also help us select positive and negative cases.<sup>7</sup>

With regards to positive cases, that is, those in which the outcome occurs, specifying the main contours of the field and its internal logic helps identify the contextual factors that the design should try to vary within and across cases. This variation is often needed to show that the argument applies in a variety of contexts – that is, it is not idiosyncratic – and, most importantly, to show that those contextual features do not necessarily explain the outcome. The latter is a good way to undermine alternative explanations. To accomplish this, we should seek variation in contextual factors that are associated with an alternative explanation and that are therefore not expected to fundamentally affect the dynamics of the field or the production of the outcome. This could be because changes in context introduce irrelevant actors, or trigger interactions outside the relevant space of action, that can be shown not to affect the outcome. Changes in context may also fail to condition the supply of productive capital to the relevant actors or the successful deployment of the tactics that, according to our theory, activate key mechanisms of change. In other words, by specifying the field along the

lines suggested here, it is possible to anticipate whether different background conditions or environmental shocks will be inert, or collide with the field and produce consequential reactions.

For example, the two positive cases in my study, Argentina and Peru, offer within and cross-case variation in the preferences of elected officials vis-à-vis transitional justice and levels of hostility of the political environment toward victim organizations. By analyzing how the process unfolds in contexts with more or less hostility, I am able to show that the implementation and success of pedagogical interventions does not depend on the tolerance for prosecutions among the political class or the strength of the military, and that the judiciary does not necessarily rule with an eye on these strategic incentives. I can also show that in favorable environments where the political class strongly supports the trials, the absence of pedagogical interventions, or the presence of interventions that are not designed carefully, leads to a weaker impetus for justice among judicial actors. This is because the logic that governs the judicial field requires legal preference change in order to observe dramatic shifts in the jurisprudential criteria applied by the courts. And the ability to induce such changes, that is, the type of capital that 'confers power' or is reactive in this field (Bourdieu and Wacquant, 1992: 101), is not dependent on political power but on skillful pedagogy.

A robust modeling exercise at the start of the research project can also help select negative cases, that is, cases in which the outcome does not occur. While the main source of inferential leverage in outcome-oriented qualitative research comes from the reconstruction of uninterrupted causal chains in positive cases, a comparison with a negative case provides empirical evidence in support of counterfactual reasoning. But negative cases present unique challenges when it comes to case selection. There are usually a huge number of cases, both historical and contemporary, where the outcome

of interest is absent. So which ones should we pick? Mahoney and Goertz (2004) argue convincingly that one should apply the ‘possibility principle’ to select relevant negative cases, that is, cases where the outcome could have occurred but didn’t. They propose two selection rules. According to the ‘rule of inclusion’, the value of at least one independent variable must be associated with a positive value in the dependent variable. According to the ‘rule of exclusion’, the value of none of the independent variables must predict a negative outcome with certainty. Cases that do not meet these criteria should be avoided.

The field approach is useful to think about how to apply these rules. If our point of departure is a clearly specified causal chain, with actors, spaces of action and interactions, it is possible to identify negative cases where at least some of the elements of the field and the causal chain are present. In particular, we should strive to select cases where the main actors are present, have some of the forms of capital needed to activate key mechanisms of change and interact with each other in various ways. This means that the outcome could have happened had the right kinds of interactions taken place, or had the actors amassed additional forms of capital. A negative case with these characteristics allows us to trace the point in the sequence at which the flow of causal energy is interrupted or stops, bolstering claims regarding the importance of all component parts of the model in producing the outcome. Moreover, the interactions between the key actors leave footprints, which enable the application of the tools of process tracing. Without these fingerprints, it is hard to apply the method.

The main negative case in my study is Mexico. I chose it because I knew that, as in the positive cases, the judiciary harbored formalistic legal preferences incompatible with human rights prosecutions, a key feature of the field of action. Moreover, Mexico has a vocal and well-organized human rights movement that fought hard to punish those responsible for the crimes

of the ‘dirty war’. In other words, the cast of central actors was complete. Finally, in the aftermath of the 2000 democratic transition, a special prosecutor was appointed to investigate these cases and subsequently filed numerous lawsuits against former politicians and military officers. This created the space for sustained interactions between victim organizations and judicial personnel, which left behind traces. During fieldwork, I found out that victim organizations focused on amassing unproductive forms of capital, such as links with political parties or mobilization capacity, instead of investing in professionalized legal teams. They therefore failed to understand the logic that makes the legal field ‘tick’, and were incapable of designing and deploying the tactics of contention to activate the re-socialization mechanism. In their encounters with judicial actors, activists did not diffuse new legal knowledge or jurisprudential standards to transform legal preferences, and this interrupted the flow of causal energy.

## CONCLUSION

In this chapter I reflected upon my experience as a qualitative analyst and provided some advice on how to design a project. The discussion is mainly relevant for research that explains political outcomes by tracing processes within cases, and that compares a small number of cases. The key message is that attention to theoretical priors from the very beginning is crucial, and has implications for data collection, the identification of alternative explanations and case selection. Importantly, the emphasis on modeling choices suggests that qualitative research is not merely about ‘soaking and poking’, but is disciplined by variables, hypotheses and concomitant observable implications. Modeling may at first seem like a daunting task, but thinking in terms of fields and their three components – actors, spaces of action and interactions – can make the job easier.

Qualitative research is not for the faint of heart. It requires the collection of copious amounts of data, often of radically different kinds within and across cases. This in turn forces researchers to become attentive to infinite sources of bias. We are also highly vulnerable to disconfirmation in the event that we end up with gaps in the evidentiary record. This is because, unlike an empty cell in a quantitative dataset, missing information about key steps or mechanisms fundamentally undermines the central promise of process tracing methods, which is to reconstruct the proposed causal pathway in its entirety (González-Ocantos and LaPorte, forthcoming). But qualitative research is also exciting and tremendously rewarding, especially during those eureka moments when we stumble upon precious smoking guns. Having well-specified models from the start structures the research process and gives it a clear direction of travel, boosting the chances of finding such evidence and generally making fieldwork more productive. It also adds rigor to the project because it provides readers with a yardstick to evaluate the logical consistency and persuasiveness of the narrative.

By way of conclusion, I would like to point out the wide range of applications of the field approach. First, the emphasis on strong theory does not mean that inductive research is off the table. While the model is designed to offer a roadmap for the research process, it still allows for detours as we enter into a dialogue with the cases and the evidence. Amending parts of the model as we go along is not 'unscientific' insofar as the goal of outcome-oriented qualitative research is precisely to provide a comprehensive explanation of political outcomes. In my case, for example, my model only specified re-socialization as a key mechanism of change. When I started collecting data about judges, however, I realized that some of them were deeply recalcitrant and not amenable to persuasion. Furthermore, I saw how victim organizations reverted to other tactics (e.g. impeachments) to get rid of them and make

way for new voices in the judicial branch. My amended model thus offered a more adequate characterization of the cast of actors in the field and the mechanisms of change, and as a result, it was better equipped to explain outcomes within and across cases.

Second, the centrality of actors and interactions may leave the impression that the field approach is not useful for probing structural accounts. This is not the case. By employing the term 'actors' I do not necessarily endorse a strong version of methodological individualism. With the right theoretical justification, the entities that act in our models could be specified at various levels of aggregation (e.g. president, cabinet, ruling party or even dominant class). Moreover, nothing stops us from assuming that individual actors are more or less determined and constrained by deeper social forces at play in any particular field. For instance, behavior by members of a social class need not be conscious or planned, but the result of atomized choices resulting from shared circumstances. The point is simply that structural factors impact reality through the beliefs, perceptions and actions of individuals and groups, so these remain a crucial component of any model. Without some reference to actors, their behavior and interactions, the story is likely to remain logically incomplete. As George and Bennett (2005: 142) explain,

[m]acrosocial mechanisms can be posited and tested at the macrolevel [...] All that commitment to microlevel consistency entails is that individuals must have been capable of behaving, and motivated to behave as the macrolevel theory states, and that they did in fact behave the way they did because of the explicit or implicit microlevel assumptions embedded in the macrolevel theory.

Or in Kitschelt's (2003: 59) words: 'weak methodological individualism is not inimical to the consideration of structural and collective phenomena. It only requires that we treat individuals' actions as critical ingredients in any account of structural transformation.'

A related concern is that an emphasis on actors and interactions favors explanations

that are too proximate to the outcome to be considered 'causal'.<sup>8</sup> This is also not the case. Thinking in terms of fields and their constituent parts does not mean that one should only look at decisions immediately prior to the outcome. In fact, causal chains usually extend back in time because we must account for the antecedent process, including the creation and transmission of necessary forms of capital to the final decision-makers, mechanisms of preference change or the entrance of the relevant actors into the space of action. How far back to go is a theoretical decision that is not necessarily constrained by the need to specify actors and their interactions.<sup>9</sup> In this sense, there are a variety of ways to locate theoretically the starting point of the causal chain. For example, in the case of outcomes characterized by some form of path dependence, we may look for relevant critical junctures (Capoccia, 2015). We may also choose to begin where rival accounts take off in order to challenge the presence of a causal pathway along the lines specified in those models. Or we may decide to start when the relevant cast of actors is complete or at a point when the main contours of the space of action are already constituted.

Finally, economicist terms such as 'capital' could also suggest that the 'field' approach is only suitable for rational choice scholars or game theorists. This is also a misperception. The primary logic of action of any particular field or actor could be consequentialist, strategic or utility maximizing, but it can also be rule oriented, role satisficing and deeply constituted by social norms. Similarly, the deployment of (or quest for) certain forms of capital could result from conscious calculations or unconscious dispositions. However we define these preferences and logics of action, the 'field' approach allows us to identify the types of behavior that are capable of making the system tick in the direction of the outcome of interest, removing blockages and shifting the balance of power against champions of the status quo. These are the moving parts of the model (which some researchers,

but not all, may see as a 'game' of strategic interactions) that explain within-case variation or stasis in the dependent variable.

## Notes

- 1 This is not to say that qualitative researchers never start with a comparative research question that doesn't feature proper names, and proceed to choose cases deductively. This is common, for example, in the case of cross-regional comparative studies (e.g. Lieberman, 2009).
- 2 For a similar approach, see Lake and Powell's (1999) strategic-choice template for theory building. I thank Branislav Slantchev for pointing this out. The 'field' approach, however, does not necessarily rely on rational-choice assumptions. For a more detailed discussion, see the concluding section.
- 3 Waldner (2014) refers to these as 'causal graphs'.
- 4 For Bourdieu's own application of the field approach to the legal field, see Bourdieu (1986).
- 5 The model also includes a second tactic, namely personnel replacement strategies. For ease of exposition, I will not refer to this part of the argument.
- 6 Important contributions include Lijphart (1971), Mahoney and Goertz (2004), George and Bennett (2005), Lieberman (2005), Gerring (2007), Gerring and Seawright (2008), Slater and Ziblatt (2013), Beach and Pedersen (2016), Seawright (2016) and Goertz (2017).
- 7 Dependent variables need not be binary (positive/negative, present/absent) or ordinal (various degrees of presence/absence). For example, many qualitative researchers work with categorical dependent variables.
- 8 For example, Coppedge (2012, chapter 3) makes the case that distant factors are analytically superior. For a defense of proximate factors, see Kiser and Hechter (1991), Capoccia and Ziblatt (2010) and Mainwaring and Pérez-Liñán (2013: 30–3).
- 9 But there might be a trade-off between causal distance or 'depth' and our ability to trace mechanisms that inhere in the minds of actors. See Kitschelt (2003).

## REFERENCES

- Beach, Derek, and Rasmus Brun Pedersen. 2013. *Process-Tracing Methods: Foundations and Guidelines*. Ann Arbor: University of Michigan Press.

- Beach, Derek, and Rasmus Brun Pedersen. 2016. *Causal Case Study Methods: Foundations and Guidelines for Comparing, Matching and Tracing*. Ann Arbor: University of Michigan Press.
- Bennett, Andrew. 2015. 'Disciplining Our Conjectures: Systematizing Process Tracing with Bayesian Analysis.' In *Process Tracing: From Metaphor to Analytic Tool*, Andrew Bennett and Jeffrey Checkel eds. Cambridge: Cambridge University Press, pp. 276–298.
- Bennett, Andrew, and Jeffrey Checkel, eds. 2015. *Process Tracing: From Metaphor to Analytic Tool*. Cambridge: Cambridge University Press.
- Bourdieu, Pierre. 1986. 'The Force of Law: Toward a Sociology of the Juridical Field.' *Hastings Law Journal*, 38(5): 814–853.
- Bourdieu, Pierre, and Loïc Wacquant. 1992. *An Invitation to Reflexive Sociology*. Oxford: Polity Press.
- Brady, Henry, David Collier and Jason Seawright. 2004. 'Sources of Leverage in Causal Inference: Toward an Alternative View of Methodology.' In *Rethinking Social Inquiry: Diverse Tools, Shared Standards*, Henry Brady and David Collier eds. Lanham: Rowman & Littlefield, pp. 229–266.
- Bunge, Mario. 1997. 'Mechanism and Explanation.' *Philosophy of the Social Sciences*, 27(4): 410–465.
- Bunge, Mario. 2004. 'How Does It Work? The Search for Explanatory Mechanisms.' *Philosophy of the Social Sciences*, 34(2): 182–210.
- Capoccia, Giovanni. 2015. 'Critical Junctures and Institutional Change.' In *Advances in Comparative-Historical Analysis*, James Mahoney and Kathleen Thelen eds. Cambridge: Cambridge University Press, pp. 147–179.
- Capoccia, Giovanni, and Daniel Ziblatt. 2010. 'The Historical Turn in Democratization Studies: A New Research Agenda for Europe and Beyond.' *Comparative Political Studies*, 43(8/9): 931–968.
- Collier, David. 2011. 'Understanding Process Tracing.' *PS: Political Science and Politics*, 44(4): 823–830.
- Coppedge, Michael. 2012. *Democratization and Research Methods*. Cambridge: Cambridge University Press.
- Gerring, John. 2001. *Social Science Methodology: A Critical Framework*. Cambridge: Cambridge University Press.
- Gerring, John. 2007. *Case Study Research: Principles and Practices*. Cambridge: Cambridge University Press.
- Gerring, John, and Jason Seawright. 2008. 'Case Selection Techniques in Case Study Research: A Menu of Qualitative and Quantitative Options.' *Political Research Quarterly*, 61(2): 294–308.
- George, Alexander L., and Andrew Bennett. 2005. *Case Studies and Theory Development in the Social Sciences*. Cambridge: MIT Press.
- Goertz, Gary. 2017. *Multimethod Research, Causal Mechanisms and Case Studies: An Integrated Approach*. Princeton: Princeton University Press.
- Goertz, Gary, and James Mahoney. 2012. *A Tale of Two Cultures: Qualitative and Quantitative Research in the Social Sciences*. Princeton: Princeton University Press.
- González-Ocantos, Ezequiel A. 2016. *Shifting Legal Visions: Judicial Change and Human Rights Trials in Latin America*. Cambridge: Cambridge University Press.
- González-Ocantos, Ezequiel, and Jody LaPorte. Forthcoming. 'Process Tracing and the Problem of Missing Data.' *Sociological Methods and Research*. doi:10.1177/0049124119826153.
- Hedström, Peter. 2008. 'Studying Mechanisms to Strengthen Causal Inference in Quantitative Research.' In *Oxford Handbook of Political Methodology*, Janet M.Box-Steffensmeier, Henry E. Brady and David Collier eds. Oxford: Oxford University Press, pp. 319–338.
- Kiser, Edgar, and Michael Hechter. 1991. 'The Role of General Theory in Comparative-Historical Sociology.' *American Journal of Sociology*, 97(1): 1–30.
- Kitschelt, Herbert. 2003. 'Accounting for Post-communist Regime Diversity: What Counts as a Good Cause?' In *Capitalism and Democracy in Central and Eastern Europe*, Grzegorz Ekiert and Stephen Hanson eds. Cambridge: Cambridge University Press, pp. 49–88.
- Kohli, Atul, Peter Evans, Peter J. Katzenstein, Adam Przeworski, Susanne Hoebler Rudolph, James C. Scott and Theda Skocpol. 1995. 'The Role of Theory in Comparative Politics: A Symposium.' *World Politics*, 48(1): 1–49.



- Lake, David, and Robert Powell. 1999. 'International Relations: A Strategic-Choice Approach.' In *Strategic Choice and International Relations*, David A. Lake and Robert Powell eds. Princeton: Princeton University Press, pp. 3–38.
- Lieberman, Evan. 2005. 'Nested Analysis as a Mixed-Method Strategy for Comparative Research.' *American Political Science Review*, 99(3): 435–452.
- Lieberman, Evan. 2009. *Boundaries of Contagion: How Ethnic Politics Have Shaped Government Responses to AIDS*. Princeton: Princeton University Press.
- Lijphart, Arend. 1971. 'Comparative Politics and Comparative Method.' *American Political Science Review*, 65(3): 682–693.
- Mahoney, James. 2012. 'The Logic of Process Tracing Tests in the Social Sciences.' *Sociological Methods & Research*, 41(4): 570–597.
- Mahoney, James, and Gary Goertz. 2004. 'The Possibility Principle: Choosing Negative Cases in Qualitative Research.' *American Political Science Review*, 98(4): 653–670.
- Mainwaring, Scott, and Aníbal Pérez-Liñán. 2013. *Democracies and Dictatorships in Latin America: Emergence, Survival, and Fall*. Cambridge: Cambridge University Press.
- Munck, Gerardo L., and Richard Snyder. 2007. *Passion, Craft, and Method in Comparative Politics*. Baltimore: Johns Hopkins University Press.
- Seawright, Jason. 2016. *Multi-Method Social Science*. Cambridge: Cambridge University Press.
- Slater, Dan, and Daniel Ziblatt. 2013. 'The Enduring Indispensability of the Controlled Comparison.' *Comparative Political Studies*, 46(10): 1301–1327.
- Waldner, David. 2007. 'Transforming Inferences into Explanations: Lessons from the Study of Mass Extinctions.' In *Theory and Evidence in Comparative Politics and International Relations*, Richard Ned Lebow and Mark Irving Lichbach eds. London: Palgrave Macmillan, pp. 145–176.
- Waldner, David. 2012. 'Process Tracing and Causal Mechanisms.' In *Oxford Handbook of Philosophy of Social Science*, Harold Kincaid ed. Oxford: Oxford University Press, pp. 65–84.
- Waldner, David. 2014. 'What Makes Process Tracing Good? Causal Mechanisms, Causal Inference, and the Completeness Standard in Comparative Politics.' In *Process Tracing: From Metaphor to Analytic Tool*, Andrew Bennett and Jeffrey Checkel eds. Cambridge: Cambridge University Press, pp. 126–152.



# Theory Building for Causal Inference: EITM Research Projects

Thomas Bräuninger and Tilko Swalve

The golden rule that any empirics without theory is, at best, descriptive is likely to be taught to any undergraduate student in her very first year. And there is nothing wrong with this. At the most basic level, the rule reminds us that whatever the co-occurrence of two events, or however strong the correlation of two variables might be, we should not stumble into a fallacy, for example, believing that babies are delivered by storks when observing an association between the number of storks and the number of human births. At a more specific level, the rule includes a warning, namely that any sort of evidence for the specific causal mechanism which a theoretical argument postulates does not rule out alternative explanations. Data may provide strong empirical evidence for a causal mechanism; at other times, the argument fits into a broader, more general, or established theory. Neither case rules out the appropriateness of any other conceivable mechanism in some hypothetical theoretical account. Practically, the question then becomes how to find a

method for inference that lets the data speak to the specific mechanism we have in mind, not to any conceivable mechanism. This is the aim of the Empirical Implications of Theoretical Models (EITM) approach. The EITM approach is a device to ‘think about causal inference in service of causal reasoning’ (Aldrich et al., 2008).

We should stress that, while the EITM approach seeks to bridge the gap between theoretical and empirical work, it is not just a missing piece to fill a gap in a research design. We think of the EITM approach more as a way to think about related questions. First: how do you improve your methodological reasoning so that empirical work is most effective and informative about theories? Second: how do you improve your theoretical reasoning to provide a larger number of useful theoretical hypotheses that can be evaluated against the evidence of empirical models? In doing so, the EITM approach provides a ‘coherent approach to evaluating information and converting it into

useful and effective knowledge' (Aldrich et al., 2008: 828).

This chapter introduces the EITM approach; discusses its methodological foundations, virtues, and challenges; and exemplifies good practices using recent academic work on topical research issues. We discuss three different research designs for how to assess the usefulness of a theoretical model to explore a causal mechanism using data: (1) equilibrium point predictions; (2) comparative statics (the evaluation of relationship predictions); and (3) structural estimation approaches. We introduce each approach with a simple example and pieces of recent research that are exemplars of how to exploit the strength of each design. Finally, we give some advice on how to strengthen research designs using the EITM methodology.

## WHAT IS EITM?

The EITM approach promotes the idea that, in the best case, there is a tight connection between a theoretical argument represented by a formal model and an empirical data analysis used to learn when and how the model finds empirical support (or not). There are ontological and methodological presumptions underlying this idea. One is that the ultimate goal of social science inquiry is the explanation of empirical events in the sense of deducing the specific event or statement from a general, theoretical statement about causal relationships. The creation and evaluation of such theoretical statements, then, is an integral part of scientific inquiry. Another assumption is that theoretical statements about states, events and causal relationships can reasonably be captured by formal models.

In making such a claim, we also assume that empirical data can be useful to identify or evaluate the model (the two are not the same; nor is empirical analysis the only way to evaluate a model: see more on this below),

and that evaluation and model identification can tell us something about the usefulness of the theoretical model, and therefore give us a better understanding of the world. Note that we shy away from using the term 'model testing' here. As Clarke and Primo (2007: 749) remark, models are representational objects; they cannot be true or false. Also, as models serve more than just predictive purposes (they may also be foundational, structural, generative, explicative), they 'should be assessed for their usefulness for a particular purpose, and not solely for the accuracy of their deductive predictions'.<sup>1</sup>

Starting from these assumptions, the diagnosis and motivation for the EITM approach is the observation that much research squanders a great deal of the enormous potential of close links between theoretical and empirical work. This criticism concerns at least one of two points.

First, much theoretical work remains at the level of theory, forgoing the chance to learn more about the *theoretical* argument by exploring empirical data. When the EITM project began in 2001, with an NSF workshop to discuss avenues for improving technical-analytical proficiency in political science, a widely shared concern was the perceived uncoupling of theory building and empirical research as a result of the fast advancement of the discipline's research methods. Since the 1960s, theory building in political science has largely benefited from and partly moved to represent and quantify abstract concepts mathematically, using tools and concepts of social choice, game theory and microeconomics. This fostered precision in theory building, and hence improved research transparency, lent credibility to this type of work and resulted in knowledge accumulation. At the same time, the evolution of novel statistical and computational methods fostered the specialization of researchers employing applied statistics and empirical modeling. It was felt that a split had developed between the two camps, or, at least, that graduate training was too often one-sided and created

sophistication in either formal or empirical modeling, but not both (National Science Foundation, 2002: 1). The critical question then became how to link rigorous theoretical reasoning and appropriate identification strategies to learn about a causal mechanism.

Second, data analysis is often not tied to the theoretical model and, therefore, it is not informative about the causal relationships postulated by the model. Moreover, the focus of analysis often is on one specific causal effect. Much more could be done with a theoretical model, however, in order to derive empirical implications that can then be combined with data to support modeling assumptions, provide a better understanding of the causal mechanism or gain insight into an entire chain of postulated causal relationships. All these types of analyses in the EITM approach offer valuable payoffs.

## FROM THEORETICAL MODEL TO IMPLICATIONS

We use the terms ‘theory’ and ‘theoretical model’ in a quite pragmatic way. A theoretical model is a representation of a causal mechanism. A theory is a set of models that is linked by hypotheses to a set of related features of the real world (Giere, 2010: 85). In this definition and in accordance with common usage in the social sciences, the term ‘theory’ means a larger structure that collects models that may address different aspects of the real world, such as the ‘theory of coalition formation’ or ‘bargaining theory’. The really important term here is ‘representation’. A theoretical model does not mirror the world on a smaller scale; nor does it simply represent a part of the world. A model simplifies and, in doing so, it focuses more on some aspects and less on others.

To be sure, there are decidedly different ideas on how the term model should be defined and, consequently, what a model really captures. One view is that there is a

*data generating process* that is behind what we observe (Morton, 1999: 33). This mechanism can be detected, at least in principle. The goal of science is, then, to reveal the mechanism, or get as close as possible to it. From this point of view, it makes sense to say that a model is valid (or invalid). Thus, model testing is an instructive scientific strategy. If the model repeatedly fails to pass some benchmark, it is considered falsified by the data, and thus rejected (Popper, 1959). In this classical falsification framework, the test against data is the relevant benchmark, and any such test probes both the model’s postulated mechanism and all (often implicitly stated) auxiliary assumptions.

Another, no less extreme, view is that real world processes are so complex that we can hardly hope to come even close to the true data generating mechanism. From this perspective, models are gross simplifications of the world that, given enough data, would always be rejected (Keane and Wolpin, 2007: 1351). The very idea of model testing thus becomes meaningless. What we can say and do, however, is gauge whether one model fits the data better than a second one – a procedure one may dub model selection. There might even be alternative models, so that each performs best in some contexts, but not in others. The real danger here is to mistake the selection of a proposed and stated mechanism with atheoretical, ad hoc curve fitting that might fit the data well, but consistently performs worse out-of-sample and hardly reveals a deep theoretical structure. A model that relates election results to pre-election polls of the days and weeks before is likely to predict the data with stunning accuracy, but is not very enlightening.

Though interesting, the epistemological debate about what is best conceived as a model is less important for our discussion. The EITM approach is useful in a wide range of contexts and with different epistemic interests (though we lean toward the second perspective in what follows). In some cases, we are interested in the effect of one specific

feature of the world, say  $X$ , on some outcome  $Y$ . Suppose there are other features,  $Z$ , that arguably have an effect on  $Y$ , but we have good reasons to assume that some separability condition holds and the relation of  $X$  and  $Y$  is unrelated to  $Z$ . Then, if our (novel) theoretical argument is captured by a model that includes  $X$  and  $Y$ , possibly in some complex way, a comparative static tells us *how*  $X$  is related to  $Y$ . A ‘test’ of the model would focus on the expected covariation of  $X$  and  $Y$ . The focus here is on the correctness of the model prediction: can we see in the data that  $X$  is related to  $Y$ ? Whether, or to what extent, the model is successful in predicting observed outcome  $Y$  is not important here.

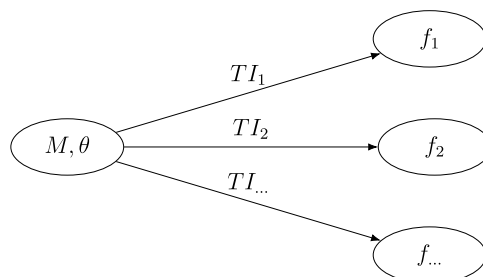
In other cases, we are interested in the substantial effect of some rule change, institutional reform or novel policy program. Does a 10% decrease in the required majority for a cloture cut the average duration of legislation in half? Does a 5% increase in child benefits have a larger effect on education equality than a 10% reduction in daycare fees? The focus with these types of research questions is on the size of the change in  $Y$  given some intervention  $X$ . A model useful to answer the latter question would take into consideration, for instance, how child benefits and daycare fees affect consumption and schooling decisions, while also taking into account expectations about the implications of these decisions in the future, and so on. While some mechanisms are better understood, other critical features of the model, like risk aversion and foresightedness of parents, have to be estimated as model parameters from the data. As the evaluation of a policy change is based on the comparison of effect sizes, a model must not only capture the key mechanism but also fit the data reasonably well. The approach here is closer to the idea of model selection (via estimation of parameters). What we are interested in is also completeness; that is, how much of the systematic variation in the data that the model aims to explain is captured by the model (Mullainathan and Spiess, 2017).<sup>2</sup>

The key value of a formal model is the clarity of its assumptions. Any theoretical argument that proposes a non-trivial link between a cause and an effect builds on a long list of assumptions about who acts and interacts when, how and with whom. It will also make auxiliary assumptions on the set of cases the model applies to, the nature of the (collective) actors, the cognitive capacities of agents, etc. Often these assumptions are seemingly innocent, as they appear to be common sense or trivial. A formal model makes these assumptions explicit.

Suppose we are studying popular protest in authoritarian regimes. While there are plausible arguments on socio-economic conditions favorable for uprisings, a formal model to capture the theoretical argument will not just explicate what motivates and constrains individuals. Instead, it will also make explicit whether the argument operates on the level of the public at large or on the level of individuals; whether, and, if so, how these differ and how individuals solve their collective action problem; or whether these problems are assumed away by auxiliary assumptions.

Auxiliary assumptions are a key part of a model, as no model can possibly capture every aspect of reality (Low and Meghir, 2017: 34). A model captures something essential of the data generating process, but not everything. A model of legislative oversight might consider variance in institutions, legislative majorities, and legislators’ preferences for one or another form of oversight while deliberately ignoring the role of a supreme court (McCubbins and Schwartz, 1984). This is not to suggest that supreme courts do not play any role in a legislator’s decision-making. It does assume, however, that decisions over police patrols or fire alarms are not substantially affected by long-term considerations. The question is: what can and cannot be left out to ensure the substantive conclusions of the model remain valid?<sup>3</sup>

A helpful term in answering the question is the term ‘separability’. Separability can



**Figure 7.1 Models, theoretical implications, and features of the real world**

Source: Adapted from Clarke & Primo (2012)

be understood as the invariance of an ordering on a subspace with respect to changes of variables outside the subspace (Blackorby et al., 1998). In fiscal policy literature, for instance, budgetary politics is often considered as either a top-down or a bottom-up process. In countries with a bottom-up process, so the story goes, individual portfolio ministers draft sectoral budgets, which are then aggregated and finally settled in bargaining. Where fiscal institutions allow for a strong finance minister, the budget process is more top-down: the finance minister decides on the total budget, and the allocation to portfolios is determined by intra-cabinet bargaining. In the first case, the total amount to be spent and the allocation are interrelated, and there is strategic interaction between agents. Whatever the ministerial draft looks like, it will affect both the total budget and the allocation. This is different in the second scenario, where bargaining over the allocation happens when the total budget is already fixed.

When modeling intra-ministerial bargaining, we therefore may consider the total budget as a fixed parameter, rather than reflecting on a full model that would model both the finance minister's decision and the allocation decision. The total budget can be thought of as *sufficient statistics*: something that summarizes decisions made outside the model (Low and Meghir, 2017: 34).<sup>4</sup> From the discussion, it should be clear that practically any political science model makes separability assumptions.

Whether or not a model has bite (we deliberately avoid terms such as right or wrong: see above) largely depends on the appropriateness of the separability assumptions made.

Figure 7.1 shows how we think about the relationship between model, theoretical implications, and features of the real world.

$M$  is the model with auxiliary assumptions: features that we assume about agents and the nature of the interaction – agents' preferences, the information they possess, individual information processing, rationality, decision-taking modes, choice sets, strategies available and equilibrium concept used. We think of these features as being fixed.

$\theta$  is a set of parameters: features that we think of as fixed in one specific case under study, but that will vary across different cases of interest. Parameters usually serve one of two purposes. Sometimes, we know or suspect that the cases under study vary greatly with respect to some marginal condition such as an individual trait, a sufficient statistic or an institutional feature that we know is critical to the mechanics of the model. Assuming one specific preference relation or utility function, a specific discount factor or voting threshold may amount to having a model that applies only to a single real-world case. If there is reason to believe that we can capture the variation in marginal conditions by some parameter, a parametrized model increases the scope of the model. Comparative statics is, then, a tool to study the effect in the

variation of these marginal conditions. At other times, it is unclear whether some marginal condition has any effect on the feature we study. In a parametrized setting, we might want to show that outcomes we are interested in do not vary in the parameter.

$f_1, f_2, \dots$  are outcomes in a very broad sense: features that are associated with the real world and that we argue or predict to occur or vary conditional on  $M$  and  $\theta$ . One common aspect of politics in which we are interested is the actions of agents: does a foreign state wage war? Does the incumbent invest in a costly campaign? Does the president delegate authority over trade policies to the legislature? This is all observable behavior, even though collecting data on actions taken by agents might be difficult. For instance, while it is pretty straightforward to verify whether a state has declared war on another, collecting data on whether the government has prepared for war might be intricate.

Observable actions must be distinguished from strategies, which are only partially known. Strategies are, using game-theoretical terminology, an agent's complete plan of what action to take at any contingency that might occur. By their very definition, strategies cannot simply be inferred from behavior. As they provide a rule for action in response to any contingency, they refer to cases that – in equilibrium – are counterfactuals. Consider the threat to fight. If the threat is credible, it will be successful and deter the opponent from making a certain move, for instance, an attack. To be credible, however, the strategy must include the plan to fight in the counterfactual contingency of an attack. Such a self-commitment is not readily observable.

Another aspect of politics in which we might be interested is outcomes, defined as consequences of the equilibrium behavior of agents. These can include policy output and outcomes, the evolution of an institution, some form of coordinated behavior and an international treaty or war. What we consider an outcome depends both on the puzzle that

we wish to understand and on the solution concept we use in our theoretical model.

$TI_1, TI_2, \dots$  are theoretical implications that link all assumptions to features of the real world. Generally speaking, we think of the usefulness of a theoretical model as weakly increasing with the number of its theoretical implications.

## SIMPLE THEORY, NON-OBVIOUS IMPLICATIONS

Let us consider a simple example of a theoretical model that we seek to combine with empirical data. Working through the example in some detail helps to clarify the steps we have to undertake, the assumptions that need to be made and the pitfalls we might end up with. Suppose we study the conditions when leaders wage a war. We have an argument that we seek to confront with empirical data. In our simple decision-theoretical model of going to war, a leader has two options. She can accept a previous offer to settle the dispute over a piece of territory yielding utility  $u_r$ .<sup>5</sup> Or she can wage war at costs  $c$ , which results in one of two events: a victory generating utility  $u_w$ , or a defeat with utility normalized to 0. The odds that the country will win the war is given by some exogenous, fixed and known probability. So, the model is indeed quite simple, as it ignores any strategic interaction with an opponent who himself might prepare for war, etc. Finally, we assume the leader values the two choice options with other criteria that are unrelated to the ones above, ones that we do not know and observe. We assume that disturbances are both unrelated to any other model parameter and known to the leader but unobserved by us (the researcher). For reasons that we will see in a second, we denote these by  $e_r$  and  $e_w$ .

To evaluate the argument empirically, provided that there is a sufficient number of cases, we would collect data on these cases and how they differ with respect to the above

conditions: the gains of winning the war and the gains of settling the dispute without waging war, but also the costs of a war, and the losses associated with losing, and so on. As the observed outcome is binary – the leader chooses war or acquiesces – we might run a standard probit or logit model with measures of gains and losses as covariates. Having estimated the statistical model, we can check whether the parameter estimates of the regressors have the expected sign, for instance, whether the costs of war have a negative effect on the occurrence of war, and so on.

However, given our description of model specification, the statistical model is not exactly capturing the mechanism that the theoretical model postulates. The theoretical model suggests that the effect of the gains of winning on waging war is not unconditional, but instead related to the probability of a victory. Obviously, if the odds of winning are nil, the size of the trophy becomes irrelevant. Thus, any estimation of the model with data that ignores the mechanism will thus not be instructive for the theoretical model. Though this may be obvious here, it may be unapparent with more complex mechanisms and interactions.

In fact, even in this trivial example, walking from the theoretical to an appropriate statistical model involves a number of steps and decisions: decisions that need to be taken and enter the analysis as auxiliary assumptions. To begin with, solving the theoretical model is straightforward here. Let  $p$  denote the probability that the leader's state wins. The leader will wage war if  $pu_w - c + e_w > u_r + e_r$ , and accept the offer otherwise.<sup>6</sup> With the assumption that the difference in the disturbance has some cumulative density function,  $F$ , the probability of waging war is

$$\begin{aligned} q &= \Pr(\text{War} \mid u_w, u_r, p, c) \\ &= \Pr(pu_w - u_r - c > e_r - e_w) \\ &= F(pu_w - u_r - c) \end{aligned}$$

In words, the *theoretical model* implies that the probability that a leader wages a war is

given by the value of the cumulative density function at the net utility of war.

The nice thing here is that the theoretical model easily translates into a statistical model, as the above equation looks like a standard binary discrete choice model. If we denote the observation of a leader going to war by  $y = 1$  and backing down by  $y = 0$ , we can think of the term  $y^* = pu_w - u_r - c$  as the latent preference for war. When the  $e_r - e_w$  are independently drawn from a normal distribution with mean zero, variance  $\sigma^2$  and CDF  $\Phi$ , the probability that the leader goes to war becomes

$$\Pr(y = 1) = \Phi\left(\frac{pu_w - u_r - c}{\sigma}\right).$$

To complete the statistical model, we need to do two more things.<sup>7</sup> First, we need to operationalize the features of the theoretical model,  $p$ ,  $u_w$ ,  $u_r$ , and  $c$ , and then get measures for them. Measurement involves issues of validity and reliability that we do not deal with here. But there is one point we want to stress: suppose that we have a reasonable measure for the leader's *ex ante* probability of winning the war, say  $\pi$ , and also that all relevant costs are material costs that can be estimated *ex ante* in monetary terms, denoted by the measure  $C$ . What is less clear is the utility that the leader associates with the two states of winning a war or accepting the deal. Some observable features of the territory – such as size, geography, natural resources, and physical capital stock – will be important, but as these goods are not easily convertible, there is no a priori rule as to how exactly different goods jointly determine a utility representing preferences.

As a consequence, we have to make auxiliary assumptions to fully specify the statistical model. To keep things simple, let us assume the following:

- $u_r$  and  $u_w$  are fully determined by two observable features of the territory: its size and its natural resources
- $X$  and  $Z$  are valid and reliable measures for size and resources



- all leaders value  $X$  and  $Z$  in the same way
- the functional form of the relationship between observable features and theoretical concepts is linear, such that  $u_r = \alpha_0 + \alpha_1 X + \alpha_2 Z$  and  $u_w = \beta_0 + \beta_1 X + \beta_2 Z$  for some unknown parameters  $\alpha$  and  $\beta$ .<sup>8</sup>

This is quite a list of assumptions. It is, however, what we do, often implicitly, when estimating models. With the above assumption, we can write the latent preferences as

$$y^* = \pi \times (\beta_0 + \beta_1 X + \beta_2 Z) - (\alpha_0 + \alpha_1 X + \alpha_2 Z) - \gamma C.$$

This is not a linear relationship:

$$y^* = -\alpha_0 - \alpha_1 X - \alpha_2 Z - \gamma C + \beta_0 \pi + \beta_1 \pi \times X + \beta_2 \pi \times Z$$

but involves two interaction terms of  $\pi$  and  $X$  or  $Z$ , respectively. The important point here is that a simple linear relationship between the regressors and the odds-ratios of going to war does not capture the mechanism behind the choice problem of the leader.

This will even be less so if there is not a single decision-maker, but two or more actors whose actions and strategies are strategically interdependent. To illustrate, consider a slightly more complex game that considers the strategic interaction of the leader and the opponent state. The opponent state can either prepare for war or not, and either action is observed by the leader. Preparing for war comes at a cost of  $d$  for the opponent, but lowers the probability that the leader wins to some  $p_0 < p$ . While the choice problem of the leader is basically the same – she would account for the opponent's war preparations with  $p_0$  or  $p$  – the opponent has a more intricate decision problem. Whether preparing (or not) is preferable largely depends on the probability that the leader wages war (or not), which again is conditional on the war preparations of the opponent. Using analogous specifications for utility functions  $v$  of the

opponent, the opponent's choice to prepare is the best strategy if

$$q((1 - p_0)v_w - d) + (1 - q)(v_r - d) + e_p > q(1 - p)v_w + (1 - q)v_r + e_n$$

In this case, the statistical model will involve simultaneous estimation of the parameters of the covariates and the choice probabilities. A specification of the statistical model that fails to capture this interaction will not be informative about the causal relationships postulated by the model.

There is a considerable literature on how to derive a statistical model from a theoretical model involving strategic interaction. In the econometric literature, the approach is referred to as structural estimation (Low and Meghir, 2017), though the political science literature is sparse (see, however, Smith, 1999; Signorino, 1999, 2003). The bad news here is that there is no vanilla or one-size-fits-all method that could readily be applied to very different types of (game-)theoretical models. Decision-theoretical discrete choice problems (as the one above) will often lead to some sort of the statistical random utility model (McFadden, 1974). These may differ in assumptions about the distribution of the unobserved, random utilities (resulting in probit or logit models), and also in the source of the uncertainty they capture (Signorino, 2003): agents may simply err when making choices; they may have private information about their own payoffs; or random utilities represent what is actually omitted variables or measurement error of the covariates that are used to capture features of the utilities.

## EITM RESEARCH DESIGNS

There is a great variety of approaches to combine theory building, model construction, and empirical work in political science research. In what follows, we distinguish between three common approaches: (1) equilibrium point predictions; (2) comparative statics; and (3) structural estimation approaches. Throughout, as a running example, we will use a simple

②			
		E	¬E
①	E	$\Pi^{Duo} - c_1$	$\Pi^{Mono} - c_1$
		$\Pi^{Duo} - c_2$	0
¬E		0	0
	¬E	$\Pi^{Mono} - c_2$	0

**Figure 7.2 Entry Game**

entry game (see Figure 7.2, and see Berry and Reiss, 2007), a strategic situation that is ubiquitous in economics and political science, to introduce each research design.

In its original version, two firms decide whether to enter a market ( $E$ ) or not ( $\neg E$ ). Likewise, we can think of politicians deciding whether to enter an electoral race or not. Entering the market comes at an entry cost  $c_1$  for firm 1 and  $c_2$  for firm 2. If only one firm enters the market, the variable profit is  $\Pi^{Mono}$ . If both firms enter the market, their profits are reduced to  $\Pi^{Duo}$ . If a firm decides not to enter the market, its payoff is 0. We will assume that  $\Pi^{Mono} = 2\Pi^{Duo} = \Pi$  and  $\Pi > c_1, c_2 > \Pi/2$ , which means that each firm will prefer to enter the market if the other firm does not, but will prefer to stay out if the other firm enters. Players know profits and costs and make their choices simultaneously. The game has three Nash equilibria: two pure strategy equilibria ( $E, \neg E$ ) and ( $\neg E, E$ ) and one equilibrium in mixed strategies

$$\left( \left( \left( 2 - \frac{2c_2}{\Pi} \right) E, \left( \frac{2c_2}{\Pi} - 1 \right) \neg E \right); \right. \\ \left. \left( \left( 2 - \frac{2c_1}{\Pi} \right) E, \left( \frac{2c_1}{\Pi} - 1 \right) \neg E \right) \right).$$

Having solved the game, a researcher can confront the theoretical model with data in at least three different ways. First, a researcher can focus on the evaluation of the point predictions of the model. Does the empirical outcome or the employed strategies of real

actors match the equilibrium outcome and the equilibrium strategies? The evaluation of point predictions is often useful in laboratory experiments, but has only limited application in large- $N$  studies relying on observational data. Another important research approach that makes use of equilibrium predictions of formal models are analytic narratives.

Second, a researcher can focus on the relationships between variables that the equilibrium of the game implies. Such comparative statics designs usually do not claim that the formal model they use captures the entire data generating process. Instead, they will seek to evaluate relationships between key parameters of the model empirically. Comparative statics is the dominant research design when combining a formal model with observational data.

Finally, instead of seeking to verify or falsify predictions of the model, a researcher might be interested in structurally estimating parameters of the model, either because they are interesting on their own or because they are useful in making predictions for counterfactual scenarios. We will next consider these approaches in more detail.

## EQUILIBRIUM POINT PREDICTIONS

Formal and game-theoretic models are solved by the application of some sort of solution

concept, a rule that defines a priori how to single out events that can be expected and, hence, are 'predicted' from the universe of events. For example, models using non-cooperative game theory usually apply the concept of the Nash equilibrium, or some refinements thereof, to make predictions about strategies players use and the resulting outcome of a strategic situation. Models in the cooperative game-theoretical tradition make use of solution concepts such as the core, the Shapley value, or the Walrasian equilibrium that focus on payoff allocations rather than strategies. An appealing characteristic of many solution concepts from a theoretical point of view is that they often make sharp, deterministic predictions.

For example, in the entry game, a sharp equilibrium point prediction can be derived using mixed strategy equilibrium. Let's assume that the total profit is  $\Pi = 7$  and firms have differing costs for entering the market  $c_1 = 5$  and  $c_2 = 4$ . Then the mixed strategy equilibrium prediction is that the first firm enters the market with probability  $\frac{6}{7}$  and stays out with probability  $\frac{1}{7}$ , whereas the odds that firm 2 enters or stays out are  $\frac{4}{7}$  and  $\frac{3}{7}$ . These are fairly sharp predictions, but how can we use them to evaluate the model?

There are at least four issues. First, as we have shown above, the model has multiple equilibria: firm 1 entering and firm 2 staying out is as much a Nash equilibrium in pure strategies as the reverse case. There is nothing that makes one superior to the other. Second, even if there were only the mixed strategy equilibrium, none of the four observable outcomes could be taken as evidence either for or against the proposed model mechanism. The issue here is that we can only observe behavior, not the actual strategies that individuals use. If both firms indeed use mixed equilibrium strategies, we should observe, for instance, a duopoly with probability  $\frac{6}{7} \times \frac{4}{7} = \frac{24}{49}$  and no firm entering the market

with probability  $\frac{1}{7} \times \frac{3}{7} = \frac{3}{49}$ . A straightforward way to evaluate mixed strategy equilibria is thus to focus on these aggregate outcome probabilities, and compare them to relative frequencies calculated from a larger number of observed cases, preferably in a controlled laboratory environment.<sup>9</sup> Third, how to measure  $\Pi$ ,  $c_1$ , and  $c_2$  in a real market? As we can see above, point predictions of strategies are vastly different when the costs of the two firms are 5 and 4 units, or 4 and 5 units. Fourth, how can we make sense of these point predictions if we are not willing to assume that the model reflects the complete data generating process? Even if the model does not miss any systemic feature of the interaction, observed behavior and outcomes will involve some randomness, leaving uncertainty as to their effect and the manner in which to account for it in the model. From an empirical point of view, sharp predictions can be less appealing since the world is generally stochastic, and sharp predictions are easily falsified. We are rarely able to study the real world 'in equilibrium'.

How should we assess point predictions? There are three basic routes that one can take (see Morton, 1999). One is to make explicit assumptions about the nature of the random effects on the point predictions that may have prevented those predictions from being observed empirically. This will transform point predictions into distributional predictions. If there is a sufficiently large number of observations, distributional predictions are then assessed against the empirical record. In the market entry game, one could assume that firms perfectly know market profits and costs, but err when making their choice. This results in the random utility framework that we discussed in the section above. In the entry game, taking into account that firms err, the probability  $p_1$  that firm 1 enters the market may be written as

$$p_1 = Pr \left( p_2 \left( \frac{\Pi}{2} - c_1 \right) + (1 - p_2)(\Pi - c_1) + \varepsilon_{11} > \varepsilon_{12} \right)$$

where  $p_2$  is the probability that firm 2 enters the market, and the  $\varepsilon$ s are individual and choice specific error terms. Let  $F$  again denote the CDF of  $\varepsilon_{12} - \varepsilon_{11}$  and  $\varepsilon_{22} - \varepsilon_{21}$ . The equilibrium predictions are the implicitly defined distributions

$$p_1 = F\left(-\frac{\Pi}{2}p_2 + \Pi - c_1\right)$$

$$p_2 = F\left(-\frac{\Pi}{2}p_1 + \Pi - c_2\right).$$

Thus, a solution to the problem – that point predictions are likely to fail to receive support because of a partially modeled data generating process – is to turn the model into something more ‘realistic’ and/or use an appropriate equilibrium concept, here a quantal response equilibrium (McKelvey and Palfrey, 1995).

A second way is to use controlled laboratory experiments. Laboratory experiments have a number of advantages for evaluating theories, in general, and equilibrium point predictions, in particular. A major advantage of experiments is that they can isolate the theoretical mechanism. The idea is to eliminate as many random effects that occur in the natural environment as possible. The third route is different from the previous ones in that it builds and evaluates a model in the context of a limited number of cases, often a unique case. Historical events such as the French revolution or the rise and fall of Venice may be instructive for our understanding of political mobilization and institutional design, but their study is plagued with the typical problem of case studies: there are different ways to interpret the historical record. Analytic narratives (Bates et al., 1998) is the attempt to provide a causal explanation for such events by building a formal model to capture the logic of the explanation and evaluating it through testable implications. In that sense, analytic narratives is less a method to simply evaluate a model, and more an approach of both model building and evaluation. We present

showcases of the latter two approaches in the remainder of this subsection.

### ***Evaluating Equilibrium Point Predictions in the Lab: Battaglini et al. (2010)***

An example of the evaluation of the point predictions of a model in a laboratory is Battaglini et al. (2010). Their starting point is the theoretical model that introduced the *Swing voter’s curse* (Feddersen and Pesendorfer, 1996). Feddersen and Pesendorfer show that it can be rational for poorly informed voters to abstain in an election, even when the cost of voting is zero. By abstaining in equilibrium, poorly informed voters avoid deciding the election in the wrong direction and, therefore, leave the decision to better informed voters. A central assumption of the model is that voters form beliefs about the probability that their vote will be pivotal. The model matches some empirical observations, such as selective abstention, but the pivotal voter assumption, along with the theory of strategic abstention, remains controversial.

Battaglini et al. use a simplified version of the model. A set of voters decides by majority rule between two alternatives, A and B. Corresponding to the two alternatives, there are two unobserved possible states of the world, A and B. In state of the world A, alternative A is optimal; in state of the world B, alternative B is optimal. There is a number of independent voters (or swing voters) who want to match the alternative and the state of the world, as well as some partisan voters who prefer alternative A or B, irrespective of the state of the world. Each voter may receive a private informative signal about the state of the world, but may also stay uninformed with some probability. After voters have received the signal, votes are cast, and an alternative is chosen by majority rule. In equilibrium, all partisan voters vote in favor of their preferred policy,

and all independent voters who received an informative signal vote in line with their signal. Uninformed independent voters, however, have incentives to abstain with at least some probability. The reason is that an uninformed independent voter knows that, with some probability, there exists an informed independent voter. By casting a vote, the uninformed independent voter would risk voting against an informed independent voter. The key equilibrium point prediction of the model is, therefore, that independent uninformed voters will abstain with some probability depending on the specific parameterization of the game.

The authors set up a laboratory experiment to evaluate the point predictions of the theoretical model. Values for three different parameters need to be chosen: the number of voters, the number of partisan voters and the prior probability for each state of the world. There are two possible jars (states of the world). The first jar contains six white balls and two red ones, while the second jar contains six white and two yellow balls. A computer randomly chooses a jar with the prior probability. On the computer screen, participants choose one of the eight balls, the color of which is then revealed. This creates a share of informed participants, namely those who see a yellow or red ball, and some uninformed participants, who see a white ball. Finally, each participant decides to either vote for jar A or jar B or to abstain. If the correct jar is chosen by a majority, a higher payoff is paid out compared to the incorrect jar. If there were partisan voters, the participants are told that the computer will cast their votes for jar A and B, respectively. Participants' choices are recorded and compared with the point predictions of the theoretical model.

Battaglini et al. find strong evidence that participants' voting behavior varies as theoretically predicted: depending on the treatment configuration, uninformed subjects abstain, seemingly delegating the choice to informed participants.

### ***Analytical Narratives: Gailmard (2017)***

In US federal and state constitutions, the separation of power is very pronounced, with governors and assemblies having independent power bases. How did this hallmark of US constitutions evolve? Gailmard (2017) argues that, to understand the origins of the separation of powers in the United States, we must look at the era of English colonies in North America. What nowadays looks like a natural institution is actually the result of an institutional choice by a strategic English Crown confronted with agency problems in its colonial governors.

Crown-appointed governors in the New World were hard to control from a distance, and as a result, they would over-tax settlers, thereby reducing the incentive for settlers to invest and reducing the revenue to the Crown. To help settlers restrain the governor, the Crown created liberal institutions, such as an empowered assembly with budget power. While this might have negative effects on the Crown's revenue in the short run, it should pay off in the long run. By turning over some political control to the colonial settlers – introducing the separation of powers – the Crown could solve the agency problem.

The theory embeds a model of the colonial economy into two alternative political environments, one with hierarchical control and one with a separation of powers within the colony. There are three players: the Crown, a colonial governor, and colonial settlers. Settlers can make either a high or a low investment in the colony, and then the governor and the Crown extract resources from the colony's economy. In the first, hierarchical model, the economy is embedded into a moral hazard model of political agency (Ferejohn, 1986). While necessary for military security and economic administration, the colonial governor has incentives to extract rents from settlers that, in turn, decrease the economic investment of settlers and the Crown's revenues. In the separation

of powers model, the Crown chooses to let the settlers determine a budget for the government. This constrains the governor and ensures high investments from the settlers, which in turn increase the Crown's revenue. Gailmard then shows that separation of powers is optimal for the Crown when returns to investment are moderate. Specifically, he establishes the conditions for the environment (parameters of the model) under which the Crown's choice of separation-of-power institutions (strategy) is the equilibrium (point prediction).

Gailmard's analytic narrative approach offers a new way of looking at the evolution of an important, widespread institution. Interestingly, the model suggests that separation of powers was neither invoked to control the Crown (by an assembly) nor invented by the Crown to tie its own hands. Rather, it was designed to empower settlers to restrain the governor. The formal modeling approach also helps address the reasons why this distinct form of separation of powers was unique to North America. An investigation of the model parameters suggests that unlike colonies that provided opportunities for natural resource extraction, economic growth in the North American colonies required settler investment in agriculture. In an even broader perspective, the approach offers insights on American institutionalism but may also be taken up by research in a range of other, colonial or authoritarian, contexts of institutional choice.

It should also be clear that, with analytic narratives, theoretical and empirical work are not delegated to distinct steps in the research process, but rather developed more in a dialogue.

## COMPARATIVE STATICS

Comparative statics asks how an equilibrium quantity of interest changes as some exogenous features change (Silberberg and Suen,

2000). As we have argued above (Figure 7.1), these quantities of interest are outcomes,  $f$ , in a broad sense: features that are associated with the real world and that we argue or predict to occur or vary conditional on parameters  $\theta$  such as actions, strategies or some policy (or outcome). For example, in the entry game above, we could be interested in how the probability that both firms enter the market will change as profits increase. The answer is straightforward. In the mixed strategy equilibrium, the probability that both firms enter is  $p_1 p_2 = 4\left(1 - \frac{c_1 + c_2}{\Pi} + \frac{c_1 c_2}{\Pi^2}\right)$ . Then the positive partial derivative of  $p_1 p_2$  with respect to  $\Pi$ ,

$$\begin{aligned} \frac{\partial(p_1 p_2)}{\partial \Pi} &= 4\left(\frac{c_1 + c_2}{\Pi^2} - 2\frac{c_1 c_2}{\Pi^3}\right) \\ &= 4\left(\frac{c_1(\Pi - c_2) + c_2(\Pi - c_1)}{\Pi^3}\right), \end{aligned}$$

tells us that, when profits increase, it is more likely that both firms enter the market. Less straightforwardly, though, it also holds that sensitivity to profits varies with entry costs. To be precise

$$\frac{\partial^2(p_1 p_2)}{\partial \Pi \partial c_1} = \frac{4(\Pi - 2c_2)}{\Pi^3} < 0,$$

as 1's costs of entry increase, the incentive provided by higher profits weakens. The simple example demonstrates the reason that comparative statics is attractive. First, it is the tool for the 'comparative analysis' of formal models, whether used in case study, comparative, experimental or statistical method design. Second, it is often difficult to exactly measure model parameters as they are unobserved or latent (see example in the third section of this chapter). As long as we know that two cases differ with respect to some parameter, comparative statics can tell us whether and how equilibria or outcomes differ.

Comparative statics is often tricky because when one parameter, such as profit or cost, changes, the resulting change in the strategy of one actor also leads to changes in the strategies of other actors. Solving these models often involves using systems of differentiated equations.

A standard technique to derive comparative statics is to assume specific functional forms about the model primitives, such as that utility is linear in profits and costs (as above), or concave in profits but convex in costs. The reason is that standard techniques to find equilibria and to study their comparative statics usually require continuity of best responses, compactness of strategy spaces, differentiability of utility functions and interior solutions, among others.

Monotone comparative statics (Milgrom and Shannon, 1994) explores properties of games that make such strong assumptions or explicit functional forms superfluous. As Ashworth and Bueno de Mesquita (2006: 214–15) note, monotone comparative statics greatly facilitates the empirical evaluation of theoretical models. On the theory side, not having to assume specific functional forms makes the model robust against many sorts of misspecification, and brings the deep structure of the model to light. On the empirical side, replacing compact strategy spaces (the standard assumption) with partially ordered spaces allows for the generation and evaluation of predictions that are based on just ordinal information. More recently, monotone comparative statics techniques have been extended to aggregate games (Acemoglu and Jensen, 2013) and distributional comparative statics (Jensen, 2017).

### ***A Full Set of Empirical Implications: Snyder and Strömberg (2010)***

Snyder and Strömberg (2010) provide an excellent example of how comparative statics can be used. They investigate the effect of political newspaper coverage on policies at the constituency level. Their argument takes

the form of a chain of relationship predictions between variables that ultimately link newspaper coverage and policies: increased newspaper coverage increases citizens' level of political information, which strengthens the monitoring of politicians who, in turn, work harder for their constituents, finally resulting in better policies. The authors could have chosen to collect data on coverage and policies to test their key hypothesis and stop there. They didn't. Instead, the authors decided to provide empirical evidence for each step in the causal chain, achieving a close connection between theoretical argument and empirics.

To identify the effect of newspaper coverage on policies, they exploit exogenous variation in the congruence between congressional districts and newspapers. The authors expect that the larger the overlap between these districts, the higher the political coverage of local politicians because there exists a higher readership share in the district. As a first empirical step, it is shown that newspaper coverage is indeed increasing in the readership share of a district. The next step in the mechanism postulates that the more extensively newspapers cover local politicians, the better informed the citizens. Using data from the American National Election Study over 20 years, the authors demonstrate that respondents in more congruent districts are more likely to receive their news from newspapers or magazines. Respondents living in more congruent districts are also informed about their local representatives: they are more likely to recall the name of at least one representative and are more willing to place them on ideological scales. Following the causal chain argument, the authors continue to show that congressmen from more congruent districts vote against the party line more often and are more likely to stand witness before congressional hearings and to serve on constituency-oriented committees. Finally, as a last step, the authors provide evidence that more congruent districts receive higher federal expenditures.

While Snyder and Strömberg do not present a formal model for their argument, their

work stands out as a demonstration of how a series of comparative statics can illuminate a theoretical mechanism.

## STRUCTURAL ESTIMATION

Structural estimation turns the logic of the previous two research designs – point predictions and comparative statics – on its head. Instead of using data to scrutinize a theoretical model by evaluating point or relationship predictions, structural approaches assume that the theoretical model is a good approximation of the real world and use data to obtain estimates of model parameters. We see at least two main reasons to derive and estimate structural econometric models.

First, so-called deep model parameters are frequently interesting and useful. For example, political scientists are often interested in the policy preferences of political actors, such as legislators or Supreme Court justices. However, these policy ideal points are generally not directly observable and have to be estimated. Ideal point models, such as Nominat (cp. Poole, 2005), posit a behavioral model (including assumptions about the dimensionality of the policy space) that links deep parameters (here: ideal points) to observables (here: roll call votes). Then, the model can be estimated using roll call data, and the actors' ideal points can be recovered and find use in many contexts (for example, the study of polarization in US politics). Other examples of interesting deep parameters are risk aversion, discount factors, marginal cost, and the value of information (see first example given below, Iaryczower and Shum, 2012).

Second, structural models can be used to assess the effect of hypothetical policy interventions by conducting simulations. Having described and estimated a structural model, a researcher can treat the estimated parameters as fixed and make predictions about equilibrium outcomes as the environment changes. For example, a researcher may be interested

in the effect of a reform that decreases the minimum required sentence on pre-trial bargaining in criminal cases (see second example below, Silveira, 2017). In many cases, it is desirable to obtain estimates of effects of policy interventions *ex ante* because small-scale field experiments are infeasible or unethical. In other cases, structural models can help to extrapolate results from a specific (field) experimental setting to other contexts (Wolpin, 2013).

To demonstrate a structural estimation approach, consider again the entry game example (also discussed in Ellickson and Misra, 2011). Imagine we have collected data of local markets, including information on entry decisions of each player as well as firm- and market-specific characteristics (e.g. population, distance to nearest distribution center) that may affect each firm's payoffs. Let  $X_k$  denote characteristics of market  $k$ ;  $Z_{ik}$  denote characteristics of firm  $i$  in market  $k$ ; and  $y_{ik}$  denote the entry decision of player  $i$  in market  $k$ . Then we can express the payoff of firm  $i$  in market  $k$  as

$$\pi_{ik} = \alpha'_i X_k + \beta' Z_{ik} + \delta_i y_{-ik} + \varepsilon_{ik}$$

where  $\varepsilon_{ik}$  is a component of a firm's payoff that is unobservable to the researcher. In equilibrium, each firm  $i$  will choose to enter the market  $k$  if  $\pi_{ik} > 0$  and stay out otherwise. Note that the payoff depends not only on firm and market-specific characteristics, but also on the entry choice of the other player,  $y_{-ik}$ .

If we further assume that firms (but not the researcher) observe the  $\varepsilon$ s and that firms make choices simultaneously, the Nash equilibrium can be characterized by the following system of inequalities:

$$y_{1k} = \mathbb{I}[\alpha'_1 X_k + \beta'_1 Z_{1k} + \delta_1 y_{2k} + \varepsilon_{1k} \geq 0]$$

$$y_{2k} = \mathbb{I}[\alpha'_2 X_k + \beta'_2 Z_{2k} + \delta_2 y_{1k} + \varepsilon_{2k} \geq 0].$$

where  $\mathbb{I}[\cdot]$  is an indicator function. Note that these outcome equations constitute a binary simultaneous equation system, an



interdependent structure that creates challenges for estimation and identification.

Estimation can usually not be achieved with econometric techniques that are implemented in standard econometric software. The reason is that, in contrast to single-agent decision-making problems, the payoff of each player depends on the action of the other player (note the  $\delta_{i|y_{-ik}}$  term in the inequalities above). Any estimation approach must therefore, simultaneously estimate both equations. In general, structural approaches often require that researchers tailor estimators precisely to the empirical case.

An identification problem arising in many structural estimation approaches – in general, and in this setting in particular – is that the estimation procedure needs to deal with equilibrium multiplicity. As shown earlier, the entry game has two pure strategy Nash equilibria if the market is only large enough to make one firm entry profitable but not two. Since, given a set of parameter values, two different outcomes are possible (either firm 1 enters the market and firm 2 does not, or vice versa), parameter estimation is not identified. To deal with this problem, the researcher could either specify an equilibrium selection rule (Tamer, 2003); aggregate to model predictions that are not affected by multiplicity (e.g. number of entrants: see Bresnahan and Reiss, 1990); make different model assumptions that yield unique predictions (e.g. sequential moves: see Berry, 1992); or adopt a bounds approach (Ciliberto and Tamer, 2009). An alternative would be to assume that firms do not perfectly observe each other's payoffs, a situation which would result in a game with incomplete information (Rust, 1994; Seim, 2006).

### ***Estimating Deep Parameters: Iaryczower and Shum (2012)***

Iaryczower and Shum (2012) are interested in the question how much information to the contrary is necessary to overcome ideological predispositions of justices in the US

Supreme Court. This 'value of information' is not directly observable in the real world, but a quantity that can only be interpreted in the context of the theoretical model. Iaryczower and Shum set up the interaction between justices as a Bayesian game in which biased justices receive a private, noisy signal about the case, update their beliefs, and vote in favor of or against the defendant. Voting can be either sincere or strategic (as in Feddersen and Pesendorfer, 1996).

The key issue to recover the value of information is to estimate two deep parameters: judge-specific preferences (bias) and information (signal precision). Iaryczower and Shum propose the following two-step estimation: first, the authors estimate, for each judge, choice probabilities for each state of the world (guilty or innocent defendant) as a function of judge and case-specific observables via maximum likelihood. In a second step, bias and precision parameters can be recovered given estimated choice probabilities. Finally, given bias and precision parameters, one can calculate the probability that a justice votes differently than he or she would have voted without case-specific information, which the authors interpret as a measure of the value of information in the court. Tracing this measure over time, the authors find that the value of information has decreased over the past 25 years, suggesting an increasing politicization of the Supreme Court.

Iaryczower and Shum's approach offers an alternative to purely ideological characterizations of Supreme Court justices' behavior by specifically modeling and estimating how ideology interacts with the information available to the justices. Doing so allows new insights into the relative weights that justices put on their preexisting ideological leanings versus the information of the case.

### ***Simulating Counterfactual Scenarios: Silveira (2017)***

Silveira (2017) investigates the effects of several hypothetical policy interventions,

namely potential sentencing reforms, on the outcomes of criminal cases. Litigation is modeled as a two-stage game. The first stage consists of pre-trial bargaining with asymmetric information between a prosecutor and the defendant, reflecting the fact that the vast majority of criminal cases in the United States are resolved by plea bargaining. Defendants, because they know the full extent of their culpability, have more precise information about potential trial outcomes than the prosecutor. The prosecutor offers the defendant a sentence to settle the case as a take-it-or-leave-it offer. If the defendant accepts the offer, the game ends. Otherwise, the case is decided at trial, at which the defendant's private information is revealed with some probability. In equilibrium, the prosecutor's offer is a function of the anticipated trial sentence, the trial costs and the perceived odds of winning. The defendant accepts the offer if he or she is comparably pessimistic about his or her chances in court.

To estimate the model, the author proceeds in two steps. First, he non-parametrically estimates the prosecutor's equilibrium settlement offer function using information on the prosecutor's settlement offers when plea bargaining was successful, on sentences assigned at trial when plea bargaining failed and on the probability that plea bargaining was successful. Having identified the prosecutor's optimal offer function, the model primitives (i.e. the distribution of potential sentences at trial, the probability of conviction and the trial costs for defendant and prosecution) can be recovered. The author uses these estimates to conduct counterfactual simulations to explore the effects of lower mandatory minimum sentences and a general ban on plea bargaining. In the first case, a decrease in the minimum mandatory sentence leads to a general decrease in incarceration time due to the fact that prosecutors offer lower settlements during plea bargaining. However, it also leads to an increase in conviction rates because shorter potential trial sentences increase the

probability of a successful plea bargain. In the case of a complete elimination of plea bargaining, all cases would go to trial. The model suggests that conviction rates would decrease substantially; however, defendants would face longer sentences. Thus, the average defendant would be worse off if plea bargaining were to be banned.

Silveira's analysis yields insights into potential sentencing and litigation reforms that would be challenging to assess otherwise *ex ante*. Furthermore, experimental research that could provide *ex post* evidence about the effects could be difficult to implement and is potentially unethical.

## PRACTICAL ADVICE

The development of internal specialization in political science is not a new phenomenon. More than 60 years ago, the APSA Committee on Standards of Instructions noted that 'the political scientist increasingly finds the body of political knowledge so great, and the tools of study so exacting, that he must specialize if he is to master and to communicate his subject matter' (APSA Committee on Standards of Instruction, 1962: 417). Nor has there been a shortage of ideas about methodological unification, with the accomplishments of the Cowles Commission in economics dating all the way back to the 1930s. The EITM approach is a genuine political science approach to break down the isolation of formal and empirical modeling, tailored to the specific questions and methodological problems in political science research.

It would be the best of all possible worlds if researchers were highly skilled in both theoretical and empirical analysis, and this was paired with in-depth knowledge of the substantive field of their interest. However, becoming a skillful formal modeler or data analyst requires an enormous amount of training and the time constraints of PhD education

usually prevent students from becoming experts in both fields. Most PhD students become one of two types: the formal modeler who would like to explore the data or the empiricist who would like to add more depth to her empirical analysis.

Also, career-wise, it can be more productive to persuade one audience than to try to appeal to two at the same time. This applies to both faculty members and reviewers. As editor and reviewers, we saw many excellent papers struggle in the review process because either the theory section did not persuade the theorists or the empirics did not convince the empiricists. Given space constraints in journals, authors often end up cutting the theoretical or the empirical part of their paper or relegating large parts to the appendix in the revisions.

For both types of students, we want to offer some brief (hopefully helpful) advice from our experience as researchers, teachers, editor, and student.

**Not every model is for empirical ‘testing’.** (Nor does the EITM approach suggest so.) Models serve different purposes. They can be foundational, organizational, explanatory or predictive (Clarke and Primo, 2012). Being apt for the derivation of testable implications is one purpose, but testability in the hypothetico-deductive methodology is not the only one. Ultimately, models are just tools to help us understand the world.

**Not every empirical analysis needs a formal model.** For some research questions, the mechanisms behind the data are not of primary interest – at least, not in the first step of research. Whether tobacco or gun control programs have causal effects is a relevant question to answer, even if we do not know how policies change individual perceptions, attitudes, and behavior. At the same time, without a proper theory of what is going on, simply relying on identification techniques such as instrumental variables or regression discontinuity designs can lead to the misinterpretation of the causal effect. As Eggers (2017) shows, there is no such thing as *the*

incumbency effect estimated from regression discontinuity.

**Collaboration.** Where true versatility is rare, collaboration seems to be a logical consequence. However, co-authoring does not mean everyone does her own thing. At a minimum, students must learn each other’s mindsets and languages, which is different than just taking a game theory and an advanced statistics class. For example, when theorists write formal models, they will often emphasize existence of equilibria, generalizability of functional forms, and mathematical elegance over whether their model generates sharp, relevant, and empirically falsifiable predictions. It will help formal modelers to understand the interests of empirical researchers and to write models that are accessible to empirical researchers, providing intuition and the reasons that we should care.

**Start simple.** Armed with lots of detailed case knowledge, the empiricist begins writing a model that resembles the case she observes in the real world as closely as possible. As a result, the model becomes quickly extensive, hard or impossible to solve and adds little to the understanding of the theoretical argument. To make it solvable, the empiricist will be forced to make some heroic assumptions that draw criticism from theorists and empiricists alike. The empiricist, we argue, should have started with the core of the theoretical argument. In many cases, the core of the argument can be phrased as a trade-off. Should I protest or stay at home? Should I turn out to vote or not? The starting point, then, is to build the simplest reasonable model that formalizes this trade-off – and make it more complicated, if necessary, from there.

## Notes

- 1 Others, like Johnson (2014), are even more skeptical, rejecting entirely the idea that models can be subjected to empirical evaluation. For him, models are fables to derive a principle from. By rendering some abstract concept like ‘power’

- or 'deterrence' more concrete, models help to understand what the concept means in the particular circumstance captured by the model (Johnson, 2019: e7).
- 2 It is important to emphasize that the (unknown) benchmark is systematic, explainable variation: a model to predict school choice that uses features such as parental education, socio-economic status and personal traits, and that has an R-square of 5%, arguably is mediocre. A stock market model with an R-square of 1% can earn you a fortune.
  - 3 Or, if parsimony is a goal, what should be left out.
  - 4 The term separability goes back to Sono (1945) and Leontief (1947) in the study of production decisions, where separability is the unaffectedness of the ease of substitution between two factors by a third factor.
  - 5 If unambiguous, we suppress indices for the unit of analysis.
  - 6 We deliberately assume that in case of equality, the offer is accepted. Such knife-edge cases are practically irrelevant, as long as (some of) the model parameters have continuous support. This is not the case in, for instance, a standard 2-by-2 pure coordination game where players are indifferent between coordinating on standing on the left or the right of an escalator.
  - 7 One other assumption we must make is a value for  $\sigma$ . As we do not directly observe  $y^*$ , the  $\alpha$ s and  $\beta$ s, together with  $\sigma$ , are not uniquely identifiable. A standard and inconsequential assumption is that  $\sigma = 1$ .
  - 8 A weaker assumption that captures the case of linear relationships would be to assume that all partial derivatives of the  $us$  with respect to  $X$  and  $Z$  are positive, and all higher-order cross-partials to be zero. The statistical model would then be  $Pr(y = 1 | \pi, X, Z) = \Phi(\pi f(X, Z) - g(X, Z) - C)$  for some appropriate  $f$  and  $g$  that is amenable to non-parametric estimation.
  - 9 Here, we think of mixed strategies as literal randomizations over pure strategies, but note that there are alternative interpretations (Harsanyi, 1973; Rosenthal, 1979).
- and D. Collier (Eds), *The Oxford handbook of political methodology* (pp. 828–843). Oxford: Oxford University Press.
- APSA Committee on Standards of Instruction (1962). Political science as a discipline. *American Political Science Review*, 56(2), 417–421.
- Ashworth, S. and Bueno de Mesquita, E. (2006). Monotone comparative statics for models of politics. *American Journal of Political Science*, 50(1), 214–231.
- Bates, R. H., Greif, A., Levi, M., Rosenthal, J.-L. and Weingast, B. R. (1998). *Analytic narratives*. Princeton, NJ: Princeton University Press.
- Battaglini, M., Morton, R. B. and Palfrey, T. R. (2010). The swing voter's curse in the laboratory. *The Review of Economic Studies*, 77(1), 61–89.
- Berry, S. and Reiss, P. (2007). Empirical models of entry and market structure. In M. Armstrong and R. Porter (Eds), *Handbook of industrial organization*, volume 3 (pp. 1845–1886). Amsterdam: North-Holland/Elsevier.
- Berry, S. T. (1992). Estimation of a model of entry in the airline industry. *Econometrica*, 60(4), 889–917.
- Blackorby, C., Primont, D. and Russell, R. R. (1998). Separability: a survey. In S. Barberà, P. J. Hammond and C. Seidl (Eds), *Handbook of utility theory*, volume 1 (pp. 49–92). Dordrecht: Kluwer.
- Bresnahan, T. F. and Reiss, P. C. (1990). Entry in monopoly market. *The Review of Economic Studies*, 57(4), 531–553.
- Ciliberto, F. and Tamer, E. (2009). Market structure and multiple equilibria in airline markets. *Econometrica*, 77(6), 1791–1828.
- Clarke, K. A. and Primo, D. M. (2007). Modernizing political science: a model-based approach. *Perspectives on Politics*, 5(4), 741–753.
- Clarke, K. A. and Primo, D. M. (2012). *A model discipline: political science and the logic of representations*. Oxford University Press.
- Eggers, A. C. (2017). Quality-based explanations of incumbency effects. *The Journal of Politics*, 79(4), 1315–1328.
- Ellickson, P. B. and Misra, S. (2011). Estimating discrete games. *University of Rochester, Working Paper*.

## REFERENCES

- Acemoglu, D. and Jensen, M. K. (2013). Aggregate comparative statics. *Games and Economic Behavior*, 81, 27–49.
- Aldrich, J. H., Alt, J. E. and Lupia, A. (2008). The EITM approach: origins and interpretations. In J. M. Box-Steffensmeier, H. E. Brady

- Feddersen, T. J. and Pesendorfer, W. (1996). The swing voter's curse. *The American Economic Review*, 86(3), 408–424.
- Ferejohn, J. (1986). Incumbent performance and electoral control. *Public Choice*, 50(1/3), 5–25.
- Gailmard, S. (2017). Building a new imperial state: the strategic foundations of separation of powers in America. *American Political Science Review*, 111(4), 668–685.
- Giere, R. N. (2010). *Explaining science: a cognitive approach*. Chicago: University of Chicago Press.
- Harsanyi, J. C. (1973). Games with randomly disturbed payoffs: a new rationale for mixed-strategy equilibrium points. *International Journal of Game Theory*, 2(1), 1–23.
- Iaryczower, M. and Shum, M. (2012). The value of information in the court: get it right, keep it tight. *The American Economic Review*, 102(1), 202–237.
- Jensen, M. K. (2017). Distributional comparative statics. *The Review of Economic Studies*, 85(1), 581–610.
- Johnson, J. (2014). Models among the political theorists. *American Journal of Political Science*, 58(3), 547–560.
- Johnson, J. (2019). Formal models in political science: conceptual, not empirical. *The Journal of Politics*, 81(1), e6–e10.
- Keane, M. P. and Wolpin, K. I. (2007). Exploring the usefulness of a nonrandom holdout sample for model validation: welfare effects on female behavior. *International Economic Review*, 48(4), 1351–1378.
- Leontief, W. (1947). Introduction to a theory of the internal structure of functional relationships. *Econometrica (pre-1986)*, 15(4), 361.
- Low, H. and Meghir, C. (2017). The use of structural models in econometrics. *Journal of Economic Perspectives*, 31(2), 33–58.
- McCubbins, M. D. and Schwartz, T. (1984). Congressional oversight overlooked: police patrols versus fire alarms. *American Journal of Political Science*, 28(1), 165–179.
- McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Frontiers in Econometrics* (pp. 105–142). New York: Academic Press.
- McKelvey, R. D. and Palfrey, T. R. (1995). Quantal response equilibria for normal form games. *Games and Economic Behavior*, 10(1), 6–38.
- Milgrom, P. and Shannon, C. (1994). Monotone comparative statics. *Econometrica*, 62(1), 157–180.
- Morton, R. B. (1999). *Methods and models: a guide to the empirical analysis of formal models in political science*. Cambridge: Cambridge University Press.
- Mullainathan, S. and Spiess, J. (2017). Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, 31(2), 87–106.
- National Science Foundation (2002). *Empirical implications of theoretical models report*. Arlington, Virginia.
- Poole, K. T. (2005). *Spatial models of parliamentary voting*. Cambridge: Cambridge University Press.
- Popper, K. R. (1959). *The logic of scientific discovery*. New York: Basic Books.
- Rosenthal, R. W. (1979). Sequences of games with varying opponents. *Econometrica*, 47(6), 1353–1366.
- Rust, J. (1994). Structural estimation of Markov decision processes. In R. F. Engle and D. L. McFadden (Eds), *Handbook of econometrics*, volume 4 (pp. 3081–3143). Amsterdam: North-Holland/Elsevier.
- Seim, K. (2006). An empirical model of firm entry with endogenous product-type choices. *The RAND Journal of Economics*, 37(3), 619–640.
- Signorino, C. S. (1999). Strategic interaction and the statistical analysis of international conflict. *American Political Science Review*, 93(2), 279–297.
- Signorino, C. S. (2003). Structure and uncertainty in discrete choice models. *Political Analysis*, 11(4), 316–344.
- Silberberg, E. & Suen, W. C. (2000). *The structure of economics: a mathematical analysis*. New York: McGraw-Hill, 3rd edition.
- Silveira, B. S. (2017). Bargaining with asymmetric information: an empirical study of plea negotiations. *Econometrica*, 85(2), 419–452.
- Smith, A. (1999). Testing theories of strategic choice: the example of crisis escalation. *American Journal of Political Science*, 43(4), 1254–1283.

- Snyder, J. M. Jr. and Strömberg, D. (2010). Press coverage and political accountability. *Journal of Political Economy*, 118(2), 355–408.
- Sono, M. (1945). The effect of price changes on the demand and supply of separable goods [in Japanese]. *Kokumin Keizai Zasshi*, 74, 1–51.
- Tamer, E. (2003). Incomplete simultaneous discrete response model with multiple equilibria. *The Review of Economic Studies*, 70(1), 147–165.
- Wolpin, K. I. (2013). *The limits of inference without theory*. Cambridge, MA: MIT Press.



# EITM: Applications in Political Science and International Relations<sup>1</sup>

John Aldrich and Jim Granato

Empirical observation, in the absence of a theoretical base, is at best descriptive. It tells one what happened, but not why it has the pattern one perceives. Theoretical analysis, in the absence of empirical testing, has a framework more noteworthy for its logical or mathematical elegance than for its utility in generating insights into the real world. The first exercise has been described as 'data dredging', the second as building 'elegant models of irrelevant universes'. My purpose is to try to understand what I believe to be a problem of major importance. This understanding cannot be achieved merely by observation, nor can it be attained by the manipulation of abstract symbols. Real insight can be gained only by their combination. (Aldrich, 1980: 4)

## INTRODUCTION

EITM, Empirical Implications of Theoretical Models, is a way to design a research project; that is, it *is* a research design. EITM is also a way to develop and integrate a stream of research. And it is a way to place a particular

research project in the context of its relevant scientific community (as per Kuhn, 1970 [1962]). It is all of those because it is a way of organizing and communicating scientific knowledge.

A completed piece of science, that is to say, an addition of new knowledge, done consistently with EITM, does two things. First, it provides an integrated set of hypotheses derived from a set of general premises in a logically consistent and coherent way. Second, it provides a set of empirical observations that reflect on the empirical plausibility of those logically coherent hypotheses. Put alternatively, for EITM, the TM part answers the central question of science: why? Why might this part of the world that is being studied be arranged in the way it is? The EI part answers the question of how likely it is that that logically coherent explanation is 'true' (empirically plausible) in the real world. EITM in total, therefore, answers the questions: 'why did this event/choice/outcome occur?' and 'is that reason for why it

happened likely to be found in the real world again, and if so how often?’

Science may be said to be a combination of art and science. The ‘art’ in its science is where science begins, which is with the asking of an important and interesting question. It is the purpose of the liberal arts in general to train people in how to ask important and interesting questions. In addition, it is also often an art, or at least it is artful, to think up what observations, experiments or other ways of gathering data might best give answers to those questions. As to the science part, it is the purpose of science to answer well thought out questions and to communicate their answers, convincingly, to others. Pre-scientific inquiry<sup>2</sup> may be thought of both as the assessment of the empirical terrain in which that question is to be asked and as the search for suitable premises from which derivation of testable hypotheses may begin.<sup>3</sup>

All of the rigorous standards of science, such as corrigibility, reproducibility and others, only begin at this point. That is, there is no reason to hold a scientist to the standard of being able to reproduce her inspiration that led her to ask the question she wants to study, nor why she believes it to be important, nor even to question whether she saw the correct set of premises she used to develop her derivations by staring at a fire. Which is only to say that the ‘context of discovery’ is not the science. The ‘context of validation’, however, does require fealty to rigorous scientific standards. These are the standards of how a potential new piece of knowledge becomes, indeed, a new addition to our corpus of scientific knowledge, and how that knowledge can be communicated convincingly to others.

If TM is providing a logically coherent answer to the question of why, and EI is the evidence that is used to say that that answer is an empirically plausible one in some significant portion of the world, then EITM is simply an acronym for science itself. Or, more accurately, it is a way to think about what it takes to create new knowledge and

tell others what that knowledge is. That is, it is a way of looking over a research agenda and assessing what has been accomplished so far and what more needs to be done to strengthen the scientific understanding of that important question/problem. EITM is not a way to ‘do’ science so much as it is a way to assess the scientific status of a stream of research.<sup>4</sup>

If you are willing to buy these claims, at least for the moment, then the only real way to proceed is to point to accomplishments – in our case, to do so in the political, social and behavioral sciences, including in the study of international relations. We provide two examples to demonstrate the case. But first, we briefly discuss these general points we made about EITM in the context of its history, made possible by the impetus and continuing support of the National Science Foundation’s (NSF) Political Science Program in the Social, Behavioral, and Economic Sciences. There would be no EITM program without the NSF (and the Political Science Program at NSF in particular).

### ***The Role of NSF***

On July 9 and 10, 2001 the Political Science Program at the National Science Foundation (NSF) convened a workshop seeking ways to improve technical–analytical proficiency in political science.<sup>5</sup> This workshop, termed the Empirical Implications of Theoretical Models (hereafter EITM) Workshop, suggested constructive approaches so the NSF Political Science Program could develop linkages and a dialogue – both methodological and interpersonal – between formal and empirical modeling and modelers.<sup>6</sup>

To put it in more technical terms, the 2001 Workshop highlighted an important contribution the EITM approach brought to both political science and the social sciences:

If one were to summarize in one word what bridging the divide between formal and empirical modeling means for the political, social and behavioral



sciences, that word would be *identification*. The ability of a researcher to *identify* or parse out specific causal linkages among the many factors is fundamental to the scientific enterprise. Specifying a model that links both formal and empirical approaches alerts researchers to outcomes *when specific conditions are in place* – and is also one of the best ways to determine an *identified* relationship (EITM Report, 2002: 1–2).<sup>7</sup>

Moreover, by integrating the two approaches, students would be exposed to the strengths of both approaches:

At the most basic level, formal modeling assists in the ‘construction of valid arguments such that the fact or facts to be explained can be derived from the premises that constitute the explanation.’<sup>8</sup> An important virtue of formal modeling is that it often yields surprising implications that would not have been considered had they not emerged from formal analysis. Conversely, if practiced correctly, applied statistical and case study analysis shows researchers where a model went wrong and leaves open the possibility that a more accurate model can be constructed. (Granato and Scioli, 2004: 314)<sup>9</sup>

With this brief background in mind, EITM Workshop participants recommended, in both written and spoken commentaries, that the NSF Political Science Program address the technical–analytical divide between formal and empirical approaches in three priority areas:

- Education: Training and Retraining<sup>10</sup>
- Dissemination of Knowledge: Conferences and Workshops
- Research: Establishment of Research Work Groups.

In addition to the NSF sponsored initiatives above, an EITM certification program has also been created at the University of Michigan’s ICPSR Summer Program in Quantitative Methods of Social Research (see <http://www.icpsr.umich.edu/icpsrweb/sumprog/> and <http://www.eitminstitute.org/index.html>). The certification program requires that students focus on a set of ‘approved’ courses that provide background for using the EITM approach and attending

the EITM Summer Institutes. Between 2011 and 2013 more than 270 students received certification (see <http://www.eitminstitute.org/recipients.html>).

## EXAMPLES

### ***QRE: Proposed as a Solution to a TM Question and as Applied to an EI Question***

**QRE as a TM question:** Richard McKelvey and Thomas Palfrey developed Quantal Response Equilibrium (QRE) as a proposed solution to something they found puzzling in practice (McKelvey and Palfrey, 1995, 1998). That is, they sought a solution to a question that was theoretical. However, it was empirical observations that led them to ask the theoretical question.

The problem that led to QRE was that empirical observations did not align with the theoretical derivations. In particular, in some repeated games – the centipede game being focal for QRE<sup>11</sup> – there is a well-defined result.<sup>12</sup> In the centipede game, there is a unique Nash equilibrium (even more, it is particularly easy to implement: play the ‘take’ strategy at the first opportunity, ending the game).<sup>13</sup> The theoretical problem with this solution is easy to state: very few people play the centipede game that way. Instead they choose almost every imaginable outcome in the centipede game. And they do so even when they know and understand the ‘correct’ outcome. This was a well-known and serious challenge to game theory, per se, because these were among the few potentially interesting and ‘playable’ games with such clear and certain answers, and yet people clearly choose not to make their ‘best’ choices.

In EITM terms, the context of the problem was game theory, so the theoretical answer to the question was likely to be (and was) a mathematically presented theory.<sup>14</sup> And the TM answer was the derivation from

well-stated premises of what they called the QRE. The empirical observations that led to the TM question were essentially ‘case study’ observations that even knowledgeable players very often rejected the Nash equilibrium solution. Testing whether QRE is a potentially ‘correct’ answer was done via more quantitative approaches using game-experimental data.<sup>15</sup> But just as theory need not be mathematical or formally logical, empirical implications need not be tested with statistical methods.

The formal derivation of QRE itself is open to many interpretations and applications. Their interpretation is but one, and it was theoretical. By changing the specification of one of the parameters in the decision-making part of the game model, they changed the game form from one that had a unique Nash equilibrium to one that not only had, but also was better understood as ‘solved’ by the use of, quantal response equilibriums. There were several ways in which that new specification of the premise could be interpreted. The specific one that McKelvey and Palfrey employed from the outset (see 1995) is that the specified term is a parameter able to be estimated as a statistic, and thus enables QRE to be considered as a statistical problem.

They defined QRE as the ‘statistical version of the Nash equilibrium’ (p. 9), one that results from assuming that, instead of utility being a fixed quantity – where, say,  $u_i(p)$  denotes  $i$ ’s utility for outcome  $p$  – it is instead subject to random variability, where, say  $u_i(p) = u_i(p) + e_i$ . McKelvey and Palfrey (1995) prove the existence of a QRE under general conditions for this random utility function that they call the ‘statistical reaction function’ (p. 10). They then specify a more precise (logit) distribution assumption to prove more specific results which they then can apply to answer their theoretical question: might players be playing according to a QRE, with a random utility function, rather than a fixed utility and playing some Nash or refinement?

**QRE as an EI Question:** Given the setup of the original McKelvey–Palfrey QRE as a ‘statistical equilibrium’, its extension to developing empirical applications was foreseeable. Signorino (1999) not only saw how to exploit that opportunity but also had the insight to see that the uncertainty assumed about the players’ utility functions, when plugged into the sequence of play and solved as a condition of uncertainty, yielded a statistical result, such that the unknown parameters in the functions were distributed according to ‘QRE’ uncertainty as a probit. Signorino thus developed a method of estimation that one might call ‘strategic probit’ and wrote a program, ‘STRAT’, for its estimation (2001). This was a brilliant insight that both allowed for a fealty to the principles of game theory and yet also implied a statistical estimation strategy for estimating the parameters in such a game.

Consider Leblang’s paper on the ‘political economy of exchange rate policy’, estimating a game theoretic model that includes strategies of speculative currency attacks (2003; see also Granato et al. (n.d.)).<sup>16</sup> The players are speculators in the currency market and policy makers in what are (in this paper at least) governments in developing nations. The strategies are fairly simple sets of strategies. Market actors can either do nothing and preserve the status quo, or undertake speculative attacks. The policy makers in the government have nothing to do unless attacked, but if they are attacked, they can either defend against the attack or devalue their currency. While these choices are simple, the set of factors that go into their utility functions are lengthier and more complex, and it is this that provides the uncertainty (e.g., for potential speculators in determining whether the government will defend or devalue). And it is these forces that are estimated in the strategic probit estimation (done with data from 95 developing nations). The government’s decision is shown to depend on many factors, including institutional features, electoral characteristics and partisan forces.<sup>17</sup>

### ***RBC Modeling and its Potential Applications in IR***

**Background:** This EITM application – based on a Real Business Cycle (RBC) approach – relies on modeling and testing techniques tracing back to the 1930s and the Cowles Commission. The Cowles Commission approach gave rise to what we now call econometrics.<sup>18</sup> Among their contributions was the *probability approach*. This approach highlighted the issues of *identification and invariance*.<sup>19</sup> Identification was central since a goal of econometrics is to determine the true values of the parameters consistent with the data and with the known or assumed model properties. The second issue was the invariance of a relation. If the structure is known to remain in the future as it was in the past, and if the auxiliary variables have constant values through both periods, then the path of each variable will be predictable from the past, apart from random disturbances.

The Cowles Commission approach also linked both theory and empirics within a system of equations. Rules for identification (such as the rank and order conditions) were based on the number of equations and unknowns. While standard hypothesis testing was one basis for evaluating theories, the other attribute of this approach was to use out of sample forecasts and simulation (resting on invariance).<sup>20</sup> The latter feature meant that rich theoretical ‘experimental worlds’ (what we now call ‘simulations’) were feasible and allowed for various interventions – policy and otherwise – for evaluation.

In sum, and speaking more broadly, the Cowles Commission approach – in addressing the issues of identification and invariance (and the linkage of formal and empirical analysis) – provides a connection to falsifiability, predictive precision and the workings of a system.<sup>21</sup> It should be noted that models possessing these properties also facilitate comparison between rival theories about the same phenomena and thus can enhance efforts for scientific cumulation (Kuhn, 1979).

While the contributions of the Cowles approach are well known and endure to this day, it also drew criticism. In the 1970s fundamental criticisms arose regarding invariance and identification. The criticisms were leveled on both TM and EI grounds. Solutions were proposed that literally blended both theory and empirics.

Regarding TM, in 1976, Robert Lucas questioned invariance durability, based on how people form expectations when the Cowles approach is used. The problem, he argued, is that in-sample estimation provides little guidance in predicting the effects of policy changes because the parameters of the applied statistical models are unlikely to remain stable under alternative stimuli.<sup>22</sup> Or, to put it another way, there was a failure to link micro level findings in a rigorous way (and accounting for agent expectations and their response) with macro level modeling.

As for EI, the results of Lucas’s theoretical challenge was borne out in sustained periods of inaccurate forecasts that contributed to policy failure. Lucas (1981) notes:

Keynesian orthodoxy or the neoclassical synthesis is in deep trouble, the deepest kind of trouble in which an applied body of theory can find itself: It appears to be giving seriously wrong answers to the most basic questions of macroeconomic policy. Proponents of a class of models which promised 3 1/2 to 4 1/2 percent unemployment to a society willing to tolerate annual inflation rates of 4 to 5 percent have some explaining to do after a decade such as we have just come through. A forecast error of this magnitude and central importance to policy has consequences, as well it should. (pp. 559–60)

### **RBC as a Solution to the TM and EI Challenge:**

One response to Lucas’s TM and EI criticisms was Kydland and Prescott’s (1982) RBC approach. This method involves computational experiments that are based on fusing micro and macro level analysis. The initial RBC program was focused on the role of economic shocks; politics and policy were excluded.

The RBC method involves a sequence of steps including: (1) deriving the equilibrium laws of motion for the model economy from

‘well-tested theory’ on agent income, households and firms; (2) ‘calibrating’ the model using parameter values derived from historical data; (3) generating simulated realizations of the equilibrium processes; (4) determining the sampling distributions of the statistics computed from the simulated data; and (5) comparing these statistics to those computed for actual data. Kydland and Prescott’s (1982) ‘computational experiments’ are often referred to as ‘calibration’ because of the use of parameter values derived from simple measures (such as averages) of historical time series to ‘calibrate’ the theoretical models.<sup>23</sup>

This research methodology evolved in the ensuing years to include a variety of factors, making them amenable to policy interventions (and political factors) as well. This update to RBC modeling is typically referred to as *dynamic, stochastic general equilibrium* (DSGE)<sup>24</sup> modeling.

RBC (and DGSE) establishes an important EITM linkage between theoretical predictions (using various behavioral equations and the merging of the micro and macro levels) with actual observations. The formal tools for RBCs – the EITM linkage – rest in the linkage of utility maximization (decision making) with model calibration (prediction).

In political science, Freeman and Houser (1998) introduced the first RBC application but its focus was in comparative political economy. On the other hand, we find an IPE ‘policy’ example in Ghironi and Melitz (2005). They build on Melitz’s (2003) model and develop a DSGE model of international trade that includes the effects of (de)regulation on entry.

Instead of specifying a purely economic explanation based on the behavior of traded and nontraded sectors,<sup>25</sup> Ghironi and Melitz’s model provides an endogenous – policy based – explanation for the Harrod-Balassa-Samuels (HBS) effect (Kravis and Lipsey, 1983). Specifically, policy shocks (i.e., deregulation) influence firms’ decisions to enter or exit in both domestic and export markets. This behavior, in the aggregate, affects real exchange rate dynamics – and macroeconomic outcomes.

A statement of the model revealing the role of policy can be found in the Appendix. The complete system (and the calibrations and simulations that follow) can be summarized below (the variable list is given in Table 8.1):

### 1 Household: utility maximization

Euler equation (bonds)

$$(C_t)^{-\gamma} = \beta(1+r_{t+1})E_t \left[ (C_{t+1})^{-\gamma} \right], \quad (1)$$

$$(C_t^*)^{-\gamma} = \beta(1+r_{t+1}^*)E_t \left[ (C_{t+1}^*)^{-\gamma} \right], \quad (2)$$

Euler equation (shares)

$$\tilde{v}_t = \beta(1-\delta)E_t \left[ \left( \frac{C_{t+1}}{C_t} \right)^{-\gamma} (\tilde{v}_{t+1} + \tilde{d}_{t+1}) \right], \quad (3)$$

$$\tilde{v}_t^* = \beta(1-\delta)E_t \left[ \left( \frac{C_{t+1}^*}{C_t^*} \right)^{-\gamma} (\tilde{v}_{t+1}^* + \tilde{d}_{t+1}^*) \right], \quad (4)$$

### 2 Firm: profit maximization

Profits

$$\tilde{d}_t = \tilde{d}_{D,t} + \frac{N_{X,t}}{N_{D,t}} \tilde{d}_{X,t}, \quad (5)$$

$$\tilde{d}_t^* = \tilde{d}_{D,t}^* + \frac{N_{X,t}^*}{N_{D,t}^*} \tilde{d}_{X,t}^*, \quad (6)$$

Free entry (Policy)

$$\tilde{v}_t = w_t \frac{f_{E,t}}{Z_t} \quad (7)$$

$$\tilde{v}_t^* = w_t^* \frac{f_{E,t}^*}{Z_t^*} \quad (8)$$

Zero-profit export cutoffs

$$\tilde{d}_{X,t} = w_t \frac{f_{X,t}}{Z_t} \frac{\theta-1}{k-(\theta-1)}, \quad (9)$$

$$\tilde{d}_{X,t}^* = w_t^* \frac{f_{X,t}^*}{Z_t^*} \frac{\theta-1}{k-(\theta-1)}, \quad (10)$$

**Table 8.1 The list of variables for the model**

<i>Variables</i>	<i>Definitions</i>
$t$	Period
$C$	Consumption
$B$	Bond holdings
$r$	Consumption-based interest rate on bonds
$x$	Mutual fund share holdings
$L$	Labor
$w$	Real wage
$Z$	Aggregate labor productivity
$\bar{z}_x$	The average productivity level for all home exporters
$z_{\min}$	The lower bound of relative productivity $z$
$N_H$	The number of total home firms
$N_D$	The number of home-producing firms
$N_X$	The number of home exporters
$N_E$	The number of new entrants
$Q$	Consumption-based real exchange rate
$\tilde{v}$	Expected post entry value
$\tilde{d}$	The average firm profit levels
$\tilde{d}_D$	The average firm profit earned from domestic sales for all home producers
$\tilde{d}_X$	The average firm profit earned from export sales for all home exporters
$\tilde{p}_D$	The average relative price of producers
$\tilde{p}_X$	The average relative price of exporters
$f_E$	Entry cost
$f_X$	Fixed cost
$k$	Productivity distribution shape parameter
$\beta$	Subjective discount factor
$\gamma$	The inverse of the intertemporal elasticity of substitution
$\theta$	The symmetric elasticity of substitution across goods
$\delta$	Probability of exit-inducing shock

\*Note: These are home variables and corresponding foreign variables are denoted by an asterisk.

Source: Ghironi and Melitz (2005).

Share of exporting firms

$$\frac{N_{X,t}}{N_{D,t}} = (Z_{\min})^k (\bar{z}_{X,t})^{-k} \left[ \frac{k}{k - (\theta - 1)} \right]^{k/(\theta-1)}, \quad (11)$$

$$\frac{N_{X,t}^*}{N_{D,t}^*} = (Z_{\min}^*)^k (\bar{z}_{X,t}^*)^{-k} \left[ \frac{k}{k - (\theta - 1)} \right]^{k/(\theta-1)}, \quad (12)$$

Number of firms

$$N_{D,t} = (1 - \delta)(N_{D,t-1} + N_{E,t-1}), \quad (13)$$

$$N_{D,t}^* = (1 - \delta)(N_{D,t-1}^* + N_{E,t-1}^*), \quad (14)$$

3 Equilibrium conditions:

Price indices

$$N_{D,t} (\tilde{p}_{D,t})^{1-\theta} + N_{X,t}^* (\tilde{p}_{X,t}^*)^{1-\theta} = 1, \quad (15)$$

$$N_{D,t}^* (\tilde{p}_{D,t}^*)^{1-\theta} + N_{X,t} (\tilde{p}_{X,t})^{1-\theta} = 1, \quad (16)$$

Aggregate accounting

$$C_t = w_t L + N_{D,t} \tilde{d}_t - N_{E,t} \tilde{v}_t, \quad (17)$$

$$C_t^* = w_t^* L^* + N_{d,t}^* \tilde{d}_t^* - N_{e,t}^* \tilde{v}_t^*, \quad (18)$$

Balanced trade

$$Q_t N_{x,t} (\tilde{p}_{x,t})^{1-\theta} C_t^* = N_{x,t}^* (\tilde{p}_{x,t}^*)^{1-\theta} C_t, \quad (19)$$

**Comparing Model Predictions with Actual Data (Calibration):** Ghironi and Melitz test the model via calibration and compute the moments from the simulated data and compare them with the actual moments in the United States.<sup>26</sup> They reproduce key economic features of the United States under exogenous shocks to aggregate productivity and contrast it with the actual data for the period 1954 to 1989 (Backus et al., 1992). The parameter values in their calibrations are presented in Table 8.2. Here we focus on how closely the model mimics actual data for various economic variables: Net Exports/Output, Investment, Consumption, and Aggregate Output.

Figure 8.1 illustrates one comparison between the simulated and actual volatility (e.g., standard deviation). Ghironi and Melitz's model comparisons reveal that their model underpredicts the volatilities of aggregate output, consumption and investment, while it generates a more volatile ratio between net exports and output.

**Policy Effects: Simulation:** Next, we examine the policy regulation effects on entry ( $f_E$ ). Specifically, Ghironi and Melitz examine the trade policy parameters (entry or trade costs) in (7). The expectation is that the policy shocks will affect the real exchange rate dynamics, explaining the HBS effect through endogenous channels.

Figure 8.2 illustrates the impact of a permanent shock to deregulation ( $f_E$ ) – reducing entry costs – on the microeconomic response and, then, the macroeconomic outcomes. The vertical axis indicates the percent deviation from the steady state (response to a permanent deregulation shock) and the horizontal axis shows the number of years after the shock.

We present the results for total labor cost and the real exchange rate. The process in this example is as follows: the deregulation shock (a decline in the deregulation parameter ( $f_E$ )) in the home market raises domestic demand. Subsequently, domestic wages increase, decreasing the relative effective labor cost ( $TOL$  in Figure 8.2) in the long run. The result is that the real exchange rate ( $\tilde{Q}$  in Figure 8.2) appreciates (the cost of a foreign basket of goods falls) in the long term, as it would in the HBS model.

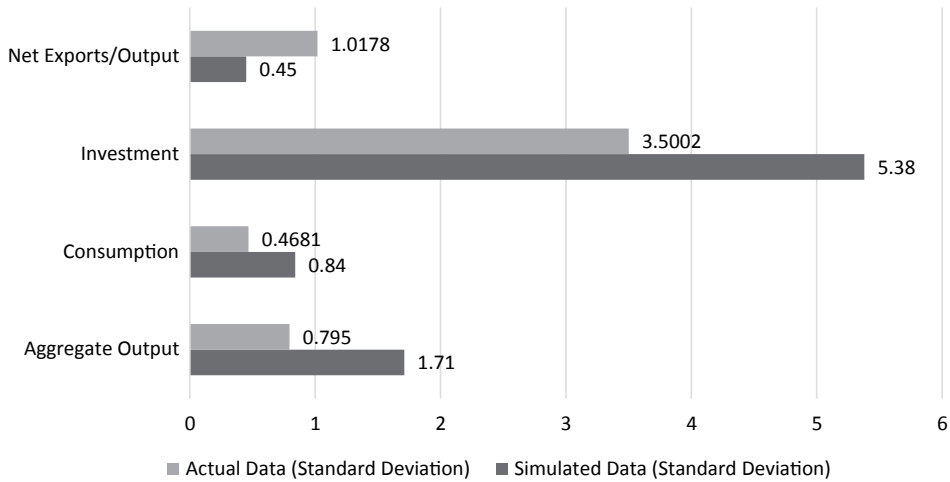
We also find that policy deregulation's influence on the decline in total labor cost and the appreciation of the real exchange

**Table 8.2 Parameter values**

Parameters	Values	Definitions
<b>Policy<sup>1</sup></b>		
$f_E$	1	Entry cost: equation (7), firm profit maximization
<b>Economy</b>		
$\beta$	0.99	Subjective discount factor: equations (1) and (3), household utility maximization
$\gamma$	2	The inverse of the intertemporal elasticity of substitution: equations (1) and (3), household utility maximization
$\delta$	0.025	Probability of exit-inducing shock: equation (3), household utility maximization
$\theta$	3.8	The symmetric elasticity of substitution across goods: equations (9) and (11), firm profit maximization
$k$	3.4	Productivity distribution shape parameter: equations (9) and (11), firm profit maximization

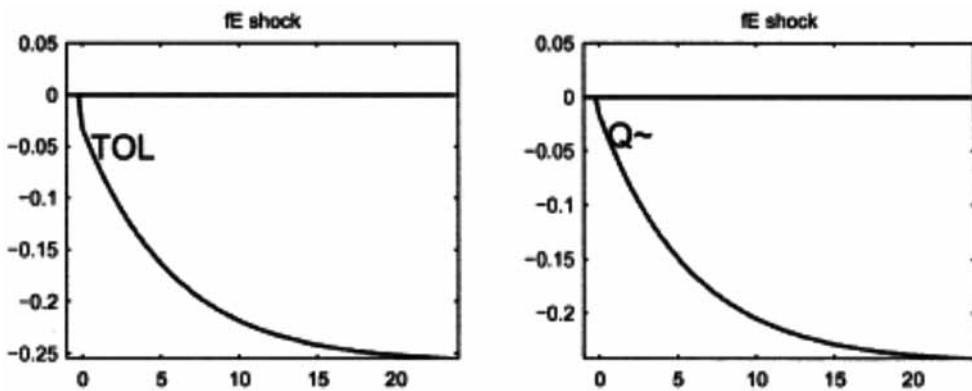
<sup>1</sup> Another policy parameter — fixed cost  $f_x$  — is calculated based on the entry cost  $f_E$ . It equals 23.5 percent of the per-period, amortized flow value of the entry cost, which is roughly 0.008 as the entry cost  $f_E$  equals 1.

Source: Ghironi and Melitz (2005).



**Figure 8.1 Simulated and actual volatility**

Source: Ghironi and Melitz (2005).



**Figure 8.2 Deregulation effects**

Source: Ghironi and Melitz (2005).

rate is not immediate. Within five years of the shock (a 1% decline in the entry cost), less than 20% of the deviation from the steady state emerges regarding the total labor cost. For the real exchange rate, there is about a 15% deviation, but in the correct direction.

**Incorporating Political Factors:** What role for politics in RBC? One avenue has been provided by Freeman and Houser (1998). They built on the work of Chappell

and Keech (1983) and took up the challenge presented by the Cowles tradition. Freeman and Houser set up an RBC model to study the joint equilibrium in the political environment and the macro-economy. As with Chappell and Keech, Freeman and Houser argue that economic decisions and political decisions are often studied separately in the literature. They state that there is ‘a lack of theoretical balance between economic and political theory’ (Freeman and Houser, 1998: 628).

Freeman and Houser's computable general equilibrium model combines the Ramsey-type RBC model with the policy maker's objective function on approval ratings. A summary of their model – without formalizing it here – is as follows: the government is a representative agent who optimally chooses tax rates and gross interest rate on public bonds in order to minimize the present value of expected sum of deviations of approval ratings. Hence, the government chooses optimal policy to minimize the approval target deviations given the optimal decisions made by households and the firms. The more general point of their model is that it opens the way for many rich political models – partisan, institutional and more – and how these models influence fiscal, monetary, trade and regulatory policy within the RBC methodology and EITM framework.

## SUMMARY

EITM was established through the NSF initiative in 2001. As the example in the last section indicates, the motivation of EITM-like research predates the 21st century considerably. Its realization in political science was due in large part to the great successes that both the development of formal modeling and political methodology had achieved in the discipline. It was the concern that they were developing on separate tracks that led to EITM as a way to align scientific development in political science, in the social and behavioral sciences and even in science more generally.

We argued here that EITM is best understood as a way to generate and communicate new scientific knowledge. It is, in that sense, a research design problem rather than a precise set of constraints on research. And, we argued, we see it as a research design applied to a stream of research.

We described two examples of EITM-consistent research. We take two basic

conclusions from these examples. First, EITM can be applied both to theory and to empirics. Indeed, the best of EITM research uses the design to solve both sides of the scientific enterprise. Second, both examples unfolded and became richer and broader in their applications over time. That is, pursuing EITM-consistent lines of inquiry in these cases – and, we believe, more generally – facilitates the accumulation of knowledge. Perhaps this is indeed the greatest promise of EITM.

## APPENDIX

**The Model:** The formal model consists of two countries, home and foreign, where households and a continuum of firms exist in each country. In the home country, the representative household supplies  $L$  units of labor inelastically in each period  $t$  at the nominal wage rate  $W_t$  (in home currency). Households make optimal decisions on consumption ( $C$ ) to maximize their utility. Firms are heterogeneous with different relative productivity,  $z$ , that translates into differences in production cost. The same holds true for foreign representative households and firms.<sup>27</sup>

The dynamics of the model *and for policy* focus on entry. Prior to entry, home (foreign) firms are identical; facing a sunk entry cost of  $f_{E,t}$  ( $f_{E,t}^*$ ) effective labor units, equal to the home (foreign) consumption goods. Upon entry, home firms draw their productivity level  $z$  from a common distribution  $G(z)$ , bounded by  $[z_{\min}, \infty)$ , while foreign homes draw theirs from an identical distribution.

All firms can serve both domestic and export markets. The cost of exporting includes a trade cost  $\tau_t \geq 1$  ( $\tau_t^* \geq 1$ ) and a fixed cost  $f_{X,t}$  ( $f_{X,t}^*$ ). Home (foreign) firms decide whether to export depending on the profits, which is determined by the firm's productivity level. A firm will export if and only if a nonnegative profit is earned from exporting,



which they assume will be the case provided  $z$  is above a cutoff level  $z_{X,t}(z_{X,t}^*)$ . They also assume the lower bound of  $z$ ,  $z_{\min}$ , is small enough relative to the export costs and that  $z_{X,t}(z_{X,t}^*)$  is above  $z_{\min}$ . This assumption determines that a nontraded sector exists. These firms can change their decision over time if profitability changes in the export market.

Firms interact with households in the following way.<sup>28</sup> On the demand side, the model assumes households in each country maximize expected intertemporal utility from consumption:

$$E_t \left[ \sum_{s=t}^{\infty} \beta^{s-t} C_s^{1-\gamma} / (1-\gamma) \right], \quad (1a)$$

where  $\beta \in (0, 1)$  is the subjective discount factor and  $\gamma > 0$  is the inverse of the intertemporal elasticity of substitution. In period  $t$ , the household consumes the baskets of goods:  $C_t = \left( \int_{\omega \in \Omega} c_t(\omega)^{\theta-1} d\omega \right)^{\theta/(\theta-1)}$ , where

$\Omega$  is a continuum of goods and  $\theta > 1$  is the symmetric elasticity of substitution across goods. The model also assumes households possess two types of assets, including risk-free bonds ( $B$ ) and shares ( $x$ ) in a mutual fund of domestic firms. Taking the home country as an example (at time  $t$ ) the representative household buys  $x_{t+1}$  shares in a mutual fund of  $N_{H,t} \equiv N_{D,t} + N_{E,t}$  home firms (where  $N_{D,t}$  is the number of home-producing firms,  $N_{E,t}$  is the number of new entrants).

Among the firms producing in the home country, there are  $N_{X,t}$  exporters. The household receives gross interest income (from bond holdings), dividend income (from mutual fund share holdings), as well as the value of selling its initial share position and labor income. They allocate this income stream to the purchase of bonds and shares to be held into next period and consumption. Thus, the period budget constraint for the household is:

$$\begin{aligned} B_{t+1} + \tilde{v}_t N_{H,t} x_{t+1} + C_t \\ = (1+r_t)B_t + (\tilde{d}_t + \tilde{v}_t) N_{D,t} x_t + w_t L, \end{aligned} \quad (2a)$$

where  $\tilde{v}_t$  is the expected post entry value given by the present discounted value of prospective home entrants' expected stream of profits,  $r_t$  is the consumption-based interest rate on bonds holdings from  $t-1$  to  $t$ ,  $\tilde{d}_t$  represents average firm profit levels, and  $w_t \equiv W_t/P_t$  is the real wage ( $P_t$  is the consumption-based price index).

For the supply side, aggregate labor productivity is indicated as  $Z_t(Z_t^*)$ . Firms are heterogeneous with different technologies measured in relative productivity  $z$ . A home firm, for instance, produces  $Z_t z$  units of output per unit of labor employed – the production function. It is assumed that  $z$  follows a Pareto distribution with lower bound  $z_{\min}$  and shape parameter  $k > \theta - 1$ :  $G(z) = 1 - (z_{\min}/z)^k$ . All firms produce – in any given time  $t$  – until an exit-inducing shock occurs with the probability  $\delta \in (0, 1)$  that is independent of  $z$ . All firms also face a residual demand curve with constant elasticity  $\theta$  in both domestic and export markets. They set flexible prices reflecting the same proportional markup  $\theta/(\theta - 1)$  over marginal cost.

The nominal domestic and export prices are indicated by  $p_{D,t}(z)$  and  $p_{X,t}(z)$  for a home firm; the export prices are denoted in the currency of the export market. Therefore, the relative domestic and export prices are  $\rho_{D,t}(z) \equiv \frac{p_{D,t}(z)}{P_t} = \frac{\theta}{\theta-1} \frac{w_t}{Z_t z}$  and  $\rho_{X,t}(z) \equiv \frac{p_{X,t}(z)}{P_t^*} = Q_t^{-1} \tau_t \rho_{D,t}(z)$ , respectively,

where  $Q_t$  represents the real exchange rate. As a consequence of the fixed cost of exporting, less productive firms may decide not to export at time  $t$ . To make the decision, a firm considers the total profit  $d_t(z)(d_t^*(z))$  in terms of the profit generated from domestic sales  $d_{D,t}(z)(d_{D,t}^*(z))$  and that earned from potential export sales  $d_{X,t}(z)(d_{X,t}^*(z))$ . All these profits are denominated in real terms in units of the consumption basket in the firm's country.

Prices are also assumed to be affected by productivity levels. Following Melitz (2003),

Ghironi and Melitz define two ‘special “average” productivity levels’ (Ghironi and Melitz, 2005: 874),  $z_D$  and  $\tilde{z}_{X,t}(\tilde{z}_{X,t}^*)$ , for all producing firms  $N_{D,t}(N_{D,t}^*)$  and exporters  $N_{X,t}(N_{X,t}^*)$ , respectively. Accordingly, in the home market, the average relative price of producers is indexed by  $\tilde{\rho}_{D,t} \equiv \rho_{D,t}(\tilde{z}_D)$ . The average relative price of foreign exporters is  $\tilde{\rho}_{X,t} \equiv \rho_{X,t}(\tilde{z}_{X,t}^*)$ .  $\tilde{d}_{D,t} \equiv d_{D,t}(\tilde{z}_D)(\tilde{d}_{D,t}^* \equiv d_{D,t}^*(\tilde{z}_D^*))$  denotes the average firm profit earned from domestic sales for all home (foreign) producers.  $\tilde{d}_{X,t} \equiv d_{X,t}(\tilde{z}_{X,t})(\tilde{d}_{X,t}^* \equiv d_{X,t}^*(\tilde{z}_{X,t}^*))$  represents the average firm export profits for all home (foreign) exporters.

Consequently, the average total profits are  $\tilde{d}_t \equiv \tilde{d}_{D,t} + [1 - G(\tilde{z}_{X,t})]\tilde{d}_{X,t}$  for home firms and  $\tilde{d}_t^* \equiv \tilde{d}_{D,t}^* + [1 - G(\tilde{z}_{X,t}^*)]\tilde{d}_{X,t}^*$  for foreign firms. At any given period,  $t$ , there is an unlimited amount of prospective entrants  $N_{E,t}$  in both home and foreign countries. These firms are forward looking and anticipate their future expected profits  $\tilde{d}_t(\tilde{d}_t^*)$  and the probability of the occurrence of the exit-inducing shock  $\delta$ . Firms discount future profits based on the household’s stochastic discount factor  $\beta$ , adjusted for the probability of firm survival  $1 - \delta$ . The entry occurs when the average postentry value (given by the current discounted value of the expected profits) equals the entry cost, resulting in the free entry condition  $\tilde{v}_t = w_t f_{E,t} / Z_t$ . Note, this condition holds so long as  $N_{E,t} > 0$ . Moreover, for this condition to hold in every period, macroeconomic shocks are assumed to be sufficiently small.

## Notes

- 1 We thank Cong Huang and M. C. Sunny Wong for their assistance.
- 2 This phrase can be meant in the way used by Kuhn, 1970 [1962], but here we mean ‘before doing the science’ in a more general sense.
- 3 The last is a paraphrase of a half of a lecture Prof. Arthur Goldberg gave to one of us in the University of Rochester’s PS 401, scope and methods class, in its Political Science PhD pro-

gram, or at least it is paraphrased as well as one of this paper’s co-authors can remember hearing it.

- 4 No one piece of research, say an article in a leading peer-reviewed journal, needs to accomplish all aspects of an EITM contribution (indeed, perhaps it *should* not, or at least should not be expected to do so). However, attention to where it fits in an EITM-consistent stream of research, what it contributes in those terms and what remains to be accomplished in that stream should be a part of any such paper.
- 5 The 2001 NSF EITM Workshop was recorded and transcribed. The written transcript is available on the NSF Political Science Program Web Site: [www.nsf.gov/sbe/ses/polisci/reports/eitm709.pdf](http://www.nsf.gov/sbe/ses/polisci/reports/eitm709.pdf) and [www.nsf.gov/sbe/ses/polisci/reports/eitm710.pdf](http://www.nsf.gov/sbe/ses/polisci/reports/eitm710.pdf). A report of the EITM initiative – based in part on the 2001 EITM Workshop (EITM Report 2002) – is also available at: [www.nsf.gov/sbe/ses/polisci/reports/pdf/eitmreport.pdf](http://www.nsf.gov/sbe/ses/polisci/reports/pdf/eitmreport.pdf).
- 6 The participants in the workshop – with diverse methodological backgrounds – were senior scholars with research experience in various technical-analytical areas and proven track records in activities that have improved technical-analytical expertise in various social sciences. Participants were primarily from political science, but economics and mathematics were represented as well. For background on the motivation for the EITM Workshop see Granato et al. (2015).
- 7 The EITM initiative is part of a multi-method approach. Recall that the motivation for the use of EITM has quantitative roots, but consistent with the arguments of Poteete et al. (2010) it is recognized that qualitative approaches have various strengths, including highlighting the importance of context (Granato and Scioli, 2004: 314–15). But, as with quantitative tools, Granato and Scioli (2004) do highlight shortcomings of qualitative approaches (e.g., Sartori, 1991; Goldthorpe, 1997).
- 8 See Wagner (2001: 3).
- 9 Recall from the introduction, that EITM is a method – even a mindset – where researchers treat formal and empirical analysis as linked entities intended to create a dialogue between theory and test. But how to define EITM? What does it look like when we try to implement the dialogue? Implementation involves defining the elements of EITM – a framework – and showing how one does EITM research and how one trains students to do such research. The development of a framework was deemed important at NSF since ‘without a framework to organize relevant variables identified in theories and empirical research, isolated knowledge acquired from studies ... by ... social

- and behavioral scientists is not likely to cumulate' (Ostrom, 2009: 420). See Granato (2005) and Granato et al. (2010) for a description and examples of the EITM framework.
- 10 A key outcome of the EITM initiative has been the EITM Summer Institutes. The Summer Institutes have taken place at:
    - Harvard University (2002)
    - The University of Michigan (2003, 2006, 2009, 2015, 2018)
    - Washington University, St. Louis (2003–9)
    - Duke University (2004, 2008, 2014, 2016)
    - UC-Berkeley (2005, 2010, 2013, 2017)
    - UCLA (2007)
    - University of Chicago (2011)
    - Princeton University (2012)
    - University of Houston (2012–17, 2019)
    - Emory University (2019).
  - 11 The finitely repeated prisoner's dilemma is the other major example.
  - 12 In the centipede game, the (usually two) players move sequentially. The first player can take an initial sum of money or pass. If the player takes the money, the game ends. If, instead, the player passes, the initial amount increases (say, doubles), and the second player can either take this larger amount of money or pass. If 'take' is played, the game ends. If pass, the money is increased again (say, doubled a second time) and the choice goes back to the first player. The game continues with the play of 'pass' and ends as soon as a player decides to take the money.
  - 13 Of course, the finitely repeated prisoner's dilemma also has a unique (and easily implemented) Nash equilibrium (every player always defects). Like the centipede game, perhaps even more so, actual players rarely follow that strategy under a wide range of conditions, even when they understand that always defect is the unique Nash equilibrium.
  - 14 While there are many good reasons for choosing a (formal) logic or mathematical representation of TM, those are all conveniences made possible through the use of math or logic. There is nothing in EITM that requires formalization of the theory. The requirement is logical consistency.
  - 15 We mean 'game-experimental' in the sense as used in Kinder and Palfrey (1993).
  - 16 Carson published a paper on candidate decisions to run for office, developing a game theoretic model of ambition theory and estimating it via the same method as Leblang in the same year (2003). Thus, the QRE and resulting STRAT-like estimation models have been applied in IPE and American politics shortly after Signorino developed the original.
  - 17 More recently, Palfrey and co-authors have closed the circle, as it were, in developing QRE as what they call a 'statistical theory of games' (Goeree et al., 2016).
  - 18 Econometric research associated with the Cowles Commission includes (but is not limited to): Cooper (1948), Haavelmo (1943, 1944), Hood and Koopmans (1953), Klein (1947), Koopmans (1945, 1949, 1950), Koopmans and Reiersol (1950), Marschak (1947, 1953) and Vining (1949). For further background on the Cowles Commission consult the following URL: <http://cowles.econ.yale.edu/>.
  - 19 The intuition behind the terms *identify* (i.e., identification) and *invariant* (i.e., invariance) is as follows. For applied statistical models *identification* relates to model parameters (e.g.,  $\beta$ s) and whether they indicate the magnitude of the effect for that particular independent variable. Or, in more technical terms, 'A parameter is identifiable if different values for the parameter produce different distributions for some observable aspect of the data' (Brady and Collier, 2004: 290). In applied statistical practice, *invariance* refers to the constancy of the parameters of interest. More generally, 'the distinctive features of causal models is that each variable is determined by a set of other variables through a relationship (called "mechanism") that remains invariant (constant) when those other variables are subjected to external influences. Only by virtue of its invariance do causal models allow us to predict the effect of changes and interventions' (Pearl, 2000: 63).
  - 20 Underlying this approach was the argument that any change in model specification would be traced out through the entire system with new results known *ex-ante* – a fundamental feature of EITM too. Indeed, this is one improvement over single equation models where specification changes would lead to new results but with the *ex-ante* origins of the effects known.
  - 21 Gabaix and Laibson (2008: 295) argue that falsifiability and predictive precision are among the key properties of useful models: 'A model is falsifiable if and only if the model makes nontrivial predictions that can in principle be empirically falsified.' Furthermore:
 

Models have *predictive precision* when they make precise – or 'strong' – predictions. Strong predictions are desirable because they facilitate model evaluation and model testing. When an incorrect model makes strong predictions, it is easy to empirically falsify the model, even when the researcher has access only to a small amount of data. A model with predictive precision also has greater potential

to be practically useful if it survives empirical testing. Models with predictive precision are useful tools for decision makers who are trying to forecast future events or the consequences of new policies. (p. 295)

In the language of econometrics, falsification and predictive precision require the mechanisms relating cause and effect to be identified. There is a large literature devoted to identification problems (see, e.g., Koopmans, 1949; Fisher, 1966; Manski, 1995), but we use identification in the broadest sense for purposes of attaining some order and underlying cause as well. Since we as social scientists do not have controlled environments to conduct our inquiry, our efforts to achieve order and cause in our models can only come about probabilistically – by chance.

- 22 The Lucas *critique* is based on the following intuition: 'given that the structure of an econometric model consists of optimal decision rules ... and that optimal decision rules vary systematically with changes in the structure of series relevant to the decision maker, it follows that any change in policy will systematically alter the structure of econometric models' (Lucas, 1976: 41).
- 23 Here the focus is on isolating parameters and on making greater explicit use of theory at both the individual and aggregate level of analysis. Where RBCs especially differ from the Cowles Commission is in the use of standard statistical significance testing.
- 24 The DSGE evolution is due in part to the inclusion of policy into the RBC modeling process. Indeed, Prescott precisely foresaw this possibility:

The models constructed within this theoretical framework are necessarily highly abstract. Consequently, they are necessarily false, and statistical hypothesis testing will reject them. This does not imply, however, that nothing can be learned from such quantitative theoretical exercises. I think much has already been learned and confidently predict that much more will be learned as other features of the environment are introduced. Prime candidates for study are the effects of public finance elements, a foreign sector, and, of course, monetary factors. The research I review here is best viewed as a very promising beginning of a much larger research program. (Prescott, 1986: 10)

For our purposes we will continue to use the terms RBC and DSGE when one or the other is more appropriate.

- 25 See Melitz (2003).
- 26 To reiterate, these moments are based on the findings in Backus et al. (1992). They introduce

the resource cost of trade in their international RBC model and include the shocks to aggregate productivities to replicate second moments of the United States and international data.

- 27 All foreign variables are denoted by an asterisk in Ghironi and Melitz's work.
- 28 Trade policy outcomes are based on aggregating the budget constraint across home households and imposing the equilibrium conditions under financial autarky. This also holds in a foreign country.

## REFERENCES

- Aldrich, John H. *Before the Convention: Strategies and Choices in Presidential Nomination Campaigns*. Chicago: University of Chicago Press, 1980.
- Backus, David K., Patrick J. Kehoe and Finn E. Kydland. 'International real business cycles.' *Journal of Political Economy* 100, no. 4 (1992): 745–775.
- Brady, Henry E., and David Collier, eds. *Rethinking Social Inquiry: Diverse Tools, Shared Standards*. Lanham, MD: Rowman and Littlefield, 2004.
- Carson, Jamie L. 'Strategic interaction and candidate competition in US house elections: empirical applications of probit and strategic probit models.' *Political Analysis* 11, no. 4 (2003): 368–380.
- Chappell, Henry W., and William R. Keech. 1983. 'Welfare Consequences of the Six-Year Presidential Term Evaluated in the Context of a Model of the U.S. Economy.' *American Political Science Review* 77, no.1: 75–91.
- Cooper, Gershon. 'The role of econometric models in economic research.' *Journal of Farm Economics* 30, no. 1 (1948): 101–116.
- Fisher, Franklin M. *The Identification Problem in Econometrics*. New York: McGraw-Hill, 1966.
- Freeman, John R., and Daniel Houser. 'A computable equilibrium model for the study of political economy.' *American Journal of Political Science* 42, no. 2 (1998): 628–660.
- Gabaix, Xavier, and David Laibson. 'The seven properties of good models.' In *The Foundations of Positive and Normative Economics: A Handbook*, pp. 292–299, eds Andrew Caplin and Andrew Schotter. New York: Oxford University Press, 2008.

- Ghironi, Fabio, and Marc J. Melitz. 'International trade and macroeconomic dynamics with heterogenous firms.' *The Quarterly Journal of Economics* 120, no. 3 (2005): 865–915.
- Goeree, Jacob K., Charles A. Holt and Thomas R. Palfrey. 'Quantal response equilibria.' In *Behavioural and Experimental Economics*, pp. 234–242, eds Steven N. Durlauf and Lawrence E. Blume. London: Palgrave Macmillan, 2016.
- Goldthorpe, John H. 'Current issues in comparative macrosociology: a debate on methodological issues.' *Comparative Social Research*, 16 (1997): 1–26.
- Granato, Jim. 'Scientific progress in quantitative political economy.' *The Political Economist* 12, no. 4 (2005): 1, 3, 9, 11–13, 20.
- Granato, Jim, Cong Huang, Kwok W. Wan, Ching-Hsing Wang and M. C. Sunny Wong. 'EITM: an assessment with an application to economic voting.' *Electoral Studies*, 40 (2015): 372–393.
- Granato, Jim, Melody Lo and M. C. Sunny Wong. 'A framework for unifying formal and empirical analysis.' *American Journal of Political Science* 54, no. 3 (2010): 783–797.
- Granato, Jim, Melody Lo and M. C. Sunny Wong. *The Empirical Implications of Theoretical Models: Unifying Formal and Empirical Analysis in the Political, Social, and Economic Sciences*, n.d.
- Granato, Jim, and Frank Scioli. 'Puzzles, provverbs, and omega matrices: the scientific and social consequences of the empirical implications of theoretical models (EITM).' *Perspectives on Politics* 2, no. 2 (2004): 313–323.
- Haavelmo, Trygve. 'The statistical implications of a system of simultaneous equations.' *Econometrica* 11, no. 1 (1943): 1–12.
- Haavelmo, Trygve. 'The probability approach in econometrics.' *Supplement to Econometrica*, 12 (1944): S1–115.
- Hood, W. M. C., and Tjalling C. Koopmans, eds. *Studies in Econometric Method, Cowles Commission Monograph No. 14*. New York: John Wiley and Sons, 1953.
- Kinder, Donald R., and Thomas R. Palfrey, eds. *Experimental Foundations of Political Science*. Ann Arbor, MI: University of Michigan Press, 1993.
- Klein, Lawrence R. 'The use of econometric models as a guide to economic policy.' *Econometrica* 15, no. 2 (1947): 111–151.
- Koopmans, Tjalling C. 'Statistical estimation of simultaneous economic relations.' *Journal of the American Statistical Association* 40, no. 232 (1945): 448–466.
- Koopmans, Tjalling C. 'Identification problems in economic model construction.' *Econometrica* 17, no. 2 (1949): 125–144.
- Koopmans, Tjalling C., ed. *Statistical Inference in Dynamic Economic Models. Cowles Commission Monograph No. 10*. New York: John Wiley and Sons, 1950.
- Koopmans, Tjalling C., and O. Reiersol. 'The identification of structural characteristics.' *The Annals of Mathematical Statistics* 21, no. 2 (1950): 165–181.
- Kravis, Irving and Robert E. Lipsey. *Toward an Explanation of National Price Levels*. Princeton, NJ: Princeton University International Finance Section, 1983.
- Kuhn, Thomas S. *The Structure of Scientific Revolutions*, 2nd enlarged ed. Chicago, IL: University of Chicago Press, 1970 [1962].
- Kuhn, Thomas S. *The Essential Tension: Selected Studies in Scientific Tradition and Change*. Chicago, IL: University of Chicago Press, 1979.
- Kydland, Finn E., and Edward C. Prescott. 'Time to build and aggregate fluctuations.' *Econometrica* 50, no. 6 (1982): 1345–1370.
- Leblang, David. 'To devalue or to defend: the political economy of exchange rate policy.' *International Studies Quarterly* 47, no. 4 (2003): 533–559.
- Lucas, Robert E., Jr. 'Econometric policy evaluation: a critique.' *Carnegie-Rochester Conference on Public Policy*, 1 (1976): 19–46.
- Lucas, Robert E., Jr. 'Tobin and monetarism: a review article.' *Journal of Economic Literature* 19, no. 2 (1981): 558–567.
- Manski, Charles F. *Identification Problems in the Social Sciences*. Cambridge: Harvard University Press, 1995.
- Marschak, Jacob. 'Economic structure, path, policy, and prediction.' *American Economic Review* 37, no. 2 (1947): 81–84.
- Marschak, Jacob. 'Economic measurements for policy and prediction.' In *Studies in Econometric Method*, eds W. M. C. Hood and Tjalling C. Koopmans. New York: John Wiley and Sons, 1953.
- McKelvey, Richard D., and Thomas R. Palfrey. 'Quantal response equilibria for normal form

- games.' *Games and Economic Behavior* 10, no. 1 (1995): 6–38.
- McKelvey, Richard D., and Thomas R. Palfrey. 'Quantal response equilibria for extensive form games.' *Experimental Economics* 1, no. 1 (1998): 9–41.
- Melitz, Marc J. 'The impact of trade on intra-industry reallocations and aggregate industry productivity.' *Econometrica* 71, no. 6 (2003): 1695–1725.
- Ostrom, Elinor. 'A general framework for analyzing sustainability of social-ecological systems.' *Science* 325, 5939 (2009): 419–422.
- Pearl, Judea. *Causality*. New York: Cambridge University Press, 2000.
- Poteete, Amy R., Marco A. Janssen and Elinor Ostrom. *Working Together: Collective Action, The Commons, and Multiple Methods in Practice*. Princeton, NJ: Princeton University Press, 2010.
- Prescott, Edward C. 'Theory ahead of business cycle measurement.' *Federal Reserve Bank of Minneapolis Quarterly Review* 10, no. 4 Fall, (1986): 9–22.
- Sartori, Giovanni. 'Comparing and miscomparing.' *Journal of Theoretical Politics* 3, no. 3 (1991): 243–257.
- Signorino, Curtis S. 'Strategic interaction and the statistical analysis of international conflict.' *American Political Science Review* 93, no. 2 (1999): 279–297.
- Vining, Rutledge. 'Methodological issues in quantitative economics: Koopmans on the choice of variables to be studied and of methods of measurement.' *The Review of Economics and Statistics* 31, no. 2 (1949): 77–86.
- Wagner, R. Harrison. 'Who's afraid of "rational choice theory"?' Typescript. University of Texas at Austin, 2001.

*This page intentionally left blank*

PART II

# Methods of Theoretical Argumentation





*This page intentionally left blank*



# Political Psychology, Social Psychology and Behavioral Economics

Rose McDermott

This chapter will examine the differences, as well as the overlaps, between the procedures, assumptions, criteria for evaluating quality and accepted standards for empirical research across several different disciplines. Particular attention will be paid to how these characteristics vary according to the different questions and aims pursued by each discipline. In addition, underlying differences in disciplinary culture that affect these intellectual pursuits will also be mentioned where relevant.

Although clearly this endeavor could be undertaken around a wide variety of disciplines, here the focus will be on political psychology, social psychology and behavioral economics, primarily because these disciplines share some important overlaps regarding the operative unit of analysis lying with the individual actor. These areas share some overlap with other disciplines that may intersect with political science, such as sociology and anthropology, but those areas tend to focus more on group dynamics and thus are more distinct in their assumptions and approaches. This does not make them better

or worse for analyzing any given political problem. Rather, as with any research question, the particular approach and methods should be driven by what is best suited for properly addressing the central research question at hand, and not by any special fixation on a particular theory or method.

This chapter will proceed by discussing each of the areas mentioned above for each of these disciplines. Although space and scope limitations prevent any substantive analysis of the content of these areas, interested readers will be referred to critical texts in each area that can provide such background. The discussion here is necessarily limited to an analysis of the similarities and differences across fields in each of these areas.

## **ACCEPTED PROCEDURES**

'Accepted procedures' really refers to those particular methodological standards that are considered standard for empirical

examination and analysis within a discipline. Accepted procedures across disciplines typically differ primarily around methodological issues and not theoretical approaches. In particular, dominant methods in each discipline largely inform the criteria by which quality work is evaluated, and which statistical as well as procedural standards are considered acceptable in each area as well.

The gold standard for all of psychology, whether political or social, revolves around the carefully controlled laboratory experiment, where subjects are submitted to conditions that vary between controlled and treatment manipulations. Investigators strive to keep as much of the background environment similar between all conditions and only change the variable of interest. This careful control and manipulation allows observers to determine the extent to which any differences that might emerge in subject responses are attributable to the variable that was manipulated across conditions. These kind of experiments can measure various attitudinal or behavioral responses, but the method of administration remains fairly standard. There are several good reference books on how to conduct experiments appropriately within the accepted standards of social psychology; Aronson et al. (1990) provide a superlative overview with useful examples for how to implement such procedures.

Political psychology encompasses a much broader array of methods drawn in part from other disciplines, including political science, such as qualitative case studies, archival work, surveys, interviews and narratives (McDermott, 2004). This is partly because political psychology is often interested not only in living people but also in historical figures, whose life and influence may only be accessed through archives, oral histories or interviews with people who knew them. In addition, because the focus tends to be more exclusively on political as opposed to broader social phenomena, as is the case with social psychology, political psychology will often examine levels of analysis that go beyond

the individual to examine group behavior, or to explore the influence of larger structural forces on individuals, or vice versa.

Experiments, particularly involving standard games, also populate behavioral economics, which tends to employ these games in interaction with formal models (Camerer et al., 2011). Indeed, the standard formula for a published paper in behavioral economics tends to begin with a formal model that is then tested empirically using experiments with real life subjects. The results of these studies are then used to adjust the formal model to become more consistent with real world behavior. The value of employing standardized games in experimental contexts is that they both simplify and standardize challenges and responses in ways that make it easier to compare across individuals. In this regard it is important to note, however, that not all experimental economics takes place in the area of behavioral economics. There has been, and continues to be, a great deal of work in experimental economics testing such things as equilibrium predictions in games (Goeree et al., 2016) that does not fall under the rubric of behavioral economics, which tends to focus more on how individuals violate assumptions of classical economics in ways that appear as failures of rationality.

The use of standard games derived from behavioral economics is not simply restricted to economics. Political scientists are increasingly employing these procedures particularly in the context of field experiments (Gerber and Green, 2012). Field experiments can mean different things in different fields. In most political science contexts, they involve conducting experiments in public settings – either within American domestic contexts, such as large voting studies (Arceneaux, 2005), or abroad in other countries to explore a wide variety of phenomena, such as using social media to track violence in Gaza (Zeitsoff, 2011) and exploring clientelism in Benin (Wantchekon, 2003). In these field contexts, the use of standard games

derived from economics can be particularly useful in the context of cross-cultural work. For example, work in this vein undertaken by political scientists has explored the ways in which individuals from different ethnic groups might respond to one another by seeing how much money people pass along to others from the same or different groups in a standard dictator or trust game (Bahry et al., 2005; Whitt and Wilson, 2007).

Field experiments are not, however, restricted to behavioral economic games to investigate their phenomena of interest. In economics, field experiments are most often used to examine issues related to economic development and inequality, particularly in places like India (Duflo, 2006). Such work often, but not always, utilizes either survey or so-called list (Harrison and List, 2004) experimental designs. Survey designs (Mutz, 2011) employ a between subjects design by administering slightly different versions of the exact same question to different respondents and then examining the differences in their responses. The value of such designs is that they can be administered in the context of nationally representative populations, allowing for a combination of the internal validity offered by lab experiments with close control over the variables and manipulations of interest, with the external validity provided by large and representative samples of the population as respondents. In political science, survey experiments have also been used in field contexts to examine a wider variety of topics, including such issues as political corruption in Brazil (Weitz-Shapiro and Winters, 2017). I am currently involved in large-scale cross-cultural work using survey experiments to explore differences in attitudes toward gender inequality by sex and religion of respondent across a variety of substantive domains.

List experiments ask subjects to give the number of items on a list to which they might respond in a certain way, for example negatively or positively. Their responses are then compared with a control group where

the sensitive item is not included, allowing investigators to examine the extent to which a socially sensitive issue like race or gender might be causing responses without forcing the individual to feel 'outed' by having to state which item affected their response. Such designs have been used to explore a wide variety of topics in economics, including labor markets and charity, and in political science they tend to be used to examine such topics as support for government forces in Afghanistan (Blair et al., 2014).

## STARTING ASSUMPTIONS

The starting assumptions of each discipline remain largely implicit and are often recognized by practitioners only when they are violated. As mentioned above, some of these assumptions relate to the preferred operative level of analysis. In the three subfields of political psychology, social psychology and behavioral economics, unlike much of the larger field of political science, the default level begins and ends with the individual and very little work concentrates on the level of the group, the domestic state, the regime type or economic system or the international system.

In addition, implicit assumptions regarding what constitutes an interesting or viable question, particularly around issues deeply inculcated with existential identification such as 'rationality', often cause practitioners steeped in different disciplines to talk past each other while ostensibly exploring the same issue. Indeed, many of the differences which unite these areas, but divide them from their respective larger disciplines, revolve around these assumptions of human rationality – assumptions which these areas largely reject, although dominant models in both political science and economics continue to embrace this descriptively inaccurate assumption. As a result, the following discussion will concentrate on these alternative

constructions of rationality because they have been so influential in driving work across fields – similar differences could be shown in other areas as well, but notions of rationality constitute the political ideological divide across disciplines and thus provide the greatest insight into divergences in starting assumptions across fields.

To be clear, the larger disciplines of both political science and classical economics tend to buy into traditional notions of human rationality at both a descriptive and a normative level. Specifically, they tend to start with an assumption of rational choice, arguing that when confronting a choice, the vast majority of healthy individuals will assess, either implicitly or explicitly, the worth and probability of each option and then decide on the one that offers the best combination of odds of getting what they want, along the lines of classic subjective expected utility theory (von Neumann and Morgenstern, 1947, based on earlier 18th-century work on expected utility by Bernoulli). This has proven to be a very useful assumption for many areas of classical economic theory.

However, work in the area of psychology, pioneered by Amos Tversky and Daniel Kahneman (1974; Kahneman and Tversky, 1979, 1984), called these assumptions into question from an empirical standpoint in groundbreaking and hugely influential work on judgmental biases and Prospect Theory, which has been cited over 50,000 times across a wide variety of fields. The work on Prospect Theory won Kahneman the Nobel Prize in Economics in 2002, which he shared with Vernon Smith, another founding father of behavioral economics. In many ways, this work interrogating the nature of human biases in decision making created the subfield of behavioral economics, and has continued to deeply influence the development of the field at large. Indeed, the 2017 recipient of the Nobel in Economics, Richard Thaler, is another behavioral economist famous for his use of methods and theories drawn from psychology to explore and intervene in various

economic phenomena, including the savings rate (Thaler, 2015). The origin of the use of behavioral economics in psychology is part of the reason why behavioral economics relies so heavily on experimental procedures to generate and test its theories and assumptions; it grew out of a field where only experiments are considered credible for clearly demonstrating cause and effect in human attitudes and behavior.

The importance of the biases that Kahneman and Tversky rigorously demonstrated in their elegant experimental work documented the ways in which individuals do not conform to the assumptions of rationality assumed by economists working in the traditional mold. This was very important because it forced many economists to have to confront the empirical accuracy of their starting assumptions around the nature of individual rationality. Kahneman and Tversky were not the first to show limitations in human cognitive processing; indeed, Herbert Simon's (1982) very influential work on satisficing and bounded rationality preceded their work.

While this more empirically grounded model of decision making forced many economists to have to grapple with the discrepancies between their models and reality (Gächter et al., 2009), this was not the only attack mounted against established approaches to rationality. Subsequent challenges have come from the area of evolutionary psychology, which points not only to inadequacies in models of human rationality offered by economists, but also to models which presented deviations from optimal rationality as errors or mistakes. This alternative model (Gigerenzer and Goldstein, 1996) suggested that human rationality remained necessarily contingent on environmental circumstances. As a result, people possessed a kind of 'ecological' rationality that uses very effective and efficient psychological mechanisms developed over the course of evolution by natural selection to render near optimal decision-making strategies from the perspective of survival. This perspective draws on

various strands of research for support, from ethological work showing similar strategies among chimpanzees (Lakshminarayanan et al., 2011) to work showing that what might appear to look like 'irrational' or 'biased' behavior actually makes sense within the context of humans who evolved in small-scale societies. For example, abstract inability to make associations becomes possible in the context of trying to detect cheaters (Cosmides and Tooby, 1992), and the well-known risk insensitivity demonstrated in the classic disease paradigm reverses when applied to very low level population sizes (Rode and Wang, 2000). And certainly it is possible to provide an evolutionary explanation for the modern divergences in risk sensitivity demonstrated in the classic Tversky and Kahneman experiments (McDermott et al., 2008).

These differences in approach do not preclude the possibility that convergences cannot be reached or found. Certainly, theoretical approaches drawn from evolutionary models have reached near consensus across most areas of the life and behavioral sciences, including psychology, anthropology and medicine, although they have so far proved far less influential in the fields of sociology, political science and classical economics. However, other disciplinary integration has occurred and more is possible. For example, Herbert Gintis (2009) has suggested that greater unification across disciplines, including the incorporation of anthropological approaches that study human interaction in small-scale societies, might benefit behavioral economics in particular, and help improve notions of rationality more generally. This kind of approach has already yielded very interesting interdisciplinary work on the nature of leadership (von Rueden and van Vugt, 2015; von Rueden et al., 2014) and status (Cheng et al., 2010), among other topics of great interest to political scientists.

In addition, methodological synthesis is becoming increasingly possible. Camerer et al. (2005) have argued for and shown how neuroscientific advances, particularly in the

use of magnetic resonance imaging (MRI) technology, can be used to enlighten our understanding of how the brain operates, and provide a methodological foundation for a more interdisciplinary approach to the study of decision making. Such work might unite behavioral economics and various strands of psychology, including social, cognitive and evolutionary approaches. By contrast, political science has been remarkably slow to adopt or incorporate such methods into investigations of political behavior (for an exception see Dawes et al., 2012). This likely results primarily from the costs and challenges associated with learning how to use MRIs, but likely also derives from an overall skepticism within the discipline regarding the importance of individual action and behavior for understanding the large-scale organizational and structural phenomena that constitute the primary area of interest and investigation for the vast majority of political scientists.

## **CENTRAL RESEARCH ISSUES AND QUESTIONS**

Perhaps most important of all the divisions noted here, and certainly presaged by the discussion of the notion of rationality just above, all these sources of division and divergence are hugely shaped by the central research issues and questions that preoccupy each discipline. This drives, in part, the methodological differences that divide these fields. Clearly, different substantive areas of inquiry demand different procedures and methods for proper investigation. However, unlike many trends supporting the value of interdisciplinary study, not all questions and methods are considered equally legitimate across disciplines, and often they are not interchangeable. Further divides around the proper level of analysis for investigation can separate scholars who might otherwise share substantive interests, concerns and

questions. Each discipline concentrates on a slightly different yet particular set of aims and questions, and these divergent intellectual intentions infuse the culture and status hierarchies implicit (and sometimes explicit) both within and across fields. Yet different fields such as political science and behavioral economics do continue to share an interest in particular substantive areas, such as voting and deliberation, although they may each take quite different approaches to the same problem.

Political psychology tends to be a quite diffuse field, with practitioners interested in a wide variety of topics that span theory, method and substantive area of investigation. Topics of legitimate inquiry run the gamut from historical psychobiographies (Runyan, 1982) to genetic work on the relationship between political ideology, emotion and attitudes toward important political events (Hatemi and McDermott, 2011). As a result of the diversity in relevant topics, theories and methods, and the lack of clear disciplinary boundaries as identified, contained and restrained within departmental structures, there is less policing around what is and is not relevant or appropriate work in political psychology. The downside of this freedom is that much of this work is often of lower quality because of the relative lack of rigor and discipline.

Social psychology tends to be most interested in issues surrounding individual behavior in the social world, often around topics related to social affiliation (Ross and Nisbett, 2011). This can include a wide range of topics, from impression management to questions surrounding the nature of conformity (Asch, 1955), obedience (Milgram, 1974) and prejudice and discrimination (Dovidio and Gaertner, 1986). Several areas offer tremendous overlap with issues of primary interest to political scientists. These often involve the investigation of attitudinal measure to examine, explore and predict problematic or sensitive aspects of human social behavior, including topics related to

prejudice and discrimination on the basis of categorization (Rosch, 1999; Adolphs, 2003), which may be by age, gender, race, religion, sexual orientation, ethnicity or a host of other factors (Tajfel, 1981). However, these are not the only areas of potential overlap. Others relate to work on in-group preference and outgroup discrimination (Tajfel, 2010) which has implications not only for domestic politics but also for inter-group conflict, including war. Such implications are of interest to international relations and comparative politics scholars.

Behavioral economics, being focused more on economic issues, tends to examine ways in which individuals systematically violate the assumptions inherent in classical economic models. These examinations include further explorations of the iconic Kahneman and Tversky (1979, 1984) work on framing effects and loss aversion in particular. Interesting and important work extending these original insights has explored the nature and function of reference points (Rabin, 2002), including the way that contracts can function as reference points (Hart and Moore, 2008). Some of the big research agendas include questions related to mental accounting (Thaler, 2008) and changing behavior using small or so-called nudge interventions (Thaler and Sunstein, 2008). In the latter area, efforts have included work on health, such as stop smoking campaigns. In addition, there is a great deal of interest in issues surrounding the discounting of time in making decisions (Loewenstein and Prelec, 1992), and increasing attention is being concentrated on the influence of emotion on decision making (Lempert and Phelps, 2014). Topics of interest that provide the greatest area of overlap with area of concern to social psychologists and political scientists involve those areas where welfare decisions remain fundamentally contingent on social variables and choices. These include topics related to trust (Fehr, 2009), fairness and reciprocity (Fehr and Gächter, 2009), equity (Fehr and Schmidt, 1999), third party

or altruistic punishment for enforcing social norms toward cooperation (Fehr and Gächter, 2002) and how preferences and belief formation can effect behavior such as free riding in public goods provision and usage (Fischbacher and Gächter, 2010).

In short, the preoccupation of each discipline, in terms of the topics each finds interesting, engaging and worthy of serious and sustained investigation, closely mirrors the content of the disciplinary focus. Political psychologists, like political scientists, tend to focus on issues related to political organization and ideology, including topics related to leadership and conflict. In particular, political psychologists, like their colleagues in the larger discipline, tend to focus on issues of power and the influence of its imposition on a wide variety of individuals and phenomena. Psychologists, including social psychologists, tend to find issues of social behavior most important, particularly those related to processes of social affiliation. Behavioral economists, by contrast, tend to focus on topics related to money, finances and the social processes that affect exchange.

Clearly, areas of overlap exist between these disciplinary cultures and can be explored within the context of overlapping acceptable methods of inquiry, especially in the area of experimental work. The topics that could easily be explored in interdisciplinary contexts using experimental methods might most easily investigate the effect of particular attitudes or emotions on behaviors related to conflict resulting from power differentials. This includes not only social areas of conflict such as prejudice and discrimination, but also the causes and consequences of economic inequality that both cause and result from such processes. Projects such as this, or work on many other possible topics, could clearly engage individual scholars from across disciplines using mutually acceptable theories and methods to investigate social and political phenomena of great impact and importance for all of society.

## **DIFFERENCES IN DISCIPLINARY CULTURE**

Understanding how to negotiate and translate the various theoretical, methodological and content divides which permeate each sub-field can prove challenging, but being able to do so helps facilitate more efficient and effective interdisciplinary work, allowing for greater progress in illuminating critical areas of social inquiry that cross disciplines.

There are a couple of important cultural differences which are worth further explicit mention since they can often, in and of themselves, prevent scholars coming from different traditions from working together. In some ways, the two most significant ones, related to incentives and deception, reflect different ethical perspectives, but not in the traditional way in which ethics is understood by, say, institutional review boards, who might find all approaches equally acceptable under specific conditions.

The first issue relates to incentives where the norms and rules across disciplines remain contentious. Psychologists often use undergraduate student populations as their subject pool; the students receive course credit for their participation in experiments. These students may be required to participate in a certain number of experiments in order to receive course credit in an introductory class, or may receive extra credit for participating in experiments in upper division courses.

Economists, by contrast, do not believe that honest results can be obtained from subjects without providing monetary incentives to give individuals a stake in the outcome of their behavior and to encourage sufficient attention and honesty in the given tasks (Smith et al., 1993). Indeed, most economics journals will not publish results that come from studies where subjects have not been properly remunerated for their participation. This does not mean, however, that behavioral economists are always ignorant of the costs and consequences associated with only using



monetary incentives to investigate behavior. Indeed, Loewenstein (1999) argues that one of the consequent problems associated with this norm of monetary incentives is that it does not allow for observers to understand the complex ways in which monetary and non-monetary rewards and incentives might interact in the real world: if investigators only ever interrogate the former, they will never be able to provide an empirical foundation for understanding the latter.

These perspectives emerge, if only implicitly, from quite divergent notions of human nature. Psychologists believe that people do all sorts of things for all kinds of reasons, including biological ones. So, for instance, they would agree that people will find sex, food, drink, sleep or other biological inducements to be as compelling, if not more so, than money; indeed, they would argue that individuals simply use money to obtain these other goods and services, and so leaving money out of the equation serves only to simplify the reinforcement. And, indeed, some behavioral economists might agree. For example, Camerer (2007) deprived subjects of food and liquid for a few hours prior to an experiment and then gave them salty treats prior to the experiment. The experiment on time discounting on savings rate in retirement then allowed people to play for money or juice. Not surprisingly, more people chose the juice, but this choice also affected other aspects of their financial behavior, showing the utility of varying incentives in order to understand the variety of human behavior in the real world. However, by and large, economists believe that money serves as a unified proxy of value that can incentivize subjects across different backgrounds and goals. This assumption depends on an extremely narrow view of human nature, and the drives and goals that motivate the diversity of human behavior, but as long as the disciplinary incentives for publication in economics remain narrowly focused on using money as an incentive, all experiments in that area will continue to do so.

It is likely that an increasing number of psychological experiments will use money as an incentive, as the use of web based platforms such as mturk or qualtrics increasingly dominates the experimental space. Because these platforms offer much wider subject pools in terms of age and other characteristics, and because they often provide much faster rates of response among larger groups of people, it is likely that the monetary incentives that drive respondents on such platforms will facilitate an increasing convergence across fields toward using money as the primary incentive in experimental studies. This does not mean that more creative forms of incentives such as those employed by Camerer cannot be used in laboratory contexts.

The issue of deception is, if anything, even more divisive between subfields than that posed by divergent incentive structures. Standard economic textbooks caution that under no circumstances should deception be allowed in or introduced into experiments (Friedman and Sunder, 1994). Usually, the justification revolves around contaminating the subject pool, assuming that subjects who have been deceived will be less likely to be willing to participate in experiments and less honest when they do so (Jamison et al., 2008). Note that, as with psychologists who have not traditionally used monetary incentives having to shift their behavior in the wake of the increasing dominance of platforms such as mturk, economists will have a very difficult time sustaining arguments about crossover effects within population samples that number in the millions where communication – while not impossible over the web – is highly unlikely to affect people whose primary motivation for participation is the money which economists highly prize. In addition, as Gneezy (2005) points out, the financial world is filled with deception, including the huge impact of tax evasion on many other aspects of social welfare: to refuse to study deception renders a great deal, if not the majority, of economic behavior outside the realm of experimental investigation.

Furthermore, while ultimately deciding that lying has consequences, Gneezy notes that the entire model of 'homo economicus' is founded on the notion of the value of a selfish economic actor, and lying is often an intrinsic part of acting selfishly. This observation further highlights the discrepancy between economic models and economic methods of investigation. Moreover, Dreber and Johannesson (2008) describe significant gender differences in deception, showing that men are much more likely than women to lie for purposes of financial gain, demonstrating that the role of deception is not gender neutral in origin or consequence.

Psychology has been grappling with the use of deception for quite a while (Kelman, 1967). Interestingly, most of the work from behavioral economics condemning deception depends on arguments based in moral philosophy, with little research to back up the substantive claims regarding damage to participants. Christensen (1988) provides some experimental data on this very outcome variable, finding that 'subjects who have participated in deception experiments versus nondeception experiments enjoyed the experience more, received more educational benefit from it, and did not mind being deceived or having their privacy invaded'. This is likely due, at least in part, to experiments involving deception being more engaging or involving topics of greater interest to subjects. This is because even experimenters in psychology treat deception in the way that Hillary Clinton once talked about abortion: 'legal, but rare'. Psychologists tend to use deception when investigating those circumstances or situations that remain socially sensitive, such as race and gender discrimination, where it would be easy to expect subjects to lie for reasons of impression management. Yet, most psychologists believe that these topics are so important that they should not be entirely dismissed as potential sources of investigation. In this way, the use of deception in carefully designed experiments provides one way to research socially sensitive topics in a manner

designed to get at individuals' thoughts, feelings and behaviors in accurate ways. There is some work showing that one way in which any negative consequences deriving from the use of deception in experiments can be ameliorated is through employing careful debriefing procedures (Smith and Richardson, 1983). There is also some evidence that forewarning (Allen, 1983) subjects that deception might be used can mitigate some of the ethical issues involved in deception, while not necessarily precluding the possibility that subjects can still be properly deceived as to the real purpose of the experiment or source of the manipulation.

Social psychologists and behavioral economists are likely to converge in unexpected, and possibly undesired, ways in the coming years, as online platforms such as mturk and qualtrics take over more and more of the experimental space. Psychologists who use these populations will have to provide monetary compensation for participation. In addition, economists who use these platforms will have to acknowledge that reputational costs and contamination are extremely unlikely to occur within such wide and diverse populations, where the incentives for communication around these issues are extremely low and costly in terms of time and energy, with very little subjective expected reward resulting from such communication. That said, it will prove much more difficult to deceive people via online platforms, and so the ability to do so will decline as well. In addition, it should be noted that there is a very real cost associated with outsourcing the majority of experimental work to online platforms. They may prove quick and efficient in obtaining large numbers of subjects at low cost, albeit with unknown and likely variant quality of responses, but such studies will necessarily need to focus more on attitudes and less on behavior. While this does not render attitudes unimportant or insignificant, meaningful but difficult to observe aspects of human behavior may increasingly fall outside areas of investigation, leaving us knowing more and

more about fewer and fewer pervasive and powerful aspects of human experience.

## CONCLUSION

This chapter has attempted to highlight some of the significant overlaps, similarities and divergences in method, assumption and content between political psychology, social psychology and behavioral economics. Dominant theories and methods of investigation differ across these sub-disciplines according to the differences in the primary questions and aims which preoccupy each field.

Meaningful and substantive differences exist across these fields. Some of these involve fundamental underlying assumptions. For example, economists and political scientists espouse a fundamentally different understanding of the nature of human rationality, and indeed often possess opposing definitions of what constitutes it. In addition, the primary focus of interest differs across fields, with political scientists primarily interested in power, psychologists in affiliation and economists in money. Furthermore, economists and psychologists have very different values regarding the nature of incentives and the role of deception in experiments.

However, important points of convergence and overlap do exist and are worth cultivating in service of greater opportunities for mutually enriching interdisciplinary work. These involve the obvious convergence around the value and utility of experimental methods of various sorts, including field, survey and list experiments, for investigating important aspects of human social, political and economic behavior. In addition, there is a great deal of overlap in substantive areas of consideration, including issues related to attitudes and emotions involving discrimination, inequality and conflict. These areas offer enormous potential for integration across fields to illuminate critical aspects of human decision making and behavior.

## REFERENCES

- Adolphs, R. (2003). Cognitive neuroscience: cognitive neuroscience of human social behaviour. *Nature Reviews Neuroscience*, 4(3), 165–178.
- Allen, D. F. (1983). Follow-up analysis of use of forewarning and deception in psychological experiments. *Psychological Reports*, 52(3), 899–906.
- Arceneaux, K. (2005). Using cluster randomized field experiments to study voting behavior. *The Annals of the American Academy of Political and Social Science*, 601(1), 169–179.
- Aronson, E., Carlsmith, J. M., Ellsworth, P. C. & Gonzales, M. H. (1990). *Methods of Research in Social Psychology* (2nd ed.). New York: McGraw-Hill.
- Asch, S.E. (1955). Opinions and social pressure. *Scientific American*, 193(5), 31–35.
- Bahry, D., Kosolapov, M., Kozyreva, P. & Wilson, R. K. (2005). Ethnicity and trust: evidence from Russia. *American Political Science Review*, 99(4), 521–532.
- Blair, G., Imai, K. & Lyall, J. (2014). Comparing and combining list and endorsement experiments: evidence from Afghanistan. *American Journal of Political Science*, 58(4), 1043–1063.
- Camerer, C., Loewenstein, G. & Prelec, D. (2005). Neuroeconomics: how neuroscience can inform economics. *Journal of Economic Literature*, 43, 9–64.
- Camerer, C. F. (2007). Neuroeconomics: using neuroscience to make economic predictions. *The Economic Journal*, 117(519), C26–C42.
- Camerer, C. F., Loewenstein, G. & Rabin, M. (eds). (2011). *Advances in Behavioral Economics*. Princeton University Press.
- Cheng, J. T., Tracy, J. L. & Henrich, J. (2010). Pride, personality, and the evolutionary foundations of human social status. *Evolution and Human Behavior*, 31(5), 334–347.
- Christensen, L. (1988). Deception in psychological research: when is its use justified? *Personality and Social Psychology Bulletin*, 14(4), 664–675.
- Cosmides, L., & Tooby, J. (1992). Cognitive adaptations for social exchange. In J. H. Barkow, L. Cosmides & J. Tooby (eds.) *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*, pp. 163–228. Oxford University Press.

- Dawes, C. T., Loewen, P. J., Schreiber, D., Simmons, A. N., Flagan, T., McElreath, R., Bokemper, S. E., Fowler, J. H. & Paulus, M. P. (2012). Neural basis of egalitarian behavior. *Proceedings of the National Academy of Sciences*, 201118653.
- Dovidio, J. F., & Gaertner, S. L. (1986). *Prejudice, Discrimination, and Racism*. Academic Press.
- Dreber, A., & Johannesson, M. (2008). Gender differences in deception. *Economics Letters*, 99(1), 197–199.
- Duflo, E. (2006). Field experiments in development economics. *Econometric Society Monographs*, 42, 322–348.
- Fehr, E. (2009). On the economics and biology of trust. *Journal of the European Economic Association*, 7(2–3), 235–266.
- Fehr, E., & Gächter, S. (2000). Fairness and retaliation: the economics of reciprocity. *Journal of Economic Perspectives*, 14(3), 159–181.
- Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature*, 415(6868), 137.
- Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics*, 114(458), 817–868.
- Fischbacher, U., & Gächter, S. (2010). Social preferences, beliefs, and the dynamics of free riding in public goods experiments. *American Economic Review*, 100(1), 541–556.
- Friedman, D., & Sunder, S. (1994). *Experimental Methods: A Primer for Economists*. Cambridge University Press.
- Gächter, S., Orzen, H., Renner, E. & Starmer, C. (2009). Are experimental economists prone to framing effects? A natural field experiment. *Journal of Economic Behavior & Organization*, 70(3), 443–446.
- Gerber, A. S., & Green, D. P. (2012). *Field Experiments: Design, Analysis, and Interpretation*. New York: W.W. Norton.
- Gigerenzer, G., & Goldstein, D.G. (1996). Reasoning the fast and frugal way: models of bounded rationality. *Psychological Review*, 103(4), 650–669.
- Gintis, H. (2009). *The Bounds of Reason: Game Theory and the Unification of the Behavioral Sciences*. Princeton University Press.
- Gneezy, U. (2005). Deception: the role of consequences. *American Economic Review*, 95(1), 384–394.
- Goeree, J. K., Holt, C. A. & Pfaffrey, T. R. (2016). Quantal response equilibria. In S. N. Durlauf and L. E. Blume (eds.) *Behavioural and Experimental Economics*. Palgrave Macmillan.
- Harrison, G. W., & List, J. A. (2004). Field experiments. *Journal of Economic Literature*, 42(4), 1009–1055.
- Hart, O., & Moore, J. (2008). Contracts as reference points. *The Quarterly Journal of Economics*, 123(1), 1–48.
- Hatemi, P. K., & McDermott, R. (eds). (2011). *Man Is by Nature a Political Animal: Evolution, Biology, and Politics*. University of Chicago Press.
- Jamison, J., Karlan, D. & Schechter, L. (2008). To deceive or not to deceive: the effect of deception on behavior in future laboratory experiments. *Journal of Economic Behavior & Organization*, 68(3–4), 477–488.
- Kahneman, D., & Tversky, A. (1979). Prospect theory: an analysis of decision under risk. *Econometrica*, 47(2), 263–292.
- Kahneman, D., & Tversky, A. (1984). Choices, values, and frames. *American Psychologist* 39(4), 341–350.
- Kelman, H. C. (1967). Human use of human subjects: the problem of deception in social psychological experiments. *Psychological Bulletin*, 67(1), 1–11.
- Lakshminarayanan, V. R., Chen, M. K. & Santos, L. R. (2011). The evolution of decision-making under risk: framing effects in monkey risk preferences. *Journal of Experimental Social Psychology*, 47(3), 689–693.
- Lempert, K. M., & Phelps, E. A. (2014). Neuroeconomics of emotion and decision making. In P. W. Glimcher and E. Fehr (eds.) *Neuroeconomics: Decision Making and the Brain* (2nd ed.), pp. 219–236. Academic Press.
- Loewenstein, G. (1999). Experimental economics from the vantage-point of behavioural economics. *The Economic Journal*, 109(453), F23–F34.
- Loewenstein, G., & Prelec, D. (1992). Anomalies in intertemporal choice: evidence and an interpretation. *The Quarterly Journal of Economics*, 107(2), 573–597.
- McDermott, R. (2004). *Political Psychology in International Relations*. University of Michigan Press.
- McDermott, R., Fowler, J. H. & Smirnov, O. (2008). On the evolutionary origin of prospect theory preferences. *The Journal of Politics*, 70(2), 335–350.

- Milgram, S. (1974). *Obedience to Authority*. New York: Harper & Row.
- Mutz, D. C. (2011). *Population-Based Survey Experiments*. Princeton University Press.
- Rabin, M. (2002). A perspective on psychology and economics. *European Economic Review*, 46(4–5), 657–685.
- Rode, C., & Wang, X. T. (2000). Risk-sensitive decision making examined within an evolutionary framework. *American Behavioral Scientist*, 43(6), 926–939.
- Rosch, E. (1999). Principles of categorization. In E. Margolis & S. Laurence (eds.) *Concepts: Core Readings*, pp. 189–206. Cambridge, MA: MIT Press.
- Ross, L., & Nisbett, R. E. (2011). *The Person and the Situation: Perspectives of Social Psychology*. London: Pinter & Martin Publishers.
- Runyan, W. M. (1982). *Life Histories and Psychobiography: Explorations in Theory and Method*. Oxford University Press.
- Simon, H. A. (1982). *Models of Bounded Rationality*. Cambridge, MA: MIT Press.
- Smith, S. S., & Richardson, D. (1983). Amelioration of deception and harm in psychological research: the important role of debriefing. *Journal of Personality and Social Psychology*, 44(5), 1075–1082.
- Smith, V. L., & Walker, J. M. (1993). Monetary rewards and decision cost in experimental economics. *Economic Inquiry*, 31(2), 245–261.
- Tajfel, H. (1981). *Human Groups and Social Categories: Studies in Social Psychology*. Cambridge University Press Archive.
- Tajfel, H. (ed.) (2010). *Social Identity and Intergroup Relations*. Cambridge University Press.
- Thaler, R. H. (2008). Mental accounting and consumer choice. *Marketing Science*, 27(1), 15–25.
- Thaler, R. H. (2015). *Misbehaving: The Making of Behavioral Economics*. London: Allen Lane.
- Thaler, R. H., & Sunstein, C. (2008). *Nudge: Improving Decisions about Health, Wealth, and Happiness*. Yale University Press.
- Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: heuristics and biases. *Science*, 185(4157), 1124–1131.
- von Neumann, J., & Morgenstern, O. (1947). *Theories of Games and Economic Behavior*. Princeton University Press.
- von Rueden, C., Gurven, M., Kaplan, H. & Stieglitz, J. (2014). Leadership in an egalitarian society. *Human Nature*, 25(4), 538–566.
- von Rueden, C., & van Vugt, M. (2015). Leadership in small-scale societies: some implications for theory, research, and practice. *The Leadership Quarterly*, 26(6), 978–990.
- Wantchekon, L. (2003). Clientelism and voting behavior: evidence from a field experiment in Benin. *World Politics*, 55(3), 399–422.
- Weitz-Shapiro, R., & Winters, M. S. (2017). Can citizens discern? Information credibility, political sophistication, and the punishment of corruption in Brazil. *The Journal of Politics*, 79(1), 60–74.
- Whitt, S., & Wilson, R. K. (2007). The dictator game, fairness and ethnicity in postwar Bosnia. *American Journal of Political Science*, 51(3), 655–668.
- Zeitsoff, T. (2011). Using social media to measure conflict dynamics: an application to the 2008–2009 Gaza conflict. *Journal of Conflict Resolution*, 55(6), 938–969.



# Institutional Theory and Method

Maxfield J. Peterson and B. Guy Peters

## INTRODUCTION

### *Institutions Matter!*

In the period following the Second World War, political science was characterized by the rapid growth of the behavioral research program, which emphasized the role of individual attributes and decisions in political outcomes. For a time, the ‘behavioral revolution’ seemed to eclipse what had been the dominant paradigm of political science research – institutionalism. Yet despite the relevance of behavioralism, the current wave of populism sweeping the United States and Europe signals the enduring impact of institutions in the political world. When politicians and their supporters rail against the ‘political establishment’, they are articulating a rejection of institutions, but also recognizing the significance of those institutions. If the study of individual political identities merits the amount of scholarly attention it has received, a focus on the institutions that

structure and inform those identities, intentionally or unintentionally, seems equally, if not more, important.

Of course, the contention that institutions play a central role in the construction of political identities is itself an assumption that occupies a central role in the structure–agency debate. The goal of this chapter is to identify and investigate this and other assumptions of the institutional perspective, review the literature that has constructed and advanced institutional theory and discuss how institutionalism, and its perennial focus on structure, remains one of the most relevant approaches to social inquiry.

This chapter proceeds as follows. We begin with a review of the central place of institutions and institutionalism in political science, beginning with their role in ancient political thought, tracing this theoretical tradition through the ‘Old Institutionalism’ of the late nineteenth and early twentieth centuries, and connecting this work to the ‘New Institutionalism’. We then review the major

approaches in institutional research, highlighting the connections between the distinct theoretical foundations of different types of institutionalism and their respective research strategies. Having reviewed institutional theory and approaches, we argue that the theoretical assumptions of institutionalism inform a set of methodological approaches and tools, which we review and discuss. We then face a set of paradoxes in institutional theory, exploring some possible resolutions. Finally, we review some central challenges facing scholars of institutionalism, and offer some concluding remarks.

### ***Central Place of Institutions, and Institutionalism, in Political Science***

The first scholars of politics were students and architects of institutions. One of the oldest political texts, Plato's *Republic* (1992), is a discourse on how to properly structure institutions so as to produce a 'just' society. Plato's student (and perhaps the first empiricist) Aristotle continued this ancient institutionalism in *Politics* (1998), in which he constructed a taxonomy of state systems for purposes of comparison. The wide dissemination of these ancient texts some two millennia later ushered in the Age of Enlightenment and its seminal political philosophers. Thinkers such as Thomas Hobbes, John Locke and J. S. Mill focused on central questions concerning the relationship between state and society, the rights of the individual, and the form of government best suited to serve their respective conclusions on such issues. Enlightenment liberalism as articulated by John Locke, Jean-Jacques Rousseau and Baron de Montesquieu was at its core an institutionalist argument for the establishment of a particular form of limited government that structured a society favorable to property rights, political and religious freedom and constrained, non-radical democracy. Such institutional thought laid the canvas for early American statesmen such

as James Madison, Alexander Hamilton, John Jay and Thomas Jefferson. The detailed arguments for ratification of the constitution found in Madison, Hamilton and Jay's *Federalist Papers* (Hamilton et al., 2009) represents a classic work of institutionalist reasoning.

While the institutionalism of Enlightenment liberalism and early American thought resonates with the Anglo-contractualism of John Locke and Thomas Hobbes, the thinkers of 19th and early 20th-century continental Europe created a distinct strand of institutional thought. Although initially more associated with sociology than political science, the works of Hegel, Max Weber and Émile Durkheim have come to occupy such an important place in the interdisciplinary study of institutions that they merit some brief review here. In contrast to the anti-statism of Anglo-American liberal thought, the state remained central to the political life of continental Europe and its thinkers. For example, Weber saw the state as the organic evolution of authority from traditional forms. Charting the development of the German bureaucracy from its monarchic birth to its role as the meritocratic administrator of modern government, Weber argued institutions were fundamental to the life and character of the state. These arguments would be central to bringing the state back into the work of American scholars.

The German school of sociological institutionalism is key to understanding the growth of political science as a distinct academic discipline in the United States. The earliest American political scientists, such as Woodrow Wilson, formulated a legal-structural understanding of American democracy and its institutional foundations at least in part on the basis of Weberian thought. Wilson's construction of a politics-administration dichotomy, which argued for the insulation of the bureaucratic, executive apparatus from partisanship, is directly influenced by Weber's theory of autonomous, meritocratic bureaucracy. Wilson's contemporaries exuded a similarly state-centric view of political science, as can be seen in T. D. Woolsey's

*Political Science, or, The State Theoretically and Practically Considered*. These early American scholars were architects of the 'Old Institutionalism', a period of institutional thought that balanced an American intellectual heritage of legal-formal analysis and a more theoretical study of the state from the German school. But the emphases on legalism and structuralism (or the elevation of institutions, rather than individuals, as the object of study) supply only a partial understanding of the Old Institutionalism.

Together with legalism and structuralism, Old Institutionalism is characterized by holism, historicism, and a normative rather than positive frame of analysis. With regard to holism, Old Institutionalists tended to compare entire political systems rather than their component parts (i.e. British versus American democracy rather than British versus American courts). Holism assumes that institutions work together to comprise a coherent political system; thus a thorough understanding of the operation and reason of any component must be set in its political and cultural context. Old Institutionalism's historicism meant that institutions must be understood as embedded within historical processes that could account for their existence and functionality. As a mode of analysis, historicism fit well with holism in that it maintained the primacy of context and eschewed the idea that political institutions could be understood in isolation. Finally, Old Institutionalism was set in a normative framework where useful ideas and theories were developed for the purpose of improving society rather than positivist investigation. While this may come as a surprise to today's political scientists given the relegation of normative theory to an isolated subdivision of the discipline, the goal of developing theories for implementation in the real world made sense at a time when a political scientist was US President. Indeed, Woodrow Wilson's studies (1887, 1892) of the bureaucracy informed his brand of early American progressivism, demonstrating the direct link between scholarly

study of political institutions and the exercise of political power.

Further influence of the German sociological school can be seen in the work of early American organizational theorists such as Philip Selznick and Chester Barnard. While 'organizations' may not always be institutions, depending on one's definition (see below), the contributions of organizational theorists are essential to understanding institutions. For example, Selznick's argument (1943) that individuals within organizations function as independent agents with 'dual goal-sets' impacts the extent to which we assign institutions power over their members, and vice versa. Barnard's *The Functions of the Executive* (1968) defined the role of executives relative to the institutional apparatus in which they operate, another way of understanding the relationship between individual leaders and institutions. Further, his theory that organizational survival is predicated on effectiveness and efficiency is clearly an institutional analysis.

Old Institutionalism faded from importance upon the midcentury arrival of the behavioral revolution. In the post-war period, American political science became dedicated to leveraging reams of information to understand individual political behaviour such as vote choice and party identification. This was driven in part by the availability of information, and in part by the interest in individual decision-making. Publication of Campbell, Converse, Miller and Stokes' *The American Voter* in 1960 established the Michigan School at the center of the behavioral movement, under which a generation of American political scientists would be educated. It was not until Theda Skocpol's *States & Social Revolutions: A Comparative Analysis of Social Revolutions in Russia, France, and China* (1979) that institutionalism would return to center stage, albeit co-starring with the ongoing behavioral program. Skocpol's remarkable work demonstrated the value of historical institutional analysis in understanding a question of critical importance to



political scientists. The resounding influence of this work can be seen in the 1970s and 1980s resurgence of institutional analysis, or what James G. March and Johan P. Olsen would deem ‘The New Institutionalism’ (1984, 1989).

While Skocpol’s work can be seen as a bellwether of the coming influence of the new institutionalism more broadly, an extremely productive branch known as rational-choice institutionalism would be instrumental in the reiteration of the classic paradigm. Following Shepsle’s (1989) observation that the study of institutions could benefit from a methodological and perspectival synthesis with the rational-choice approaches, scholars began seeking ‘structure-induced equilibria’, or outcomes that could be seen as the aggregate product of utility-maximizing individuals operating within institutional constraints. By factoring institutional constraints into the mathematical models originally appropriated to predict voting behavior, one could produce outcomes in a much more reliable manner than by focusing on the unconstrained behaviors of individuals.

March and Olsen (1984, 1987) argued that institutional theory blending modern behaviorist influences with an older concern with institutions could produce a distinct and promising agenda for research. This new institutionalism included a rejection of behavioralism’s assumption that politics can aggregate individual preferences efficiently. New Institutionalism revived an understanding of history that appreciated the influence of institutions on political behavior and change. In this view, societal preferences were at a minimum constrained, and at a maximum informed and directed, by the institutional context. The New Institutionalism stood to benefit from the behavioral revolution by channeling gains in understanding individual behavior through an institutional prism. Thus, the New Institutionalism did not imply a return to the normative, holist or even necessarily historicist characteristics of its older form, but rather was a more

modern application of its emphasis on formal structures to the study of politics. One might consider a bowling metaphor, in which behaviorists attempted to understand the movement of the bowling ball with no attention paid to what sort of barriers might be erected on either side of the lane. The New Institutionalism asserted that renewed attention to those barriers would better our understanding of the possibilities for the ball’s movement in the first place.

This rejuvenated emphasis on the roles of institutions in shaping political outcomes is evident in work on legislatures (Tsebelis and Money, 1997; Tsebelis, 2002; Krepel, 2002; Saiegh, 2016), political economy (North, 1990; Fish, 1998; Johnson, 1982; Nölke and Vliegenthart, 2009) and regime change and democratization (Lijphart, 1999; Linz and Stepan, 1996; Evans and Skocpol, 1985; Herbst, 2000), among many others. As can be seen in these works, the new institutionalism is not monolithic in its approach. Some scholars have embraced the rational-choice assumptions common of behavioral work and applied them in institutional contexts, while others have hewed closer to the old institutionalism in their reliance on historicism. A more detailed discussion of the new institutionalism and its subgenres will follow later in this chapter. Here what is important to understand is that institutionalism, and its assumptions about the role of the state institutions – as well as institutions such as political parties – in guiding behavior, rules and incentives, has returned to a central place within the discipline. The new institutionalism represents a maturation of the discipline in its ability to incorporate some of the gains from behavioral research while restoring important theoretical assumptions that remain critical to understanding the political world.

### ***The Major Approaches***

Before entering a discussion of the relationship between institutional theory and

methodology, it is important to understand that institutionalism comprises a set of approaches built upon distinct ontologies which lend themselves to different methods. We now turn to a review of the major approaches within the institutionalist paradigm: normative, rational choice, historical, empirical and discursive. This will be followed by a discussion of how these different approaches (and ontologies) lend themselves to particular methodological choices. And we will also need to inquire whether there is one contemporary version of institutionalism, or many disparate strands of theorizing about institutions.

### *Normative*

When March and Olsen (1984) issued the call for a new institutionalism, they were essentially advocating what has come to be called normative institutionalism. Normative institutionalism attempts to understand the social and political world through analysis of the normative underpinnings of institutions; that is, the values, ideas and norms that form their basis. Normative institutionalism exists in sharp opposition to atomistic individualism, which assumes attitudes and actions to be exogenous to political systems and a product of the ideas, values and norms of individuals. Thus, rather than understanding action by understanding individuals, we should try to understand how institutions shape their behavior, and what the norms and values of that institution are. This does not mean, however, that the values and attitudes of any individual or group of individuals is easily equated to the values of a given institution. Individuals tend to be embedded in multiple overlapping institutions, which compete for influence. Some institutions might be more salient than others depending on the issue at hand. For example, membership in the American Medical Association may have little impact on how a doctor in a state legislature chooses to vote on a gay marriage bill, while that same doctor's membership in the Southern Baptist Convention

may have a profound influence on such a decision. Either way, in the normative institutional view, the doctor's decision to vote one way or the other is seen as a product of the influence of institutional membership rather than a rational, calculated decision.

The roots of normative institutionalism lie in the sociological school pioneered by Émile Durkheim and Max Weber. Durkheim argued that institutions, or 'the beliefs and modes of behaviour instituted by the collectivity', were at the heart of understanding society (1982 [1895]: 45). The study of such institutions could, in Durkheim's view, be leveraged to discover structural social facts – universal truths constructed by the existence and constancy of social and political institutions. A resounding early exercise in this approach is Weber's *The Protestant Ethic and the Spirit of Capitalism* (1958), in which Weber argued that Protestant ethics in Western Europe encouraged Protestants to engage with the secular market economy, forming a normative basis for the form of Western capitalism that continues to dominate the world's economic systems.

Philip Selznick argued for a distinction between two types of institutions: those that could be considered an aggregation of individuals' rational behaviors, and those which influence individual behavior through individuals' commitments to those institutions' goals. The former may be thought of as instrumental institutions: for example, individuals join a labor union to lobby for collective benefits that would be unattainable through purely individual advocacy. The latter may be thought of as consistent with March and Olsen's 'logic of appropriateness': individuals agree to donate a monthly membership fee to the American Civil Liberties Union (ACLU) because, although it may not yield clear individual benefits, they are committed to the preservation of civil rights. In this latter form, individuals may adopt beliefs and behaviors that the institution considers appropriate. They do so out of an understanding that the institutions' rules and recommendations serve values that they themselves hold.

### *Rational-choice*

Rational-choice institutionalism represents an approach that is distinct from the rest of the paradigm in its accommodation of the atomistic individual. Rational choice asserts that individuals maintain independent preferences that may be bounded by institutions. So although an individual may be a member of an institution, we cannot (as normative institutionalism might have it) understand their behavior solely by understanding that institution (or the combination of institutions) that may be influencing action). Instead, we should view institutions as impacting otherwise exogenous individual behavior by limiting the available choices for the individual and modifying the incentives that inform preferences (see Katznelson and Weingast, 2005).

Whereas much of institutionalism owes its intellectual heritage to sociology or contractual political philosophy, rational-choice institutionalism derives its assumptions from economics. The rational-choice model has become a powerful presence in economics and behavioral political science because of its ability to consistently predict outcomes across contexts based on a calculus of human activity. Individuals are seen as 'utility maximizers' who aim to gain the most benefit out of any situation. Thus, political outcomes can be understood as a function of the interactions between utility maximizers, who must bargain among each other to achieve optimal outcomes. Looking at the political world this way has several advantages; for example, it enables complex phenomena to be collapsed into abstract models, and these models can then be used to identify equilibria, or stable outcomes. A rational-choice model of behavior theoretically applies just as much in a back alley poker match in Calcutta as on a Wall Street trading floor.

Unlike pure rational-choice analyses, rational-choice institutionalism invites politics back in to understand how institutional rules, incentives and values may impact otherwise individual utility maximizing

behavior. As with institutionalism more broadly, rational-choice institutionalism is characterized by several different sub-categories that are defined as follows: historical, principal-agent, transaction-cost, veto-players. The historical tradition applies a rational-choice framework to historic periods in order to explain political, social or economic outcomes. A classic example is North and Weingast's analyses (1989; see also Greif, 2006) of the evolution of public institutions in 17th-century England, which details how institutional reforms drove rational actors to construct and invest in an open and regulated system of capital markets. More recently, the work of Elinor Ostrom, Larry Kiser and others involved in the Institutional Analysis and Development (IAD) framework has contributed to modern historical rational-choice institutionalism. This framework emphasizes the need for an interdisciplinary approach to institutional analysis that embeds the rational-choice assumptions as applied in the 'new institutional economics' into the accumulated knowledge of historians and anthropologists. This perspective has led to a series of works in political economy that emphasize the role of institutional development in the promotion of contract enforceability and communication in otherwise thinly institutionalized contexts where violence and cheating were commonplace (La Porta et al., 2008).

The principal-agent (PA) literature applies rational-choice assumptions to institutions' interactions with one another and with individuals (Miller, 2005; Carpenter and Krause, 2015). A principal (some individual, entity or group) with a desired outcome assigns an agent (another individual, entity or group) to be responsible for achieving that outcome. Classic principal-agent relationships include the public (principal) and their executive (agent), the legislature and the bureaucracy or the organizational leader and their subordinates. The principal agent formulation may provide a useful way of understanding

seemingly unorthodox behavior; agents may be allowed to act in ways inconsistent with the proximate goals or values of the principal if the principal believes their long term goal will be accomplished by allowing such latitude. For example, intelligence agencies (particularly in the United States) have been known to have a cavalier attitude toward legal authority, a stance which is often tacitly allowed by their principal (the president) because of the principal's belief in their ability and expertise in neutralizing security threats. Agents may also act in inappropriate ways because the institutional design of the PA relationship enables them to keep the principal in the dark.

The principal-agent model has been applied to understand a diverse array of institutional environments. In the study of public administration and bureaucracy, PA models have been applied to understand the relationship between leaders of organizations attempting to maximize resources for their organizations (wherein the leader serves as the agent and the organization as the principal) (Wildavsky, 1992; Savoie, 1990). PA approaches have become fairly standard in the analysis of regulatory policy in the United States, wherein Congress is cast as the principal and independent regulatory agencies are cast as the agents (McCubbins et al., 1989). In these applications, the object of curiosity is how institutions impact Congress' ability to keep the activities of independent regulatory authorities (the EPA, BLM, SEC) in line with legislative goals. Institutional variation can occur via incentive structures that motivate agents to actuate policy goals, or 'fire alarms' which empower the public or interest groups to alert Congress when agencies are in dereliction of their duties.

Another strand of the rational-choice institutionalism literature views institutions as mechanisms for reducing 'transaction costs'. Transaction costs can be understood as the inherent risks and inefficiencies involved in any transactions between two actors or

entities, the certitude that one receives payments for services rendered, the difficulty of communicating goals and plans across cultural or linguistic barriers or simply the time it takes for two actors to arrive at a mutually beneficial bargaining arrangement (Durant, 2006). Some of the previously discussed historical rational-choice institutionalist literature emphasizes this role, but it is perhaps most common in the comparative literatures on legislatures and the international relations literature on international cooperation. The comparative legislatures literature emphasizes how the committee system reduces transactions costs by (1) allowing legislators to specialize in policy arenas, enhancing their informational resources and thus reducing uncertainty about consequences; and (2) enabling legislators to 'credibly commit' to returning a policy favor, as their committee seat ensures continuing influence over a given policy arena.

In the international relations literature, the ability of institutions to make credible commitments has been important for the neoliberal institutionalist paradigm. Most notably advanced by Robert Keohane (2005), these scholars argue that institutions such as the World Trade Organization provide a forum for communication and mutual enforcement of deals. While nations could presumably renege on a multilateral trade agreement, reaping short-run benefits while evading long-run costs, such behavior is likely to be mutually punished by other institutional members, imposing a cost for such actions. Because nations recognize the costly 'shadow of the future' posed by renegeing on deals, they generally comply, or refuse to enter agreements that they would otherwise violate. Thus, institutions are seen as a way of reducing uncertainty and enabling otherwise risky forms of international cooperation. Again, this rather functionalist understanding sees institutions as emerging to serve the exogenous preferences of atomistic utility-maximizing actors rather than informing their initial preferences.

A final and important category in the rational-choice institutionalism literature is the veto players framework. Developed by Tsebelis (2002; but see Ganghof, 2003), the veto players framework views outcomes as a function of the number of 'veto players', or actors with the ability to block legislation, and their preferences. Veto players may be committee chairs, executives or any member of a decision-making body whose approval is required for a motion to progress. Veto players' preferences are represented spatially with 'indifference curves', or shaded circles surrounding a veto player's ideal point along a multidimensional policy space. The overlap between the necessary veto players indifference curves is deemed the 'winset', or the set of potential actions given what is acceptable to the players.

The veto players framework reveals an important dynamic of group decision-making in institutional contexts. Frequently, especially in legislatures, the winset is defined primarily by a single player known as a 'gatekeeper', who has the power to unilaterally block legislation. Thus, the likelihood of a particular motion moving forward is defined more by what this individual is either supportive or indifferent to than by the preferences of the other members.

Another important value of the veto players framework lies in its parsimony and versatility: it can be applied in essentially any institutional setting where decisions are made, and effectively limits the range of potential outcomes in a methodologically consistent fashion. Further, it can allow us to understand outcomes without a thorough understanding of every member of an organization, because it directs our focus to those members who matter most. The veto players' framework has been applied to understand the development of the rule of law in emerging democracies (Andrews and Montinola, 2004), the impact of globalization on welfare spending (Ha, 2007), the conclusion of civil war (Cunningham, 2006) and many other

problems of comparative and international politics.

### *Historical*

Historical institutionalism differs from other categories in its admission of historical-contingent factors as important aspects of institutional evolution and impact. Historical institutionalism is explicitly structuralist; that is, actors are seen as reacting to, and operating within, macro-structural constraints posed by the interaction of social, economic, political and institutional forces. In this view, history is viewed 'inefficiently', in that political outcomes cannot be seen as an efficient function of the preferences of the masses or elites. Instead, changes in technology, the sociological groupings of individuals, legal conventions and other historic factors are seen as advancing or constraining change (David, 2001).

For example, consider Theda Skocpol's (1979) classic work on revolutions. In this analysis, revolutions are seen as the product of imperial monarchic institutions' inability to rise to the financial and organizational challenges of increased international military competition. The historical evolution of international military competition is seen as having an inefficient relationship with domestic institutional development, priming domestic pressures for institutional change. Again, all of these changes are embedded in historical-contingent shifts rather than the preferences of individuals.

More recently, Steinmo et al. (1992) have made the modern case for historical institutionalism, arguing that a focus on the policy choices made when an institution is formed have a long-term influence on its policy choices into the future. While this pattern is conventionally referred to as 'path-dependency', Mahoney and Thelen (2010) and Streeck and Thelen (2005) have made the case for a historical institutionalism that rejects the 'punctuated equilibria' model of historical institutional change in which 'critical

junctions' (see Capoccia and Keleman, 2007) are seen as the singular temporal mode of reform. Instead, they argue that the distribution of interests within an institution can leverage a variety of techniques for gradual and endogenous institutional change. These contributions suggest an affinity between the rational-choice and historical schools that runs counter to their traditionally distinct ontological commitments (to individual agency and structure, respectively) (but see Katznelson and Weingast, 2005). Even with the addition of these mechanisms for gradual change, the historical institutionalists can be criticized because of a seemingly excessive emphasis on stability within institutions.

### *Empirical*

Empirical institutionalism is closely related to the old institutionalism in its focus on formal governmental structures and rules. Privileging constitutional documents over parties and interest groups, the empirical school attempts to infer the long-term empirical consequences of discrete choices in governmental design. Arend Lijphart's *Patterns of Democracy* (1999) argues that sets of institutional conditions lead to 'consensus' or 'majoritarian' democracies, each with their own consequences for policy, economy and conflict. Horowitz has been a consistent voice of dissent from Lijphart's views, arguing that electoral rules designed to foment universalistic electoral appeals offer more promise for divided societies (Horowitz, 2004). Shugart and Carey (1992) and Linz and Stepan (1996) debate the impacts of presidential versus parliamentary systems in two now foundational texts in the empirical institutional literature.

The empirical institutionalism is not, however, confined merely to studying differences between presidential and parliamentary systems. For example, one question within this approach relates to the effect of the autonomy of institutions – whether central banks or bureaucratic agencies (Laegreid

and Verhoest, 2010) – on the performance of public sector organizations. The analysis of federalism and other patterns of intergovernmental relations also falls into the domain of empirical institutionalism (Hueglin and Fenna, 2015). Even patterns of interest intermediation – corporatism, pluralism and other alternatives – could be analyzed through the lens of empirical institutionalism.

What distinguishes contemporary empirical institutionalism from the old institutionalism is the more explicit attempt to build theory. In addition, empirical institutionalism draws from other strands of institutional theory to develop expectations about the behaviors of institutions and the individuals within them. The questions may not be so different from those of the old institutionalists – after all, Woodrow Wilson advocated parliamentary government for the United States in 1885 – but the means of addressing those questions certainly have changed.

### *Discursive Institutionalism*

Discursive institutionalism argues that the internal discourses among institutional members, and institutions' discourses with their environments, provide a useful way of understanding political outcomes. In this view, the evolution and distribution of ideas is central to institutions' roles in organizing political and social life. The preeminent member of the discursive school has been Vivien Schmidt (2008, 2010), although her arguments emerge in part from the constructivist tradition as articulated by Colin Hay (2008) and Nicolas Jabko (2006). Perhaps most fundamentally, the discursive approach emphasizes the importance of ideas and arguments in defining the institution.

Unlike other institutional theories discussed, discursive institutionalism is not causally oriented in the material world. As such, it places substantially less emphasis on formal institutional structure, which is seen as a result of causally prior ideological discourses. This has important implications for the nature

of institutions: if formal rules do not entrench predictable and stable behavior, but are rather subject to evolutionary effects of changing ideological preferences, institutions may fail to serve the purposes central to other schools of institutionalism (see Panizza and Miorelli, 2013). While normative institutionalism lies closer to this discursive view, it is still distinguished from it by its focus on organizations and their capacity to influence the views of their members in an enduring manner.

Discursive institutionalism appears to allow considerably more room for the role of individuals, as they can be seen to be instrumental in the advocacy of particular ideas within institutional discourse and as important interlocutors between institutions and their environments. This is perhaps most evident in constructivist work in international relations, which appears intimately related to discursive institutionalism. For example, Finnemore and Sikkink (1998) outline a 'norm lifecycle' whereby idea entrepreneurs may utilize institutions to infuse norms on the domestic and international levels. Similarly, Schmidt (2008) views individuals and their organizations as active agents in the local dissemination of global ideas, leveraging awareness of domestic institutions (formal and cultural) to elevate particular norms, rules and practices.

## **THE METHODOLOGICAL IMPLICATIONS OF INSTITUTIONALISM**

Institutionalism is first and foremost a set of theoretical assumptions and understandings about humanity: individuals always operate within organizational contexts, and these contexts modify behavior through rules, incentives and values. But these assumptions have important implications for methods because they preclude certain approaches while privileging others.

Why is this so? A useful starting point is to consider how institutionalist assumptions

impact the appropriate level of analysis in approaching a research question. If we are curious about the growth of right-wing populist parties in recent European elections, rather than beginning by analyzing the individual level voter demographics of their supporters, we might first look at the institutional environments in which they have been successful. One could run a regression at the individual level and conclude that the average right-wing populist party supporter is white, elderly and working class. But this would only help us understand the outcome (right-wing populist victories) if we assume a purely efficient relationship between societal attitudes and political outcomes. If instead we embrace the institutional assumptions that history and political processes are inefficient and subject to modification by institutions, we must ask a different set of questions. Further, we might ask about the decline of other organizations within the party system, which has opened more niches for these parties.

At this point, the appropriate line of inquiry begins to turn on the particular brand of institutional theory one is following. If one opts for a rational-choice or empirical institutionalist account, one can continue to accept preferences for populist parties as exogenously given, and proceed to model their success as a function of popular support and electoral rules. Do right-wing populist parties perform better in electoral systems with proportional representation? Do they perform better under electoral systems with lower thresholds for participation? Do they succeed in environments with high district magnitude? One can include controls for these questions in a familiar regression analysis, and perhaps come closer to understanding this phenomenon.

One of the more exciting innovations of the new institutionalism is the incorporation of micro-level analysis into an institutional context. That is, if the answer to a particular research question is not evident at the macro level, there are now ways to model

micro-level behaviour within known institutional constraints. Pioneers of this method such as Tsebelis, Krehbiel and Weingast allowed institutional factors to influence preferences and options, enabling the modeling of micro-level behavior without ignoring formal influences. For example, Jonathan Rodden's *Hamilton's Paradox: The Promise and Peril of Fiscal Federalism* (2006) applied a rational-choice framework to understand how varying institutional environments impact the fiscal behavior of local and national governments. While these works draw their conclusions from micro-level analyses, they do not violate the assumptions of institutionalism because they incorporate institutional values and rules into individual preferences.

If, however, the analyst adopts the ontological commitment of the embedded, rather than atomistic, individual, a separate set of questions arises that are not so easily answered through quantitative means. Returning to the example of European populism, a sociological institutionalist might consider how trends of liberalization in economic and immigration policy have influenced social cleavages. Are there cultural changes that have manifested in institutional biases that elevate anti-immigrant populism? Institutionalists in this school might begin at the level of state institutions, formal and informal, seeking relationships between these characteristics and the outcome of interest.

Historical institutional analysis permits deep focus on the roles of organization, laws and culture on political outcomes. Acemoglu and Robinson (2006) apply this method to identify four ideal-type pathways to democracy (or its failure). The authors identify changes in the political and economic institutions of Britain, Argentina, Singapore and South Africa that led to democratic consolidation or failure. Qualitative comparative analysis (QCA) allows for the systematic comparison of multiple institutional features across a small or medium number of cases to identify whether particular conditions are associated with an outcome of

interest. Process-tracing (see Collier, 2011 for a review) focuses on precise causal pathways within a single case, and can reveal path-dependent outcomes that might elude other forms of analysis. A shared asset of all of these methods is their embrace of complexity and context; they allow the researcher the opportunity to compare a range of institutional features to derive whether or not they may have a relationship with an outcome that is causally prior to or endogenous with micro-level behavior.

A more generalized process of inquiry would look something like this: (1) what is the institutional context for the phenomenon under study; (2) in what ways does that context constrain or expand the probability of a given outcome? These questions tend to be difficult to answer with strictly quantitative analysis for two reasons: (1) the process of understanding institutional contexts (Peters, 2013) often requires more attention to the detail and history of individual cases than quantitative analyses typically (although not always) allow; (2) more often than not, the higher the level of analysis (that is, legislatures versus individuals), the smaller the sample, reducing the leverage of most quantitative tools.

In addition to the methodological implications of institutionalism mentioned above, Diermeier and Krehbiel (2003) argue that institutionalism is fundamentally a methodology. This argument is based on a rational-choice assumption about the nature of institutions and the manners in which behavioral assumptions shape individual and collective behavior. In this view, perhaps the most significant element of institutional theory is its capacity to guide the design of institutions to produce equilibria in well-defined circumstances.

## **CONTRADICTIONS AND PARADOXES OF INSTITUTIONAL THEORIES**

This final section of our chapter will discuss some apparent contradictions and paradoxes



within institutionalism, and within institutions themselves. These contradictions, or apparent contradictions, can be taken as continuing challenges to scholars who seek to understand institutional theory and apply it to the real world within the public sector. But they also illuminate some of the key points of institutional theory, and can be seen almost as summaries of the fundamental points of this approach to political science.

### ***Institutions Are Stable, but Institutions Change***

One of the fundamental paradoxes of institutions, and institutional theory, results from the presumed permanence of institutions, as compared to their continuing need to change and adapt. The fundamental logic of institutions is to create some order and predictability in situations that might otherwise be uncertain. The increased predictability may be a function of rules or of normative commitments to the values of the institution, but the important point is individuals functioning within an institution behave in more regular patterns than do those outside, and those regularities persist over time.

But few if any institutions can survive without change. Any institution is embedded in an environment that will generate pressures for change. These pressures may be political, functional and social (Oliver, 1992), and will require adaptation on the part of the institution if it wishes to survive. Some institutions may change so much that they lose their identity and become in essence new institutions, while most adapt more gradually and recreate their identity in reference to their changed environment.

Further, all but the most strongly institutionalized institution will have some internal conflicts that will generate change (Peters et al., 2005). These internal conflicts may produce even greater challenges to the institution than will external challenges, given that they can undermine the functioning of

the institution or lead to its disintegration. Similarly, internal dissent can lead to ongoing subversion of the purposes of the institution and reduce its performance (O'Leary, 2006; Olsson, 2016).

### ***Institutions Have Common Goals, but Often Are Internally Diverse***

One common image of an institution, at least in the popular conception of an institution, is that it should be relatively well integrated and have a common set of values that guides its actions. And some theoretical approaches to institutions, such as normative institutionalism, stress very strongly the importance of the institution creating a uniform set of internal values that will guide the behaviors of the members of the institution. Institutions develop myths, symbols, values and routines, all intended to create predictable behaviors among the members of the institution, and to make the values of members endogenous to the institution.

The problem is that few if any institutions are that uniform and predictable. For some institutions, such as the military, there is a clear need for high levels of internal compliance and for the members to march to the same drum (literally and figuratively). For other institutions, however, that need for uniformity is not so great, and may not be obtainable even if it were desired. And in some cases internal differences may be desirable, especially when the institution is engaged heavily in innovation and problem-solving, and may want to have members advancing disparate ideas (see Peters, 2019).

Theoretically, at least two approaches within the institutionalism literature speak to the degree of internal difference and debate. Within political science, discursive institutionalism (Schmidt, 2010) assumes that there will be several internal discourses within an institution. This internal debate may result in a single communicative discourse dealing with the outside world, while internally

there is some discord. Similarly, the institutional logics literature in sociology (Thornton et al., 2012) assumes that institutions contain a variety of different logics that may be more or less dominant in any one situation. Likewise, individuals within the institution may be connected more or less strongly to the various possible logics. The presence of different logics and cultures permits the institution to innovate more readily, and also to coordinate more easily with other institutions (see below).

### ***Institutions Shape Individuals, but Are Shaped by Individuals***

The relationship between individuals and institutions is also complex. On the one hand, institutions are created by individuals, but they are often created in order to constrain the autonomy of others, or even the autonomy of the individuals who created them. The obvious example is creating political institutions – the constitutions that are designed to constrain behaviors and produce certain types of outcomes (Buchanan and Tullock, 1962) in a presidentialist system, for example, are intended to make decision-making more difficult as compared to parliamentary systems.

But institutions, as human creations, can also be changed by human intervention (Grafstein, 1992). Therefore, assuming that institutions are somehow immutable or impervious to the actions of individuals tends to understate the role of agency in defining and changing institutions. And the capacity to exercise agency is more available to individuals occupying leadership positions than to ‘ordinary’ members of the institution. When those leaders disregard the norms of the institution, as President Trump has done with the presidency (and American government more generally), these actions can produce change within the institution. The institution may have some tendency to revert to the *status quo ante*,<sup>1</sup> but in the short term can be altered by leadership.

The example of the president altering an institution by his behavior reflects, to some extent, unintentional change within the institution, but the question of designing change is perhaps more interesting. For rational-choice versions of institutionalism this would appear to be relatively easy: all one needs to do is to change the rules and incentive structures and the behavior should be changed accordingly. For other versions of institutionalism, exercising individual agency in changing the institution will be more problematic, given that change may involve altering underlying beliefs or overcoming established inertia.

### ***Institutions Are About Structures, but Also About Individual Agency***

Explanations based on institutions are almost inherently structural, or depend upon inertial forces that are not necessarily personified. And given that in many versions of institutionalism the preferences of individuals are endogenous to the institution, there may appear to be little opportunity for agency. However, any institution will depend upon the actions of individuals. The danger in some versions of the general approach, such as historical institutionalism, is that individual actions are assumed and not really understood properly (but see Pierson, 2000). Further, institutional change may depend heavily on exogenous forces, notably human agency.

Therefore, any reasonable version of institutionalism must find a way to integrate individuals and the institutions that they populate. The theoretical approaches available tend to be better at explaining how institutional rules and values shape individual behavior than at explaining how individuals shape the nature and functioning of institutions. But the individuals do shape institutions, bringing with them values and behavioral patterns to which the institution may have to adapt – or which it

may attempt to change. And individuals may also bring with them ideas that they wish to use to alter the norms of the institution.

### ***Institutions May Be Organizations, but They May Be More Than That***

One of the continuing questions in institutional theory, as well as in organization theory, is how institutions and organizations differ. Or do they? Some discussions of institutions tend to use organizations virtually interchangeably with institutions. There are indeed many important similarities. Both are more or less formalized structures populated by individuals and designed to reach goals. What makes them different – or is there really any significant difference?

At one level, institutions are discussed as broad social institutions such as law, the market and the polity. At a somewhat lower level of generality, as with most other questions about institutions, the answer may depend on the approach. For example, Selznick (1957) argues that an institution is an organization that has been infused with values. Scholars coming from economics argue that institutions are broad sets of rules while organizations are actors playing by the rules (North, 1990; Khalil, 1995). And in some discussions within historical institutionalism, institutions appear to be synonymous with policy, while organizations are those entities charged with making the policy function.

### ***Institutions Address Collective Action Problems, but May Also Create them***

Especially within rational-choice institutionalism, institutions are designed to address collective action problems. Elinor Ostrom's (1990, 2005) work on the commons and common pool resources demonstrates the importance of developing institutions to address collective action problems. Other

perspectives on institutions can also be understood as addressing collective action issues. For example, in normative institutionalism, the 'logic of appropriateness' can be used as a means to have individuals pursue the collective interest of the institution, or perhaps even the society, rather than individual self interest.

But, somewhat paradoxically, the more successful an institution is in solving collective action problems within its own domain, the more likely it may be to create collective action problems across domains. A thoroughly institutionalized structure may have the same narrow goals and conception of interest, and appropriateness, as an individual pursuing their personal utility. They may therefore be unwilling to engage in cooperative behavior that may be perceived to weaken the institution's own benefits. For example, within the public sector, one institution may pursue its own goal of maximizing its power but in the process weaken the governance capacity of the system as a whole.<sup>2</sup>

### ***Institutions Are Independent, but Share Space with Other Institutions***

Much of the literature on institutions tends to discuss institutions as largely independent, autonomous actors. We have already argued that institutions are influenced by their individual members, but they are also influenced by their environment. The environment of institutions is composed largely of other institutions and organizations. For example, the market, as an institution defined by economic principles, is shaped or even defined by other institutions, such as law and the political system. The market and market actors may be powerful, but they cannot avoid the influence of other institutional actors. These institutional and organizational fields mutually constrain the actions of all the participants, and can also serve as mechanisms for monitoring and accountability.

The constant interactions among institutions lead to isomorphic pressures (DiMaggio and Powell, 1983). The interactions among institutions may be facilitated if there is a common pattern of organization among them, and if they, to some extent, have some of the same institutional logics. As a simple example, market organizations tend to organize themselves on the basis of states and provinces in federal systems; committee systems in legislative organizations tend to mirror the structure of the executive branch. These are simple examples, but the basic process of replication of structural formats and process across institutions can be observed in more complex settings as well.

***An Institution Is an Institution, but May Be More or Less So (Institutionalization)***

Most of the literature on institutions treats institutions as a dichotomous variable: an institution exists or it does not. Thus, no matter how shaky an institution may be – see, for example, the presidency in the last days of Richard Nixon’s time in office – it is still an institution. While this simple dichotomy is to some extent true and it does simplify the analysis of institutions, it also wastes a great deal of information. Institutions are not always stable and are not all equally well organized, and therefore we can gain a great deal of analytic leverage by attempting to understand how well structured the institution is in practice.

The concept of *institutionalization* can be used to address the differential levels of structuring within the institution. The various versions of institutionalism rather naturally have different conceptions of institutionalization. For example, for normative conceptions of institutions, Selznick (1957) discussed institutionalization in terms of infusing structures with values and meaning, rather than depending upon mere mechanical achievement of tasks. Empirical studies of institutions, such

as that by Ragsdale and Theis (1997), have developed a number of measures that can provide some basis for comparing levels of institutionalization.

The causes and processes of institutionalization and deinstitutionalization are also important for understanding how institutions are formed and then wax and wane. Oliver (1992) has discussed the processes of deinstitutionalization, but if these processes are run in the other direction they can also help to explain institutionalization. She argues that deinstitutionalization occurs because of three types of pressures on the institution: functional, social and political. Further, those pressures arise both from within the institution and from the environment. Thus, institutions are constantly facing changing pressures that alter their internal capacity and their capacity to influence their environment.

***Multiple Approaches, but One Institutionalism***

Perhaps the ultimate paradox, or question, in contemporary institutional theory is whether there is one institutionalism or just a set of rather disparate approaches to institutions. To what extent is there any capacity for institutionalism to function as a paradigm for understanding political life – or is it just a set of approaches to institutions that do not mesh together in any meaningful way? In either case the institutionalism literature can be valuable, but each alternative serves different purposes and will make different contributions to the social sciences.

We can make a case that there is one institutionalism. This perspective is based on the existence of a number of points of similarity among the various approaches discussed above. The most obvious point of similarity is that all approaches focus on institutions, or more generally tend to emphasize structure over agency in explaining outcomes of the political process. Although some approaches, such as that of the empirical institutionalists,

use structure in an almost commonsense manner and other approaches use it more analytically, they all depend upon structure (see Easton, 1990).

Second, all the approaches to institutions represent attempts to reduce the amount of variance in the behavior of individuals who are members of an institution. This reduction in variance can be achieved through inculcating values, the creation of rules, the use of internal discourses or simply inertia, but all approaches to institutions attempt to explain reductions in variability of behaviors. The reduction in variance does not necessarily imply that institutions are involved heavily in controlling their members, given that many members join an institution knowing what the expectations are, and are thus prepared to comply.

Third, all versions of institutions (and organizations) have some notion of boundary maintenance. Some actors are inside the institution and others are outside, and although the environment of the institution is important for shaping and constraining its behavior, maintaining some differentiation from that environment is important for institutions. Boundary maintenance for the institution is used in part to identify levels of commitment of members and potential members, as well as a means of controlling resources.

Finally, all forms of institutionalism are faced with problems of change. The logic of institutions is to create stability, but excessively stable institutions are likely to be unsuccessful over time, as their environments and the nature of the personnel being recruited change. Although there are more robust ideas about institution change now than there were in the earlier days of the new institutionalism (see Mahoney and Thelen, 2010), change still presents problems for this body of theory. The *a priori* assumption for institutionalism appears to be that change is the exception, while in the real world of politics and governing, change appears endemic.

These similarities among the approaches are important, but in some people's minds

they may be outweighed by the differences among the approaches. Whether the glass is half full or half empty is to some extent a perceptual question, but the way in which institutionalism is perceived – both by its advocates and its critics – will affect its capacity to have broader influence in the discipline. The tendency to emphasize differences among the approaches<sup>3</sup> may have undermined the ability to identify, and to utilize, the similarities.

## CONCLUSION

Institutionalism is a major approach within contemporary political science. Although there are numerous approaches to the concept of an institution, and therefore multiple mechanisms through which institutions are argued to influence politics and governance, there is a common core that argues that institutions are important. Within the approach there is broad agreement that the structures within the political system, both formal and informal, are crucial in shaping the actions of individuals.

Like all approaches within political science, institutional theory continues to evolve. Part of that evolution is the tendency to blend some of the approaches, and to make the differences among them less stark. Rational-choice theories have taken on some of the ideas of the normative approach, and have increasingly argued that social norms are significant in, for example, overcoming collective action problems. And historical institutionalism has now long abandoned its emphasis on radical change by accepting more gradual forms that resemble other types of political change. It is too soon to tell whether these various approaches will in the long run evolve into a single, more powerful approach, but there are some hopeful signs.

## Notes

- 1 This raises an interesting point about historical institutionalism. If the path on which the

institution, or policy, is traveling is sufficiently strong to reassert itself after some changes, can we say that the assumptions of historical institutionalism are supported?

- 2 Conflicts between presidents and legislatures in presidential systems are an obvious example of this phenomenon.
- 3 One of the authors of this article must plead guilty to this charge.

## REFERENCES

- Acemoglu, Daron and James A. Robinson (2006) *Economic Origins of Dictatorship and Democracy*. (Cambridge, UK: Cambridge University Press).
- Andrews, Josephine T. and Gabriella R. Montinola (2004) 'Veto Players and the Rule of Law in Emerging Democracies', *Comparative Political Studies* 37(1), 55–87.
- Aristotle. Translated by C.D.C. Reeve (1998) *Politics* (Indianapolis: Hackett Publishing Company).
- Barnard, Chester (1968) *The Functions of the Executive* (Cambridge, MA: Harvard University Press).
- Buchanan, James M. and Gordon Tullock (1962) *The Calculus of Consent* (Ann Arbor: University of Michigan Press).
- Campbell, Angus, Philip E. Converse, Warren E. Miller and Donald E. Stokes (1976) *The American Voter* (Chicago: University of Chicago Press).
- Capocchia, Giovanni and R. Daniel Keleman (2007) 'The Study of Critical Junctures: Theory, Narrative, and Counterfactuals in Historical Institutionalism', *World Politics* 59(3), 341–69.
- Carpenter, Daniel and George A. Krause (2015) 'Transactional Authority and Bureaucratic Politics', *Journal of Public Administration Research and Theory* 25(1), 5–25.
- Collier, David (2011) 'Understanding Process Tracing', *PS: Political Science and Politics* 44(4), 823–30.
- Cunningham, David E. (2006) 'Veto Players and Civil War Duration', *American Journal of Political Science* 50(4), 875–92.
- David, P. (2001) 'Path Dependence, Its Critics and the Quest for "Historical Economics"', Ch. 7 in P. Garrouste and S. Ioannides, eds, *Evolution and Path Dependence in Economic Ideas* (Cheltenham: Edward Elgar).
- Diermeier, Daniel and Keith Krehbiel (2003) 'Institutionalism as a Methodology', *Journal of Theoretical Politics* 15(2), 123–44.
- Dimaggio, Paul J. and Powell, Walter W. (1983) 'The Iron Cage Revisited: Institutional Isomorphism and Collective Rationality in Organizational Fields', *American Sociological Review* 48(2), 147–60.
- Durant, R. F. (2006) 'Agency Evolution, New Institutionalism, and "Hybrid" Policy Domains: Lessons from the "Greening" of the US Military', *Policy Studies Journal* 34(4), 469–90.
- Durkheim, Émile. (1982 [1895]). *The Rules of Sociological Method*. (S. Lukes, Ed.; W. D. Halls, Trans.) (New York: The Free Press).
- Durkheim, Émile, Sarah A. Solovay, John H. Mueller and George E. G. Catlin (1950) *The Rules of Sociological Method* (Glencoe, IL: Free Press).
- Easton, David (1990) *The Analysis of Political Structure* (New York: Routledge).
- Evans, Peter B. Dietrich Rueschemeyer and Theda Skocpol, eds. (1985) *Bringing the State Back In* (Cambridge: Cambridge University Press).
- Finnemore, Martha and Kathryn Sikkink (1998) 'International Norm Dynamics and Political Change', *International Organization* 52(4), 887–917.
- Fish, M. Steven (1998) 'Democratization's Requisites: The Postcommunist Experience', *Post-Soviet Affairs* 14(3), 212–47.
- Ganghof, Steffen. (2003) 'Promises and Pitfalls of Veto Player Analysis', *Swiss Political Science Review* 9(2), 1–26.
- Grafstein, Robert (1992) *Institutional Realism* (New Haven, CT: Yale University Press).
- Greif, A. (2006) *Institutions and the Path to the Modern Economy: Lessons from Medieval Trade* (Cambridge: Cambridge University Press).
- Ha, Eunyoung (2007) 'Globalization, Veto Players, and Welfare Spending', *Comparative Political Studies* 41(6), 783–813.
- Hamilton, Alexander, James Madison and John Jay (2009) *The Federalist Papers* (New Haven, CT: Yale University Press).
- Hay, C. (2008) 'Constructivist Institutionalism', (56–74) in S. A. Binder, R. A. W. Rhodes and B. A. Rockman, eds, *Oxford Handbook of Political Institutions* (Oxford: Oxford University Press).

- Herbst, Jeffrey (2000) *States and Power in Africa: Comparative Lessons in Authority and Control* (Princeton, NJ: Princeton University Press).
- Horowitz, Donald L. (2004) *Ethnic Groups in Conflict: Theories, Patterns, and Policies* (Los Angeles: University of California Press).
- Hueglin, Thomas O. and Alan Fenna (2015) *Comparative Federalism: A Systematic Inquiry*, Second edition (Toronto, CA: University of Toronto Press).
- Jabko, N. (2006) *Playing the Market: A Political Strategy for Uniting Europe, 1985–2005* (Ithaca, NY: Cornell University Press).
- Johnson, Chalmers. (1982) *MITI and the Japanese Miracle: The Growth of Industrial Policy, 1925–1975*, (Cambridge, MA: MIT University Press).
- Katznelson, I. and Barry R. Weingast (2005) *Preferences and Situations: Points of Intersection Between Historical and Rational Choice Institutionalism* (New York: Russell Sage).
- Keohane, Robert O. (2005) *After Hegemony: Cooperation and Discord in the World Political Economy* (Princeton, NJ: Princeton University Press).
- Khalil, Elias L. (1995) 'Organizations versus Institutions', *Journal of Institutional and Theoretical Economics* 151(3), 445–66.
- Kiser, Larry L. and Elinor Ostrom. 1982. 'The Three Worlds of Action: A Metatheoretical Synthesis of Institutional Approaches', in *Strategies of Political Inquiry*, ed. Elinor Ostrom (Beverly Hills, CA: Sage), 179–222.
- Krepel, Amie. (2002) *The European Parliament and Supranational Party System: A Study in Institutional Development* (Cambridge, UK: Cambridge University Press).
- Laegreid, Per and Koen Verhoest, eds. (2010) *Governance of Public Sector Organizations: Proliferation, Autonomy and Performance* (Basingstoke: Macmillan).
- La Porta, Rafael, Florencio Lopez de Silanes and Andrei Shleifer (2008) 'The Economic Consequences of Legal Origins', *Journal of Economic Literature* 46(2), 285–332.
- Lijphart, Arend (1999) *Patterns of Democracy: Government Forms and Performance in Thirty-Six Countries* (New Haven, CT: Yale University Press).
- Linz, Juan J. and Alfred Stepan (1996) *Problems of Democratic Transition and Consolidation* (Baltimore, MD: Johns Hopkins University Press).
- Mahoney, J. and K. Thelen (2010) 'A Theory of Gradual Institutional Change', Chapter 1 (1–38) in James Mahoney and Kathleen Thelen, eds, *Explaining Institutional Change: Ambiguity, Agency and Power* (Cambridge: Cambridge University Press).
- March, James G. and Johan P. (1984) 'The New Institutionalism: Organizational Factors in Political Life', *American Political Science Review* 78(3), 738–49.
- March, J. G. and J. P. Olsen (1989) *Rediscovering Institutions* (New York: Free Press).
- McCubbins, Matthew D., Roger G. Noll and Barry R. Weingast (1989) 'Structure and Process, Politics and Policy: Administrative Arrangements and the Political Control of Agencies,' *Virginia Law Review* 75(2), 431–82.
- Miller, Gary J. (2005) 'The Political Evolution of Principal-Agent Models', *Annual Review of Political Science* 8, 203–25.
- Nölke, Andreas and Arjan Vliegenthart (2009) 'Enlarging the Varieties of Capitalism: The Emergence of Dependent Market Economies in East Central Europe', *World Politics* 61(4), 670–702.
- North, Douglass C. (1990) *Institutions, Institutional Change, and Economic Performance* (Cambridge: Cambridge University Press).
- North, Douglass C. and Barry R. Weingast. (1989) 'Constitutions and Commitment: The Evolution of Institutions Governing Public Choice in Seventeenth-Century England.' *Journal of Economic History* 49(4), 803–832.
- O'Leary, Rosemary (2006) *The Ethics of Dissent: Managing Guerilla Government* (Washington, DC: CQ Press).
- Oliver, Christine (1992) 'The Antecedents of Deinstitutionalization', *Organization Studies* 13(4), 563–88.
- Olsson, Jan (2016) *Subversion in Institutional Change and Stability: A Neglected Mechanism* (London: Palgrave Macmillan).
- Ostrom, E. (1990) *Governing the Commons: The Evolution of Institutions of Collective Action* (Cambridge: Cambridge University Press).
- Ostrom, Elinor (2005) *Understanding Institutional Diversity* (Princeton, NJ: Princeton University Press).
- Ostrom, Elinor (2011) 'Background on the Institutional Analysis and Development Framework', *Policy Studies Journal* 39(1), 7–27.

- Ostrom, Elinor (2014) 'Collective Action and the Evolution of Social Norms', *Journal of Natural Resources Policy Research* 6(4), 235–52.
- Panizza, F. and R. Miorelli (2013) 'Taking Discourse Seriously: Discursive Institutionalism and Post-Structuralist Discourse Theory', *Political Studies* 61(2), 301–18.
- Peters, B. Guy (1996) 'Political Institutions, Old and New', Chapter 7 (205–23) in Robert E. Goodin and Hans-Dieter Klingemann, eds. *A New Handbook of Political Science* (Oxford: Oxford University Press).
- Peters, B. Guy (2013) 'Institutions in Context and as Context', in C. Pollitt, ed., *Context in Public Policy and Management: The Missing Link?* (Cheltenham: Edward Elgar), 101–14.
- Peters, B. Guy (2019) 'Designing Institutions for Designing Policy', *Policy and Politics* 47(4), 1–17.
- Peters, B. Guy, Jon Pierre and Desmond S. King (2005) 'The Politics of Path Dependency: Political Conflict in Historical Institutionalism', *Journal of Politics* 67(4), 1275–1300.
- Plato. Translated by G. M. A. Grube (1992) *Republic* (Indianapolis: Hackett Publishing Company).
- Ragsdale, Lyn and John J. Theis III (1997) 'The Institutionalization of the American Presidency 1924–92', *American Journal of Political Science* 41(4), 1280–1318.
- Rodden, Jonathan (2006) *Hamilton's Paradox: The Promise and Peril of Fiscal Federalism* (Cambridge: Cambridge University Press).
- Saiegh, S. M. (2016) 'Lawmaking', Chapter 23 in S. Martin, T. Saalfeld and K. Strøm, eds, *Oxford Handbook of Legislative Studies* (Oxford: Oxford University Press).
- Savoie, Donald J. (1990) 'Reforming the Expenditure Budget Process: The Canadian Experience', *Public Budgeting & Finance* 10(3), 63–78.
- Schmidt, Vivien A. (2008) 'Discursive Institutionalism: The Explanatory Power of Ideas and Discourse', *Annual Review of Political Science* 11, 303–26.
- Schmidt, Vivien A. (2010) 'Taking Ideas and Discourses Seriously: Explaining Change Through Discursive Institutionalism and the Fourth New Institutionalism', *European Political Science Review* 2(1), 1–25.
- Selznick, Philip (1943) 'An Approach to a Theory of Bureaucracy', *American Sociological Review* 8(1): 47–54.
- Selznick, P. (1957) *Leadership in Administration* (New York: Harper and Row).
- Shepsle, K. A. (1989) 'Studying Institutions: Lessons from the Rational Choice Approach', *Journal of Theoretical Politics* 1(2), 131–47.
- Shugart, Matthew Soberg and John M. Carey. (1992) *Presidents and Assemblies: Constitutional Design and Electoral Dynamics* (Cambridge: Cambridge University Press).
- Skocpol, Theda (1979) *States and Social Revolutions: A Comparative Analysis of France, Russia, and China*. Cambridge: Cambridge University Press.
- Steinmo, S., K. Thelen and F. Longstreth, eds. (1992) *Structuring Politics: Historical Institutionalism in Comparative Analysis* (Cambridge: Cambridge University Press).
- Streeck, Wolfgang and Kathleen Thelen, eds. (2005) *Beyond Continuity: Institutional Change in Advanced Political Economies* (Oxford: Oxford University Press).
- Thornton, P. H., W. Ocasio and M. Lounsbury (2012) *The Institutional Logics Perspective: A New Approach to Culture, Structure and Process* (Oxford: Oxford University Press).
- Tsebelis, George (2002) *Veto Players: How Political Institutions Work* (Princeton: Princeton University Press).
- Tsebelis, George and Jeannette Money (1997) *Bicameralism* (Cambridge: Cambridge University Press).
- Weber, Max (1958) *The Protestant Ethic and the Spirit of Capitalism* (New York: Charles Scribner's Sons).
- Weber, Max. (1994) *Political Writings*. Eds. Lassman, P. Speirs, R. (UK: Cambridge University Press).
- Wildavsky, Aaron (1992) 'Political Implications of Budget Reform: A Retrospective', *Public Administration Review*, 52(6), 594–9.
- Wilson, Woodrow (1887) 'The Study of Administration', *Political Science Quarterly* 2(2): 197–222.
- Wilson, Woodrow (1892) *The State: Elements of Historical and Practical Politics* (Boston: D. C. Heath & Co.).
- Woolsey, Theodore Dwight (1877) *Political Science, or, The State Theoretically and Practically Considered* (London: Sampson Low, Marston, Searle, & Rivington).





# Applied Game Theory: An Overview and First Thoughts on the Use of Game Theoretic Tools

Adam Meirowitz and Kristopher W. Ramsay

In this chapter we put forward a framework for using game theory to study politics. In our view, the rigorous analysis of strategic interactions can make a number of contributions to our collective understanding of political phenomena. Game theory can be used to develop theoretical insights into how the structure of an interaction, the preferences of actors, the information available to different actors and some idea of purposeful choice combine to create forces that push politics toward certain outcomes, sometimes desirable and sometimes tragic. Game theory connects these fundamental elements of the political world through equilibrium (or steady state) analysis. This approach to theory building, at its core, starts from the premise that what happens in political interactions is fundamentally a consequence of the interdependence of actors' choices, their objectives and the constraints they face.

Game theory clearly is a useful tool for theory building. Since at least von Neumann and Morgenstern (1944), scholars and analysts

have built game theoretic models to study the general properties of contests, electoral competition, bargaining and nuclear deterrence, to name a few topics. This approach to applying game theory to the study of politics, while substantively motivated, tends to be general in both its mathematics and its predictions about strategic behavior, but not primarily empirical. A typical claim might be that deterrence success requires a credible threat or that bargaining favors the patient. The primary purpose of such modeling is to identify general mechanisms and forces in politics or explore the logical foundations related to a specific class of problems. The output of this work is not intended to be tested, and a test may not even make sense.

But game theory can also be used with more empirical objectives in mind. The scientific development of knowledge relating to a subject does not just depend on general theories or stylized analysis emphasizing important trade-offs and incentives; it also consists

of building a library of empirically informed models. These models are specific representations of concrete political domains, such as crisis bargaining between countries, environmental policy making or coordination in elections, that incorporate elements understood to be relevant from general theoretical analysis. Each of these domain specific models emphasizes different aspects of the political context – such as who they take to be important actors, what choices are available to players, the nature of actors' preferences and the existing information environment – and then compete to explain phenomena, often with different implications for our understanding of causation and the consequences of policy intervention. Often these models are most valuable for their predictions linking changes in fundamental characteristics of an environment, such as military power, preferences or information, to outcomes through the process of strategic interaction. The empirical plausibility of such models is often explored through statistical tests, derived from equilibrium comparative statics, linking observable characteristics of the 'real world' to changes in outcomes or choices. In this exercise, data and model are distinct objects and the value of the model is assessed based not only on the novelty or elegance of the analysis itself, but also on the quality of the tests to which it can be subjected.

Sometimes, and with increasing frequency, similar domain specific models are empirical in a completely different way. In these circumstances researchers look to quantify the theoretical model directly using observed behavior and data to estimate fundamental, and unobservable (or unobserved), theoretical quantities through the actors' revealed preferences. Here the model and equilibrium concept acts as a set of identifying assumptions specifying what observed actions must imply about the preferences, information, actions or values of important theoretical variables unobservable to the analyst *if* the modeled strategic interaction is producing the observed data. This quantifying exercise

can rely on estimation and statistical inference or calibration. The result allows both for forecasting and for the evaluation of theoretically motivated and empirically consistent counter-factual analysis.

With a randomized controlled trial, careful thought about research design prior to implementation and empirical analysis is highly recommended. Similarly, considering what kind of game theoretical analysis is best depends on its purpose, and consideration of these issues prior to analysis can be quite valuable. In what follows we flesh out this framework for thinking about applying game theory in political science and provide examples in the literature. As we are not intending to write a review, our selection of examples is incomplete and influenced by the areas in which we work.<sup>1</sup>

Before jumping into particulars, it is useful to summarize the implications of our view of the role of game theory in political science. Game theory is a valuable tool for synthesizing the effects of structure (such as institutional rules), preferences, beliefs, information, learning, dynamics and strategy into one coherent theoretical framework. It is also very flexible and permits the inclusion of conventional notions of rationality as well as a broad range of models of behavioral decision-making in these environments. Given its flexibility, there are naturally multiple ways in which game theoretic models and analysis can be useful for the scientific development of political science and, because of the breadth of potential uses, the criteria by which one evaluates the contribution of a particular piece of research varies by its targeted contribution. In some cases the choice to develop a stylized and seemingly narrow model is well justified and in other cases generating conclusions from such a model is deeply problematic. So a reasonable assessment of a piece of work requires the reader to think about how the modeling tool is being used in a given case.<sup>2</sup>

Our chapter proceeds first by briefly reviewing some key concepts and elements

of game theory. Here we do not intend to teach the reader the technical details of game theory, but rather to be clear about the terms we will use later in the chapter.<sup>3</sup> Second, we will discuss the use of game theory for building theory *per se*. Here we will develop a simple illustrative bargaining example and review some examples of this kind of theoretical work in the literature highlighting their contributions. Next we will turn to what is really the bread and butter of the modern application of game theory to political science: modeling. Again we will turn to our running bargaining example to think about how a model is used differently, review some examples of this kind of work, and point to their contributions and how the growing library of strategic models is advancing our understanding of politics. Last, we will turn to recent work that aims to quantify game theoretic models for purposes of ‘measuring’ important theoretical quantities, for counterfactual analysis and for theoretically based and empirically consistent prediction.<sup>4</sup> In this area published work is more sparse, but a substantial literature is circulating and in development.

## THE LANGUAGE OF GAME THEORY

Before we begin talking about how game theory has and can be used to study politics, we need to have a common language to discuss these models. We need to describe what elements constitute a particular model, what general assumptions are made about actors and behaviors, and how the analysis of this environment leads to explanations and predictions of political and behavioral outcomes.

A non-cooperative game consists of a set of players, a set of strategies available to each player, a structure or game form that links strategy profiles to outcomes, an information structure, and utility functions that represent the preferences of players over outcomes of the game.

For example, we could think about the strategic problem of bargaining between two states over a territory or prize that may be divided. For convenience call the states *A* and *B* and treat the prize as a unit resource,  $[0, 1]$ . A split  $x$  provides  $x$  to *A* and  $1 - x$  to *B*. The payoff to each state from their share is  $u_A(x)$  and  $u_B(1 - x)$ . Here we could imagine that offers are made back and forth, each state taking turns until someone makes a proposal which the other agrees to accept, where neither state enjoys spending time bargaining. This description of possible actions, their timing and their consequences describes a structure. States may be uncertain about each other’s patience. Here beliefs, influenced by whatever information each side of the negotiation has, will also affect their assessment of the choice they should make. And bargaining actions will potentially reveal information about players’ patience.

Given a non-cooperative game, how do we reason about how the agents play the game? The natural starting point is to assume that agents play Nash equilibrium or one of its many well-studied refinements. For example, in games of multiple periods with complete and perfect information, it is standard to focus on subgame perfect Nash equilibria which require that agent choices are based on expectation that agents would behave rationally at choice nodes that aren’t actually reached. In static games in which players possess private information it is standard to focus on Bayesian Nash equilibria which require mutual best responses for each type of each player when agents take expectation over the attributes of other players. In games with imperfect information and sequential choices perfect Bayesian Nash, sequential or some further refined equilibrium concepts which require that agents reasonably form beliefs about what they don’t know based on conjectures of strategies and the things they do observe are typically employed. In general these concepts prune away Nash equilibria that involve choices which are justified by at least one player anticipating actions that

would turn out to be difficult to justify as rational given plausible assessments of how others are playing the game. Equilibrium then becomes the link between what we observe and what that implies about unobservable elements of the political environment.

**USES OF GAMES**

With the preliminaries out of the way, we next consider the different uses of game theory in political science.

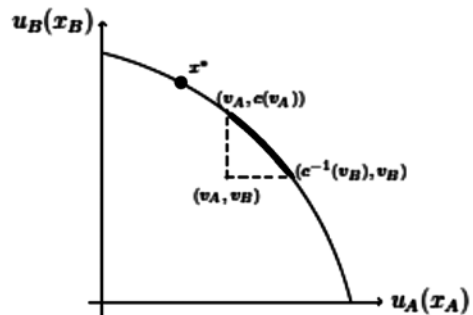
**Theory Development**

One use of game theory is theory development. What do we mean by theory development? This kind of work focuses on understanding an abstract representation of a class of problems where the focus is on mechanisms and logic, but not empirics. This kind of general theoretical development can take a number of forms. One focuses on creating stylized games that capture some intuition about a general strategic problem, and tells a story. The other attempts to find general results that apply to a very broad set of circumstances and generates theorems that can be used in future work.

Let us return to the example of bargaining between states.<sup>5</sup> The world is complicated, and models by design are less complicated; they are simplifications that ignore some, if not most, details. Why should scholars bother developing and analyzing a simplification? The first order answer is to clarify intuition. Much like how an experimentalist designs an experiment to isolate the consequences of a specific intervention or treatment, such a model looks to isolate a set of incentives and equilibrium forces in a simplified setting to better understand how some specific mechanism or set of mechanisms influence behavior and outcomes.

For example, consider any bargaining model that has two potential outcomes, either

a successful division of the pie or bargaining failure, and ask: how does the settlement that is obtained relate to what happens if the states do not reach an agreement on the split? Let  $v_A, v_B$  denote the payoffs that state A and B would receive if either player walks away from the bargaining table. Here we don't assume any specific bargaining protocol but we assume that minimally, if a bargain is reached, it is on the payoff frontier. That is, our bargaining game does not waste resources in the process of dividing them. This represents a bargaining situation where we imagine that prior to accepting a split,  $x$ , and terminating the game, each player has the option to walk away and receive their reservation payoffs, and end the game with  $v_A, v_B$ . It is not difficult to see that in any Nash equilibrium that ends with a split  $x$  we must have  $u_A(x) \geq v_A$  and  $u_B(1-x) \geq v_B$ . Consider the bargaining problem illustrated in Figure 11.1. The arc,  $c(x)$ , represents the frontier of bargaining solutions in utility space for states A and B. Suppose that the bargaining process were to produce an outcome, *without* the outside option, of  $x^*$ . Now suppose that the point  $(v_A, v_B)$  is the walk-away value for states A and B. Clearly A would walk away from  $x^*$ . In fact, we know that in this case states A and B would only both agree to a bargain that is on the arc  $c(v_A)$  between points  $(v_A, c(v_A))$  and  $(c^{-1}(v_B), v_B)$ . All these options are preferred to walking away, and which would materialize



**Figure 11.1** The bargaining problem with unilateral outside options

depends on the bargaining interaction, but we know that an efficient bargaining protocol would have to pick one such point. We thus see a strong dependence and general relationship between utility from an acceptable agreement and the players' values of walking away. Outside options matter. Generally, the better a state does from not reaching an agreement the better she must do in any voluntary settlement.

Importantly, the main contribution of this analysis is not really something to be tested. We have an example of a general finding that makes no specific prediction about how exactly the strategic interactions will end without a more detailed model of the bargaining process. Recognizing the importance of outside options is, however, clearly valuable.

### ***Breadth and Limits of an Intuition***

Not all theoretical contributions seek to develop an intuition. Some seek to say what does and does not happen. For example, in international relations it has become standard to build on our bargaining example to develop models of crisis bargaining. Suppose that we delve deeper into the outside option. Imagine that if A and B do not agree to a split then they will fight a war. Let  $v_A$ ,  $v_B$  represent the payoffs to A and B from fighting a war. One natural idea might be that  $v_A = p - k$  and  $v_B = (1 - p) - c$  where  $p$  is the probability that A wins the war and  $c$  and  $k$  are the costs to A and B from fighting. A researcher might seek to understand how uncertainty, in particular private information, influences the odds of war. In this model, war is equivalent to bargaining failure (Fearon, 1995).

A starting point might be the assumption that one of the players is unsure about the other's cost. For example, suppose that we assume that both players know  $k$  but only B knows  $c$ . The standard approach then is to assume that A treats  $c$  as a random variable with some known distribution. It might be convenient to assume a particular functional

form for this distribution, say the uniform on  $[0, 1]$ . What structure governs the interaction of A and B when they negotiate? In the language of game theory, what is the extensive form game? A scholar may seek to get the ball rolling by assuming something simple like the ultimatum (take it or leave it) protocol. Suppose then that we assume that A can make an offer. Then B either accepts or rejects it and the game ends with consumption of the offer or war. Finding equilibria of such a game is direct. Sequential rationality requires that B accept any offer with  $1 - x \geq 1 - p - c$ . So B accepts if  $c \geq x - p$ . Given this, the payoff to A from an offer,  $x$ , that has a chance of being rejected is given by  $x(1 - x + p) + (p - k)(x - p)$ . Differentiating and solving the first order condition, we find that A's optimal offer is  $x = p + \frac{1 - k}{2}$  (assuming that the equilibrium involves a risk of war – otherwise the offer is  $x = p$ ). Importantly, in an interior equilibrium, war happens with positive probability. It is tempting to conclude then that uncertainty (at least of this form) can cause war.

One could and perhaps should also ask: how general is this or any particular intuition? There were many specific assumptions in the above paragraph's worth of analysis. It is natural to focus on the assumptions that show up in the formulas: linearity of utility in shares; one-sided private information given by a uniform distribution. Relaxing these assumptions is not particularly taxing. For example, relaxing the assumption that  $c$  is uniform to the case of some smooth distribution function,  $F$  yields an implicit characterization of the equilibrium offer that is no more informative than the solution above, and the same basic insights come out of it.

But there was another important assumption: that player A makes an offer and B decides whether to accept or reject. Does this matter? How do we check? While the relevance of any particular functional forms can be tested by considering different functional forms or more generalized specifications,

testing and relaxing assumptions about the game form and informational assumptions can be more challenging. Here, one could select a particular bargaining protocol and re-solve the game. This approach could be repeated ad infinitum. An alternative approach is demonstrated in Banks (1990), and more recently in Fey and Ramsay (2011), drawing on mechanism design, the approach that has become standard in areas of auction theory and contract theory. Instead of solving for equilibria to a particular game, the analyst seeks to characterize properties of all equilibria to a well-defined class of games. In this setting we may ask about equilibria to any games in which each state has the ability to veto a deal in favor of war. Although the equilibria to the take it or leave it game involves a risk of war, there are games that achieve peace for certain.

Consider a game that offers A the share  $p$  and B the share  $1 - p$  and gives each player the choice to accept the split or veto it in favor of war.<sup>6</sup> A gains  $k$  compared to her war payoff and every type of B obtains a higher payoff here than from war (the strongest type,  $c = 0$ , is indifferent). So in this game form war does not occur in equilibrium. But this game still has the same private information that is typically described as the cause of war. What is learned? Among other things, by sticking with a particular game it is possible to misattribute the risk of war to private information. This identification is false. By varying the game, one learns that war occurred above not just because there is private information, but because of a conjunction of private information and the fact that the game form gave the uninformed player an opportunity to extract rents from commitment power.

Moving to two-sided private information, Fey and Ramsay (2011) are able to use the mechanism design approach to identify the aspects of private information that do and do not necessitate a risk of war in any equilibrium to a large class of games. The key is whether knowing about the other state's type

is relevant for predicting one's own war payoff. If it is then the problem is said to exhibit interdependent values and the kind of game exhibited in the last paragraph cannot prevent the occurrence of war.<sup>7</sup>

Taking a step back then, simple, explicit, finely detailed models have value and general treatments have value. The former can be especially useful when trying to provide clarity and insight about possible intuitions or relationships. But conclusions about what is and is not possible can be misguided when they derive from the particulars of a model, and if one only analyzes a particular model it can be hard to determine what particulars drive which conclusions. For conclusions about what must or cannot obtain, approaches that cover larger classes of problems are needed.

### ***Testing Consistency of Explanations***

Another type of analysis that is largely aimed at theory building is one that looks to put existing conventional arguments on solid rigorous foundations. Most sciences start with intuition and observation, but eventually apparent contradictions emerge and it becomes necessary to become more rigorous when making arguments. Some arguments turn out to have solid foundations, some are false and some are incomplete. Formal theoretical work contributes by determining which is which.

Take the example of mutual optimism as a cause of war. For quite some time, going back at least to Blainey (1988), scholars have argued that two states can agree to fight a costly war because their private information about the probability of victory makes them both optimistic about a war's favorable outcome. On the face of it, such an argument makes sense, but upon more careful consideration the link between mutual optimism and war is quite complicated. First, formalizing the most straightforward argument,

Fey and Ramsay (2007) show that if countries must agree to fight, that is, have mutual optimism and mutual agreement to forego an efficient settlement and proceed to war for war to occur, then there is no war with rational or even boundedly rational decision-makers.<sup>8</sup> Slantchev and Tarar (2011) respond by showing that when there is bargaining with one-sided incomplete information, mutual optimism can be both necessary and sufficient for war. Fey and Ramsay (2016) show that with two-sided uncertainty and bargaining, mutual optimism, by many possible definitions, is neither necessary or sufficient for war, yet in that environment war is extremely prevalent. Smith and Stam (2004) consider a model with subjective prior beliefs that do not originate from private information and show that in some cases mutual optimism returns as a cause of war. Bilal et al. (2017) show that extreme divergence in subjective beliefs brings us back to peace.

Another intuition that surfaces in conflict studies pertains to the necessity of fighting for learning. Wagner (2000) argued that because talk is cheap, it is only when disputants meet on the battlefield and see each other's capabilities that learning can occur. Wagner draws a sharp distinction between the type of learning in bargaining models that feature in economics and the type of learning that happens in conflict. In the former, bargaining costs allow for screening and in the latter, a key feature is that information is revealed on the battlefield. Powell (2004) develops a model that allows for direct learning while fighting and shows that in equilibrium there is still strategic information transmission from screening, that is, the standard channel survives. The basic insight which cuts against Wagner's seemingly compelling intuition can be seen by a very minor change to our running example. We walk through this exercise to show how a minimal change can shed light on a very different question. Although this model shares some similarities with Powell's, it is purposefully starker and simpler to analyze.

Return to our setting in which both players know  $k$  but only B knows  $c$ . Again the common prior is that A views  $c$  drawn from the uniform on  $[0, 1]$ . Because our analysis is focused on challenging a logic and not developing a model that has all of the relevant moving parts, there is little cost to just assuming that  $k = 0$ . The key change that we need to make is to assume that the game starts with A and B fighting. In addition we want to allow for war termination to be endogenous, and so there needs to be more than one possible period in which fighting occurs. A simple way to accomplish this is to assume that the game lasts for 3 periods. In period 1, A gets to make an offer,  $x^1$ . B can then accept ending the war, in which case the payoffs are  $p + 2x^1$  and  $1 - p - c + 2(1 - x^1)$  respectively, or reject. Following rejection another period of fighting happens and A once again makes an offer,  $x^2$ . B can either accept, resulting in payoffs  $2(p) + x^2$  and  $2(1 - p - c) + 1 - x^2$ , or reject, resulting in payoffs  $3p$  and  $3(1 - p - c)$  respectively.

The analysis of this game is fairly straightforward but somewhat tedious. Begin with the conjecture that B is using a monotone strategy when evaluating offers, so that if B accepts an offer giving her  $x$  when she has cost  $c$  then she will accept if her cost is higher or the offer is higher. These strategies can be described with cutpoints. For a fixed offer in period 1, B accepts if her cost is above some level,  $c^1$ . Suppose that we reach an information set where the offer,  $x^1$ , was rejected and according to B's strategy her type must be  $c < c^1$ . The subform starting with A's second offer looks just like the take it or leave it game analyzed above, except that  $c$  is uniformly distributed on  $[0, c^1]$ . B will accept the last potential offer  $x^2$  if  $1 - p - c \leq 1 - x^2$  and so A's problem in selecting  $x^2$  is to maximize  $x^2 \frac{(c^1 - x + p)}{c^1} + p \frac{(x - p)}{c^1} + 2p$ . The first order condition and solution are similar to the standard game: the offer is  $x^2 = \frac{c^1}{2} + p$ . Because  $k = 0$ , B will make an offer that some

types will reject. Now consider the decision to accept or reject the initial offer,  $x^1$ . The key insight is to note that the type,  $c^1$ , that is indifferent between accepting the first and second offer will accept the second offer, and thus the first period offer,  $x^1$ , is accepted if  $2(1-x^1) \geq 1-p-c+1-p-\frac{1}{2}$ . So as a function of the initial offer,  $x^1$ , B rejects if  $c \leq c^1(x^1) = 2(x^1-p) - \frac{1}{2}$ . Given this, the first period offer,  $x^1$ , is chosen to maximize

$$2x^1 \left( 1 - 2(x^1 - p) + \frac{1}{2} \right) + \left( 2(x^1 - p) - \frac{1}{2} \right) \left( p + \left( x^1 - \frac{1}{4} \right) \frac{1}{2} + \frac{p}{2} \right)$$

Taking the first order condition and solving yields  $x^1 = \frac{23}{48} + p$ . And so if  $p < \frac{25}{48}$  (which is a bit more than  $\frac{1}{2}$ ), the first offer will be accepted by some but not all types of player B. Thus in equilibrium the fact that B rejects the initial offer provides information to A. Accordingly, information is gleaned by A based on whether B decides to continue fighting or acquiesce. Put succinctly, direct learning from what happens on the battlefield is not necessary; strategic players may learn from the fact that other strategic players choose to fight. Similar exercises are carried out by Nalebuff (1991) and Gurantz and Hirsch (2017) as they explore how Jervis's (1982) claim about self-fulfilling beliefs can and cannot lead to war.

**MODELING**

In our view, modeling is a slightly different exercise than theory building. If theory building is like developing our understanding of fluid dynamics, which can also allow us to understand flight, modeling is analogous to

building wings or a scale airplane to see if we can in fact make something fly.<sup>9</sup>

Modeling is an important part of the scientific process, creating a library of domain specific abstractions that bring general theoretical forces to a specific problem in a particular way. Modeling tends to be different from theory building because it tends to make domain specific assumptions that emphasize different elements of a political environment, like deciding who are the relevant players and what their preferences are, with the objective of explaining a concrete set of observable events. These models then tie specific assumptions to testable hypotheses. They also provide a means for making predictions about correlations we should see in observable data through comparative static analysis.

To the modeler, then, the task of deriving relationships between independent and dependent variables has a place of privilege. The path taken by researchers proceeding in this way needs little explanation. The author provides enough specificity so that either an explicit or implicit characterization of the equilibrium can be obtained. Then, often using differential calculus, the relationship between parameters and equilibrium quantities are explored.<sup>10</sup> But see Ashworth and Bueno de Mesquita (2005) for a review of other useful techniques. For example, we may return to the basic bargaining game from above in which state A is empowered to make a take it or leave it offer. But we might want to be more general about utility functions. Assume that  $u_i$  is strictly monotone and continuous for both players. Further assume that with  $u_i(0) \leq v_i \leq u_i(1)$  for both players, where we let  $v_i$  denote the reservation payoff to  $i$  from not reaching a settlement. Subgame perfection then requires that B accept if and only if the offer beats her reservation value and A keeps as much as she can subject to B accepting. This reduces to the following equilibrium condition

$$u_B(1-x^*) - v_B = 0.$$



The implicit function theorem then implies that  $x^*$ , the share to player A, is decreasing in  $v_B$ , the reservation payoff of B. One could parameterize  $u_A(x, \gamma_A)$  and  $u_B(1-x, \gamma_B)$  as dependent on other covariates. The above then yields

$$u_B(1-x^*, \gamma_B) - v_B = 0$$

and  $\frac{\partial x^*}{\partial \gamma_A} = 0$  and because  $u_B(1-x, \gamma_B)$  is decreasing in  $x$ , the sign of  $\frac{\partial x^*}{\partial \gamma_B}$  is the same as the sign of  $\frac{\partial u_B}{\partial \gamma_B}$ . Sometimes the comparative statics are not monotone and pointing this out can help explain why a long series of empirical papers with linear models yield different (or consistently insignificant) results. A linear model of a non-monotone relationship hinges on what part of the parameter space the data comes from. Accordingly, revising the expectations of primarily empirical scholarship to anticipate potentially non-monotone relationships can be especially valuable. Romer and Rosenthal's (1978) treatment of budgeting hinges on showing that in the spatial model the reservation utility  $v_B$  is a non-monotone function of the status quo. Very low and very high status-quo are bad for a moderate veto-player. Equilibrium offers should, then, be a non-monotone function of the status quo.

By what criteria should we evaluate these models? One criteria is its empirical consistency. The process of testing is also one that is very familiar. Here the theoretical model and the data in the world are considered to be independent objects. In this type of exercise the empirical element of the analysis tests to see if the implied correlates of the theoretical model exist in the data. In this way, the testing process is one where the 'theorist hat' comes off and the 'statistician's hat' goes on. Determining what is the best test of a theoretically derived positive or negative correlation still involves thinking through the econometric concerns that are present in any data analysis.

Examples include Benson (2012), where he develops a model of ambiguous commitments

to manage the incentive to start adventurous conflicts when they know they have the support of powerful allies. Benson tests the insights of his model on alliance data and conflict initiation, showing that ambiguous commitments do temper the aggressiveness of allies. Similarly, Ritter (2014) develops a model of leader security in office and then uses event data from 1990 to 2004 to analyze the effects of repression using a two-stage estimator. She finds executive office security decreases the likelihood that repression will occur in the first place, that is, peace inducing policy adjustments are made by secure leaders, but secure leaders are more severe in their repression when protests arise.

Furthermore, the traditional test of the null hypothesis also makes sense in this context. Specifically, it makes sense to hold as the null hypothesis that the derived comparative static *does not* exist in the world. It is only when such a position can be rejected with the appropriate confidence that we believe the analyst has found empirical evidence that their model is describing a force possibly present in the world.

It is also the case that the larger the collection of associations that can be found in the data, and the more these relationships distinguish this model's explanations from other models, the better.

## QUANTIFYING MODELS

Sometimes, and with increasing frequency, similar domain specific models are empirical in a completely different way. Researchers look to quantify the theoretical model directly using observed behavior and other data to estimate unobservable (or unobserved) theoretical quantities through the actors' revealed preferences or beliefs. Unlike in the model-testing paradigm, when quantifying a model the analyst asks: what is most likely to be true about preferences, beliefs or other theoretical unobservables in this model if it were

the process by which the data we observe is generated?

**Estimation**

Return to our running example. A robust phenomenon is that even in settings where  $v_A = v_B = 0$  and the utilities are strictly monotone in shares, we observe offers that are not close to  $x = 1$  and we observe the rejection of offers that give B substantially more than 0. A common explanation is that players tend to have other-regarding preferences. Player B can be sufficiently bothered by inequities that she will forego positive rents to rule out an inequitable split. Alternatively, even if player A thought she could get away with taking nearly all of the pie, she might feel remorse and thus offer more equitable splits. Outside the laboratory one might believe that concerns related to the fungibility of resources lead player B to have a preference for more equity, or political factors may generate altruism in A. With this idea in mind, we can trace out the structural approach. The starting point is a sample. In the extremely simple case, suppose we observe an offer of  $x$  by A and acceptance by B. Further suppose that we are inclined to parametrize  $u_A, u_B$  to capture a degree of other-regarding preferences. Namely, assume that  $u_A(x) = x - \beta_A |x - (1-x)|$  and  $u_B(1-x) = 1-x - \beta_B |x - (1-x)|$ ,  $v_A = v_B = 0$  and that A is empowered to make a take it or leave it offer. The simplest structural econometric exercise involves estimating the parameters  $\beta_A, \beta_B$  from a sample consisting of one play of the game, namely the path:  $x$  is offered by A and B accepts. It is natural to think of this as a maximum likelihood estimation problem. But we cannot yet specify the likelihood function.

First, solve the game as a function of parameters  $\beta_A, \beta_B$ : Player B will accept the offer giving A  $x \geq \frac{1}{2}$  if

$$1 - x - \beta_B(2x - 1) \geq 0.$$

The largest  $x$  that satisfies this is  $x = \frac{1 + \beta_B}{1 + 2\beta_B}$ .

The proposer’s problem is then  $\max_x x - \beta_A(2x - 1)$  subject to the constraint that  $x \leq \frac{1 + \beta_B}{1 + 2\beta_B}$ . Given that A’s utility is linear in  $x$  it is clear that if  $\beta_A$  is low then the solution is  $x = \frac{1 + \beta_B}{1 + 2\beta_B}$ . But if  $\beta_A$  is higher than  $u_A$  is maximized by selecting  $x^* = \frac{1}{2}$ .

Accordingly, as a function of the data ( $x$ , accept) the parameters are partially identified. In particular, if  $x = \frac{1}{2}$  then we infer  $\beta_A \geq \frac{1}{2}$  and any value of  $\beta_B$  is possible. If, however,  $x > \frac{1}{2}$  then  $\beta_B = \frac{1-x}{2x-1}$  and any value of  $\beta_A \leq \frac{1}{2}$  is possible. No parameters are consistent with  $x < \frac{1}{2}$ .

Although our analysis of the structural model is stark, limited to a dataset that includes one observation, it serves to flesh out the ways in which one can very closely tie the model, equilibrium analysis and econometric tests. Here, the nature of identification is interesting. Some patterns of play have more power to narrow down the set of possible deep parameters than others. In practice it is typically convenient to add more noise to the model or introduce noise that accounts for measurement error faced by the scholar.

Consider another example: suppose we were to focus on the strategic aspects of war fighting in our bargaining model. Kenkel and Ramsay (2019) consider the case where war fighting is a strategic contest in which the probability that one side wins is

$$p_A(e) = \frac{\sum_{j \in A} m_j e_j}{\sum_{j=1}^I m_j e_j},$$

for all those states exerting effort on side A. For their purposes, substantive variables affect the effort multiplier such that  $m_j = \exp(X_j \beta)$  and each participant pays a cost for

effort level  $e_j = c_j e_j$ , with the marginal cost  $c_j = \exp(Z_j \gamma)$ .

For the two country case, we can solve for equilibrium effort levels, substitute them into the contest function and get the probability that state A wins equal to

$$\frac{\exp(X_j \beta - Z_j \gamma)}{\sum_{i \in \{A, B\}} \exp(X_i \beta - Z_i \gamma)}$$

That is, the two player strategic contest function, in equilibrium, generates logit war winning probabilities where the coefficients of the estimator are the effects of variables on the marginal return to effort and the marginal cost of war. Given these estimates, Kenkel and Ramsay (2019) can back out unobserved equilibrium efforts and bargaining offers, and make inferences about the challenger's beliefs regarding the target's resolve.

In this case the model acts as a set of identifying assumptions specifying what observed actions must imply about the preferences, information or values of important theoretical variables unobservable to the analyst if the modeled strategic interaction is producing the observed data. This exercise can be based on estimation and statistical inference or calibration. The result allows both for forecasting and evaluation of theoretically motivated and empirically consistent counter-factual analysis.<sup>11</sup> Similar work has been done with respect to crisis dynamics (Signorino, 1999), territorial consolidation (Carter, 2010) and the effects of democratic audience costs on escalation (Kurizaki and Whang, 2015).

Notice that there is nothing really to test about the theory here. What the analyst does is estimate distributions of parameter values consistent with the model and the data. Unlike in the model-testing framework, in this process of bringing the model to data, the model and estimation are part of one enterprise.

So how might we assess such a model? Informally we may look to see how well the model forecasts events of interest, either in or

outside the sample data. We might also review estimated parameter values for their intuitive plausibility. But really, what you would want to do is compare this model, including its quantitative estimates, with another model and its estimates. In this framework there is no null hypothesis test: models beat other models.

Alternative methods used for model forecasting and counter-factual analysis are fundamentally different from this kind of theory based estimation. Statistical methods such as non-parametric estimation, vector autoregression, machine learning and causal inference are all aiming to estimate some version of the policy function mapping observables to outcomes, the credibility of which is determined by the flexibility of the procedure and whether the variation in treatment variables of interest is plausibly exogenous. These methods do not, however, generate estimates of the theoretical parameters, and also assume that there is a single policy function to be estimated. We know from theory – by the existence of multiple equilibria for fixed variable values or different, but unique, mappings from observables to outcomes for different values of the parameters – that these policy functions may not be unique or be useful for counter-factual analysis. This does not mean that such estimates are not useful; it just means it takes special care to determine what is being estimated and whether counter-factuals can be performed on the policy functions obtained by these means.

Structural models, on the other hand, directly estimate theoretical parameters that allow the analysis to make theoretically informed and empirically consistent counter-factual claims. But here too, questions can arise about estimation in the face of multiple equilibria, equilibrium selection and what exactly it means to have the best estimate. The difference is that one cannot proceed with the inference without explicitly making these choices, resulting in transparent, if still debatable, analysis and results.

## CONCLUSION

In the study of politics, game theory is a tool that can be useful in the performance of many different tasks. Each of these tasks contributes to a larger exercise of theory building, model creating and empirical analysis, but they do not all have the same purpose, nor are their contributions measured by the same metric. Something that might be a virtue for one endeavor, such as generality in theory building, might be too abstract to gain traction when modeling a specific political event to think about policy counter-factuals. Game theory has brought many advances to our understanding of politics, but understanding and judging individual contributions requires thinking about the purpose of the analysis.

## Notes

- 1 We mean no offense to the many other fine examples we could have chosen instead.
- 2 In some important ways our view of models is similar to that in Clarke and Primo (2012).
- 3 Readers needing additional development should consult a standard text: Fudenberg and Tirole (1991), McCarty and Meirowitz (2007) or Osborne (2004).
- 4 See also Franzese (Chapter 31) in this *Handbook* for a similar conception of models, which can be aimed at measurement, causal inference, or prediction.
- 5 As much as is possible, we will use this bargaining model as a running example to make ideas concrete.
- 6 It does not matter if the accept/veto decisions are made sequentially or simultaneously in this private values setting.
- 7 For a similar result with correlated types see Compte and Jehiel (2009).
- 8 Note that in some, if not many, settings, conflict can begin without both sides agreeing to fight. But the key conceptual debate on mutual optimism centers on an argument that both sides would want to fight.
- 9 A note on terminology: some may prefer to use the term applied modeling. We prefer to dispense with the term applied, because in a clear sense all of this is applied. Many scholars already draw a distinction between theory and applied theory, where the former is about rationality, properties of equilibrium concepts and relationships between classes of games (and more) and the latter is about using game theory to understand somewhat more specified classes of games that match up to an area of substantive research. To the extent that everything in this chapter is therefore applied, we see reference to theorizing about narrow domains as modeling.
- 10 But see Ashworth and Bueno de Mesquita (2005) for a review of other useful techniques.
- 11 In an exceptional instance of scholars agreeing on something, see also Franzese (Chapter 31) and Slantchev (Chapter 5) in this *Handbook* for compatible views on this approach.

## REFERENCES

- Ashworth, Scott, and Ethan Bueno de Mesquita. 2005. 'Monotone Comparative Statics for Models of Politics.' *American Journal of Political Science* 50(1):214–231.
- Banks, Jeffrey S. 1990. 'Equilibrium Behavior in Crisis Bargaining Games.' *American Journal of Political Science* 34(3):599–614. URL: <https://www.jstor.org/stable/2111390>
- Benson, Brett V. 2012. *Constructing international security: alliances, deterrence, and moral hazard*. Cambridge University Press.
- Bils, Peter, Richard Jordan and Kristopher Ramsay. 2017. 'Fanatical Peace.' Working Paper.
- Blainey, Geoffrey. 1988. *Causes of war*. (3rd edn). Simon & Schuster..
- Carter, David B. 2010. 'The Strategy of Territorial Conflict.' *American [abbreviate American to correct Conflict] Journal of Political Science* 54(4):969–987. URL: <http://doi.wiley.com/10.1111/j.1540-5907.2010.00471.x>
- Clarke, Kevin A. and David M. Primo. 2012. *A model discipline: political science and the logic of representations*. Oxford University Press.
- Compte, Olivier and Philippe Jehiel. 2009. 'Veto Constraint in Mechanism Design: Inefficiency with Correlated Types.' *American Economic Journal: Microeconomics* 1(1):182–206. URL: <https://www.aeaweb.org/articles?id=10.1257/mic.1.1.182>
- Fearon, James D. 1995. 'Rationalist Explanations for War.' *International Organization*. 49(3): 379–414.
- Fey, Mark and Kristopher W. Ramsay. 2007. 'Mutual Optimism and War.' *American Journal of Political Science* 51(4):738–754. URL: <https://www.jstor.org/stable/4620097>

- Fey, Mark and Kristopher W. Ramsay. 2011. 'Uncertainty and Incentives in Crisis Bargaining: Game-Free Analysis of International Conflict.' *American Journal of Political Science* 55(1):149–169. URL: <https://www.jstor.org/stable/25766260>
- Fey, Mark and Kristopher W. Ramsay. 2016. 'Mutual optimism in the bargaining model of war.' *Unpublished manuscript, Princeton University*.
- Fudenberg, Drew and Jean Tirole. 1991. *Game theory*. MIT Press.
- Gurantz, Ron and Alexander V. Hirsch. 2017. 'Fear, Appeasement, and the Effectiveness of Deterrence.' *The Journal of Politics* 79(3):1041–1056. URL: <https://www.journals.uchicago.edu/doi/abs/10.1086/691054>
- Jervis, Robert. 1982. 'Deterrence and Perception.' *International Security* 7(3):3–30. URL: <http://www.jstor.org/stable/2538549>
- Kenkel, Brenton and Kristopher W. Ramsay. 2019. 'A Quantitative Bargaining Theory of War: Bargaining and Fighting in Militarized Interstate Disputes.' Working Paper.
- Kurizaki, Shuhei and Taehee Whang. 2015. 'Detecting Audience Costs in International Disputes.' *International Organization* 69(4):949–980.
- McCarty, Nolan and Adam Meirowitz. 2007. *Political game theory: an introduction*. Cambridge University Press.
- Nalebuff, Barry. 1991. 'Rational Deterrence in an Imperfect World.' *World Politics* 43(3):313–335.
- Osborne, Martin J. 2004. *An introduction to game theory*. Oxford University Press.
- Powell, Robert. 2004. 'Bargaining and Learning while Fighting.' *American Journal of Political Science* 48(2):344–361.
- Ritter, Emily Hencken. 2014. 'Policy Disputes, Political Survival, and the Onset and Severity of State Repression.' *Journal of Conflict Resolution* 58(1):143–168. URL: <https://doi.org/10.1177/0022002712468724>
- Romer, Thomas and Howard Rosenthal. 1978. 'Political Resource Allocation, Controlled Agendas, and the Status Quo.' *Public Choice* 33(4):27–43.
- Schelling, Thomas C. 1980. *The strategy of conflict*. Harvard University Press.
- Signorino, Curtis S. 1999. 'Strategic Interaction and the Statistical Analysis of International Conflict.' *American Political Science Review* 93(2):279–297.
- Slantchev, Branislav L. and Ahmer Tarar. 2011. 'Mutual Optimism as a Rationalist Explanation of War.' *American Journal of Political Science* 55(1):135–148. URL: <https://www.jstor.org/stable/25766259>
- Smith, Alastair and Allan C. Stam. 2004. 'Bargaining and the Nature of War.' *Journal of Conflict Resolution* 48(6):783–813.
- von Neumann, John and Oskar Morgenstern. 1944. *Theory of games and economic behavior*. Princeton University Press.
- Wagner, R. Harrison. 2000. 'Bargaining and War.' *American Journal of Political Science* 44(3): 469–484.



# The Spatial Voting Model

James Adams, Samuel Merrill III and Roi Zur

## INTRODUCTION

We first discuss spatial voting models of party-centered elections<sup>1</sup> involving a continuum that is typically a broad Left–Right (liberal–conservative) ideological dimension which encompasses debates over income redistribution and government intervention in the economy, a major electoral cleavage in most western democracies. The Left–Right continuum is *positional* in that different voters prefer – and different parties advocate – different positions along this continuum. We review the assumptions that underlie the positional spatial model of elections, and then survey spatial modeling research on parties’ positional strategies in these types of elections. We describe how a fundamental spatial modeling result is that vote- or office-seeking parties are typically motivated to advocate policies near the center of public opinion, i.e., near the middle of the distribution of voters’ preferred Left–Right positions (Downs, 1957). This prediction applies most

strongly to elections between two dominant parties (as in the United States and, in earlier periods, British politics), but important elements of this prediction extend to elections featuring three or more parties, i.e., multi-party elections (as in most other western democracies). We then identify an empirical puzzle, namely that parties typically fail to converge towards the center of the voter distribution (or towards each other) to the extent predicted by basic positional spatial models.

We then discuss two possible solutions to this empirical puzzle of party positional divergence. The first extends the spatial model of *voters’ motivations* to consider election scenarios where, in addition to their Left–Right concerns, voters also weigh parties’ reputations along character-based dimensions such as party elites’ reputations for competence, integrity and leadership – i.e., to a *valence dimension* of party competition. In the case of valence issues, it is perhaps plausible to assume that voters agree about which character-based traits

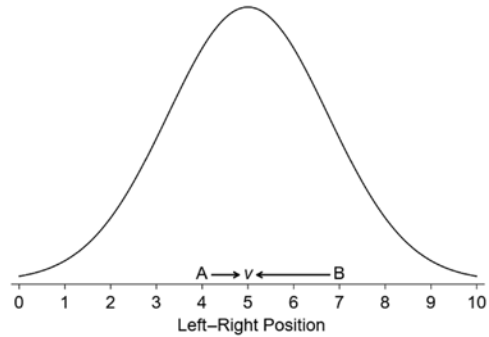
they prefer, i.e., all citizens prefer that party elites possess more rather than less competence, integrity and leadership ability, and moreover all political parties seek to publicly convey positive valence attributes.<sup>2</sup> However, the parties may be differentiated on valence because voters perceive some parties' elites possessing superior character-based valence traits compared with rival parties' elites. We show how, in a spatial model that includes both a positional and a valence dimension, vote- and office-seeking parties may rationally propose radical Left–Right positional strategies that differ sharply from rival parties' positions, and also from the mainstream of public opinion. These radical strategies are not predicted by the positional spatial model that omits valence.

We next explore an extension of the spatial model that considers an alternative *party motivation*, namely that parties – like voters – have preferences over government policy outputs. We show how this assumption can prompt policy-seeking parties to advocate sharply noncentrist positions relative to rival parties, and also relative to their expectations about public opinion.

## EMPIRICAL AND THEORETICAL RESEARCH ON ALTERNATIVE DIMENSIONS OF VOTING

### *Positional Dimensions of Voting*

The spatial model of party competition is associated with the research of Harold Hotelling (1929) and Anthony Downs (1957). The simplest spatial model represents policy debates as options along a one-dimensional continuum or line, and posits that both the policies that voters prefer and the policies that parties advocate are represented by positions along this line. The best known dimension in contemporary western democracies is the Left–Right or liberal–conservative dimension, which involves disagreements over issues such



**Figure 12.1** Illustrative placements of a voter *v* and parties *A*, *B*, on Left–Right ideology

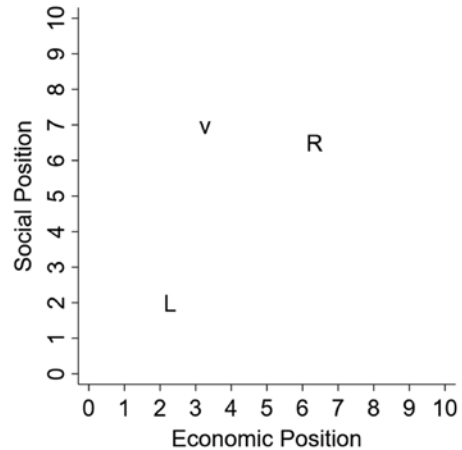
as government intervention in the economy and income redistribution. Figure 12.1 illustrates such a one-dimensional model representing the Left–Right scale, with a voter *v* located closer to party *A* than to party *B*, so we expect that, all else equal, this voter would prefer party *A* to party *B*. In this illustration the Left–Right scale runs from zero to 10, which mirrors the scale that is typically presented to respondents in national election surveys, where the convention is that higher numbers denote more right-wing positions. Empirical studies find that this Left–Right dimension is central to voter choice in western democracies, with communist, socialist and labor parties associated with leftist positions, while conservative and Christian democratic parties are associated with right-wing positions (Dalton et al., 2011).

We note that media and political elites' references to 'left' versus 'right' are shorthand for a *set of* positions on more specific issues. Thus the designation that a party (or a voter) is *left-wing* on economic issues denotes that the party advocates more progressive tax policies and expanded social welfare programs that tend to redistribute income from more to less affluent citizens, whereas *right-wing* parties (and voters) favor less progressive tax policies and a more modest social welfare safety net. These left- or right-wing labels may also encompass employment-related

issues such as the legal rights and restrictions on labor unions, the minimum wage and the degree of government intervention in the economy (Adams, 2018). Moreover, as one would expect, empirical studies find that more left-wing citizens tend to support left-wing parties while right-wing voters tend to support right-wing parties (Bølstad and Dinas, 2017; Powell, 2000).

We note that although Left–Right economic policy dominated positional debates in most western democracies at least through the 1970s, cross-cutting cleavages have emerged pertaining to debates that are not directly aligned with Left–Right economic issues. One set of debates pertains to social and moral issues including abortion, gay rights and gender equality, which cross-cut the Left–Right economic dimension because citizens’ economic views do not necessarily correlate strongly with their views on social and moral issues (Kitschelt, 1994; Marks et al., 2006). Another emerging cleavage pertains to issues involving race, religion and immigration, where, again, citizens (and parties) who share similar Left–Right economic viewpoints may disagree over issues such as affirmative action, multiculturalism and immigration policies. Well before the 1970s, the issue of civil rights in the United States, primarily associated with long-standing racial attitudes, defined a major dimension that cut strongly across the economic dimension and that split the Democratic Party into Southern and Northern branches.

Accordingly, the one-dimensional scale (from liberal to conservative or Left to Right) discussed earlier can be extended to a two-dimensional positional model, represented by a plane with  $X$  and  $Y$  axes. For example, one dimension, say  $X$ , might represent economic (traditional Left–Right) issues while the second dimension  $Y$  might represent social issues. In this two-dimensional model, a voter  $v$  has preferred positions on both the economic scale (say  $v_1$ ) and on the social issue scale (say  $v_2$ ), and is hence represented by a point  $(v_1, v_2)$  in the plane. Generally, it is



**Figure 12.2** Illustrative placement of a voter  $v$  and parties  $L$  and  $R$  in a two-dimensional space

assumed that a voter prefers a party whose expressed positions in this two-dimensional plane are nearest to their own. Figure 12.2 illustrates a voter  $v = (v_1, v_2)$  who is liberal on economic issues but conservative on social issues and is closer overall to party  $R$  than to party  $L$ , even though this voter is closer to  $L$  on the single economic dimension.

### ***A Non-Positional Dimension of Voting and Elections: 'Valence' Issues***

As discussed above, the Left–Right economic dimension – along with dimensions pertaining to social/moral issues, multiculturalism, and so on – is *positional* in the sense that voters (and parties) hold conflicting positions. Despite the continuing relevance of such dimensions, in recent years scholars have considered the effects of *valence dimensions* of voters’ party evaluations. Stokes (1963) coined this term to denote dimensions ‘on which parties or leaders are differentiated not by what they advocate, but by the degree to which they are linked in the public’s mind with conditions,



goals, or symbols of which almost everyone approves or disapproves' (Stokes, 1992: 143). Valence dimensions include such attributes as parties' (and party leaders') images with respect to honesty, competence, empathy and charisma.<sup>3</sup> These dimensions contrast with the Left–Right positional dimension, on which 'parties or leaders are differentiated by their advocacy of alternative positions' (Stokes, 1992: 143).

Valence considerations matter because although nearly all voters prefer that parties be more competent and honest, voters may perceive different parties possessing differing degrees of positive valence. In American and British politics, for instance, national political candidates including Dwight Eisenhower, Bill Clinton and Tony Blair were widely perceived as competent and likable during the periods in which they first led their parties into national elections, whereas others such as Michael Dukakis, Michael Foot, Ed Miliband and Theresa May were viewed far less positively along these dimensions. Cross-national research affirms the growing importance of valence issues across western democracies (see, e.g., Clark, 2009; Abney et al., 2013).

### **SPATIAL MODELS OF ELECTIONS WITH OFFICE-SEEKING PARTIES: THE PURELY POSITIONAL MODEL PREDICTS PARTY CONVERGENCE**

We first review purely *positional* models of party competition, while deferring valence considerations. Downs (1957) was the first to apply this framework to electoral competition, assuming, first, that both parties' positions and voters' *ideal points*, i.e., the positions voters prefer, are arrayed over a unidimensional, positional issue space. Here we illustrate the Downsian model in terms of the Left–Right continuum, but Downs' arguments apply to any positional dimension. Second, Downs assumed that voters evaluate the

parties based – as spatial modelers continue to do, at least in part – on the *proximity* of their preferred positions to the parties' positions, i.e., that voters prefer more spatially proximate parties. Third, political parties strategically announce positions that maximize their electoral prospects, i.e., parties are *office-seeking* and propose policies purely as a means of winning elected office. Downs justified this assumption by emphasizing the private benefits politicians obtain from holding office, including prestige and celebrity, their government salaries, and opportunities to distribute government jobs and contracts to political allies and family members. Hence the basic Downsian model posits that voters are purely *policy-oriented*, i.e., they invariably support the party that offers the most attractive policy positions, whereas political parties are purely *office-seeking* in that they propose policies purely as a means of winning votes, and through this winning office.

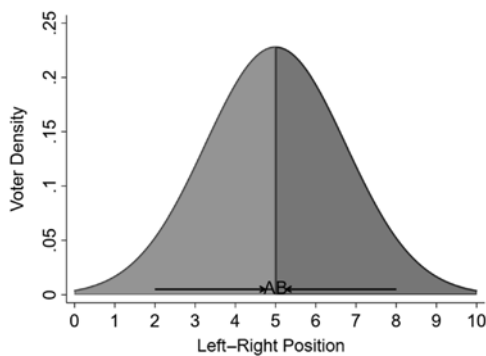
### ***Incentives for Party Convergence in the Positional Spatial Model: the Two-Party Case***

In two-party competition over a single positional dimension (here the Left–Right dimension), Black's (1948) Median Voter Theorem states that office-seeking parties converge to the *median voter's* position, i.e., the Left–Right position, such that half the electorate is located on either side.<sup>4</sup> To understand why two-party convergence at the median voter position is optimal for both parties in competition over one positional dimension, note first that, if party **A** locates at this median position and party **B** does not – say, it is at a position to the right of the median – party **B** will lose the election. This occurs because all voters to the left of the median, together with some to the right of it, will be nearer to and hence vote for party **A**, so that party **A** will win the election. However, party **B** can force a tie if it shifts in turn to also locate at the median voter position. Second, if both parties

locate away from the median voter position then either party can win the election by unilaterally shifting to the voter median (Adams, 2016).<sup>5</sup> Hence two-party, purely positional spatial competition provides centripetal incentives for parties to converge toward each other, and toward the center of the distribution of voters' ideal points.

Figure 12.3, which displays a Left–Right continuum (the horizontal axis) with a distribution of voters' ideal points (where the height of the line along the vertical axis represents the density of these ideal points at each position), illustrates this logic. In this example the voter distribution is assumed to be normal, with a median position at the center of the Left–Right scale – a distribution which, as we shall see, roughly approximates the distributions of voters' preferred positions in the electorates of many western democracies. Here we display the Left–Right scale running from zero to 10 (with higher numbers denoting more right-wing positions), which, as we discuss below, is the scale that is usually included in national election surveys to elicit respondents' ideologies. In this scenario the two political parties **A** and **B** are drawn towards the median voter position – and hence towards each other – at the center point (here 5) of the Left–Right scale.

The configuration in Figure 12.3, in which the two office-seeking parties each occupy the median voter position, constitutes a *Nash*



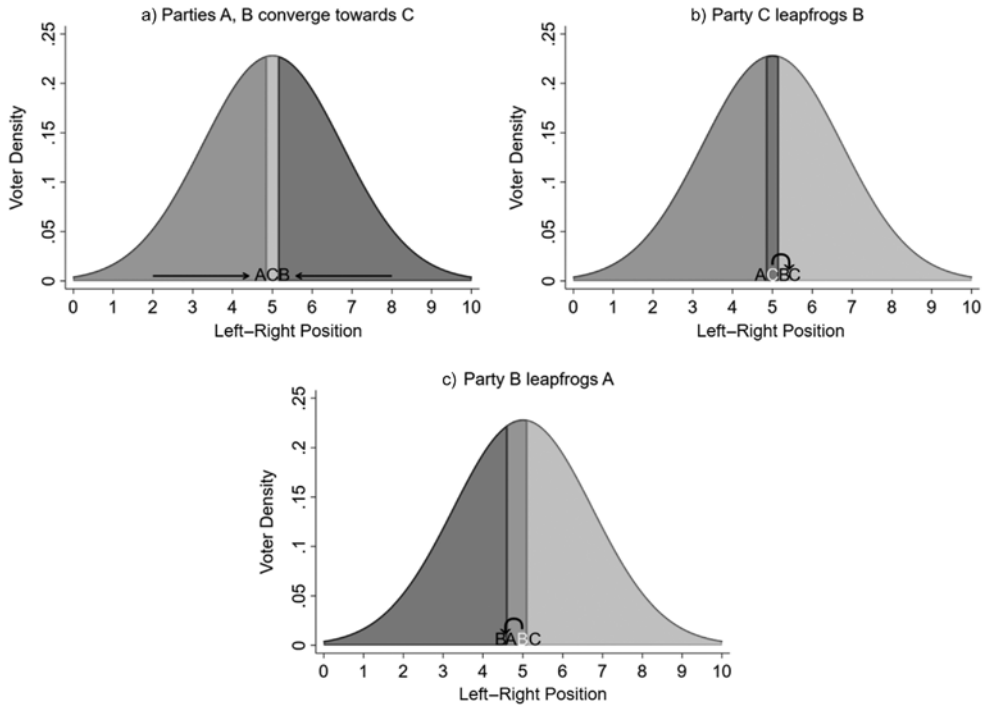
**Figure 12.3** How two-party, positional competition motivates party convergence

*equilibrium* in party strategies, i.e., a configuration of positions such that no party is motivated to unilaterally change its position, given its opponents' positions. In the Downsian two-party spatial model with a single positional dimension, the pairing of parties **A** and **B** at the median voter position constitutes a unique Nash equilibrium.

### **Extensions to Positional Spatial Competition with Three or More Parties**

Downs' arguments about incentives for two-party convergence provided an intuitive explanation for political competition in the American post-World War II party system (Downs' book was published in 1957), in which the Democratic and Republican parties both presented relatively similar and moderate Left–Right economic policies. However, virtually all western democracies outside the United States feature at least three competitive parties, i.e., they are *multi-party systems*.<sup>6</sup> The convergent Nash equilibrium for one-dimensional, two-party positional competition breaks down when additional parties compete. The research of Eaton and Lipsey (1975) – focused on competition between firms but directly translatable to elections – and Cox (1990) suggests that in multiparty elections the *centripetal incentives* motivating vote-seeking parties to converge toward similar positions – and toward the center of the distribution of voters' ideal points – are balanced by *centrifugal incentives* to differentiate their policy positions (Adams, 2018). Moreover, for the basic model we have discussed so far, a Nash equilibrium in party positions rarely exists in multiparty elections.

The above points can be grasped by considering a three-party election along the Left–Right dimension where, regardless of the distribution of voters' ideal points, the two 'peripheral' parties – i.e., the parties that announce the most left- and right-wing



**Figure 12.4 The dynamics of three-party positional competition**

positions – can increase their support by converging towards the position of the third ‘interior’ party. Figure 12.4a illustrates this incentive, with the peripheral parties **A** and **B** converging towards the interior party **C**, which causes **C** to be ‘squeezed’ and hence to win few votes.<sup>7</sup> This convergence prompts the interior party **C** to leap-frog the position of one of its rivals (Figure 12.4b), and the party that is leap-frogged will in turn be squeezed, motivating it to leap-frog another party in turn (Figure 12.4c), and so on without limit.

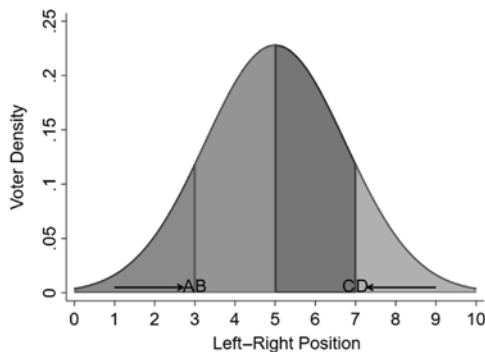
The centrifugal incentive described above, which counteracts peripheral parties’ centripetal incentives to converge toward the interior party’s position, implies that no Nash equilibrium is possible in three-party elections with vote-maximizing parties. Note, moreover, that in multiparty, one-dimensional positional competition with any number of parties,

the left- and right-most parties are invariably motivated to converge toward the positions of their immediate ‘neighbor’ parties along the positional continuum, because this maximizes the peripheral parties’ vote shares. Figure 12.5 illustrates this dynamic for a scenario involving four parties labeled **A**, **B**, **C** and **D**, located from left to right across the ideological spectrum, where the peripheral party **A** converges toward party **B** while the peripheral party **D** converges toward party **C**. (In this illustration we again depict a normal distribution of voter ideal points.)

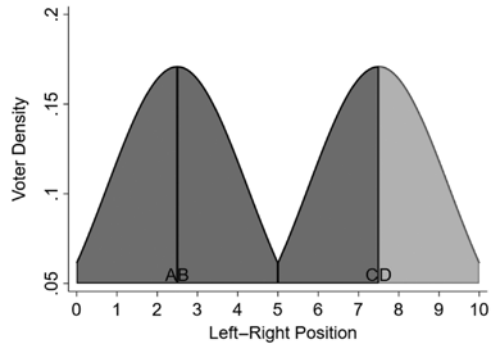
Hence for both two-party and multi-party elections with an even number of parties, the positional spatial model with vote-seeking parties predicts that either the left- and right-most parties will ‘pair’ with each other (two-party case) or “pair” with their nearest interior competitor (more-than-two-party case). In two-party elections this

implies complete party convergence to the median voter position; in multiparty elections this implies that the most extreme parties will converge towards the position of the most proximate interior party, i.e., they will ‘pair’ with an interior party if there are an even number of parties, although, depending on the number of pairs, the pairings will occur at different positions.

We briefly note that whether a Nash equilibrium exists for multiparty competition over one positional dimension depends on several technical details of the voter distribution and the number of competing parties. One condition is that, except for a uniform voter distribution, the number of parties cannot be more than double the number of ‘modes’ in the voter distribution, so that for a unimodal distribution (such as the normal distribution pictured in Figures 12.3–12.5) no Nash equilibrium exists for more than two parties, a bimodal distribution cannot support an equilibrium for more than four parties, and so on. Thus, no equilibrium in vote-maximizing strategies is possible for the scenario pictured in Figure 12.5, in which four parties compete over a unimodal distribution of voter positions. To see this, note that even if peripheral parties **A** and **D** ‘pair’ with the interior parties **B** and **C**, respectively, this cannot constitute a Nash equilibrium since the interior



**Figure 12.5 Centripetal incentives in a four-party election: the peripheral parties converge toward their ideological ‘neighbors’**



**Figure 12.6 Example of four-party Nash equilibrium configuration in competition over a bimodal voter distribution**

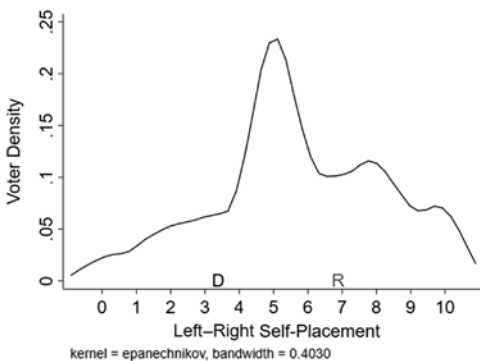
parties can increase their support by unilaterally shifting their positions towards the center of the voter distribution. On the other hand, Figure 12.6 illustrates a bimodal distribution of voters’ ideal points, where one mode is at 2.5 and a second mode at 7.5. In this case, parties **A** and **B** locate at the left-wing mode and parties **C** and **D** locate at the right-wing mode. Note, again, that these positions will constitute a Nash equilibrium only under a set of specific conditions that are beyond the scope of this chapter (but see Eaton and Lipsey (1975) for the conditions that support a multiparty equilibrium for one-dimensional competition).

### **AN EMPIRICAL PUZZLE: REAL WORLD PARTIES’ POSITIONS DO NOT CONVERGE**

The central qualitative prediction associated with the Downsian model of two-party positional competition, namely that the competing parties will offer similar, if not identical, policy positions that reflect the central tendency of public opinion,<sup>8</sup> matched the party dynamics of postwar American and British politics up through the mid 1970s. These patterns featured the Democratic and Republican parties – along with the Labour and

Conservative parties, which dominated British politics throughout this period – presenting similar, moderate Left–Right positions with respect to social welfare policy and government intervention in the economy.

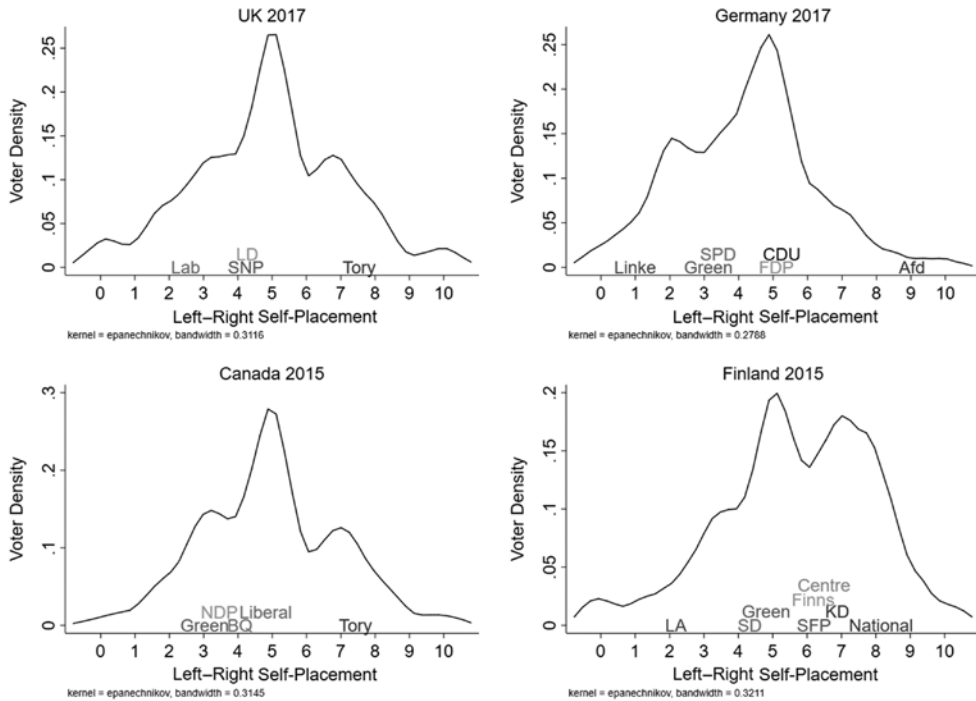
However, beginning in the late 1970s, party politics in both countries diverged from the centripetal logic of the two-party spatial model, with the British Conservatives led by Margaret Thatcher (and her successor John Major) and the American Republican Party under Ronald Reagan (and his successors) shifting their policies sharply to the right, away from their opponents' positions and from the center of public opinion (see, e.g., Adams et al., 2012). Moreover, contra the Downsian prediction that radical policy positioning is electorally damaging, both conservative parties' rightward shifts met with electoral success. And the policy divergence between Republican and Democratic Party elites has continued to widen since the 1980s – indeed, party polarization (at both the elite and mass levels) is one of the most widely studied features of contemporary American politics (see, e.g., McCarty et al., 2006).<sup>9</sup> Figure 12.7 displays this pattern based on survey responses from the 2016 American National Election Study in which survey respondents were asked to place themselves along a Left–Right scale running from 0 ('extremely left-wing') to 10 ('extremely



**Figure 12.7** Distribution of American survey respondents' Left–Right self-placements and their mean party placements, 2016 National Election Study

right-wing'), and also to place the Democratic and Republican parties similarly.<sup>10</sup> The figure displays the distribution of respondents' self-placements, which very roughly resembles a normal distribution centered near the midpoint of the scale, and also displays respondents' mean placements of the Democratic and Republican parties. These mean placements are sharply differentiated, with the Democrats' mean perceived position (3.34) well to the left of the center of the voter distribution, and the Republicans' perceived position (6.92) well to the right of the center.

At the same time that the Downsian positional spatial model appeared to illuminate two-party US and British party politics at least up to the mid 1970s, empirical patterns of party positioning in multiparty systems – which constitute the vast majority of western party systems outside the United States – have rarely displayed the degree of party convergence predicted by the Downsian model. In particular, the party systems of multiparty western democracies often feature the anomaly of radical 'peripheral' parties that present positions far more extreme than those of their nearest competitor. Figure 12.8, which displays the distributions of voter positions and of parties' (mean perceived) positions for several western party systems (the UK, Canada, Finland and Germany), illustrates this phenomenon. (The data are based on the 2017 post-election surveys in the UK and Germany and the 2015 national election surveys in Canada and Finland.<sup>11</sup>) In every country we observe at least one peripheral party whose position is perceived as substantially more extreme than that of its nearest competitor.<sup>12</sup> Yet despite the fact that the logic of the Downsian positional model implies that parties will not adopt these unduly radical strategies – and that those that do initially adopt such strategies will eventually converge towards their nearest competitor – the empirical pattern we actually observe is one where radical peripheral parties maintain stable positions over time (see, e.g., Dalton and McAllister, 2015).



**Figure 12.8** Distributions of citizens' Left-Right self-placements and their mean party placements in the UK, Germany, Canada and Finland

### A POSSIBLE SOLUTION: SPATIAL MODELS THAT INCORPORATE VALENCE DIMENSIONS

It was in the context of Left-Right party polarization in the UK and the United States that spatial modelers began exploring whether valence dimensions of party evaluation could explain this empirical puzzle. As discussed above, valence dimensions differ from positional dimensions in that nearly all voters share the same preferences with respect to valence, i.e., voters prefer that party elites display higher degrees of competence, integrity, unity, compassion and leadership ability, and moreover all political parties strive to publicly project these positive valence-based qualities. However, not all parties succeed in conveying positive valence images to the public: some parties – but not others – become enmeshed in scandals; some

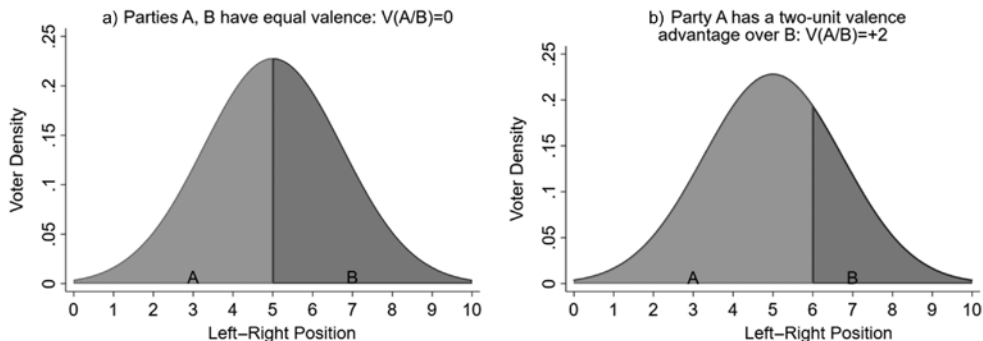
parties' elites appear internally divided, while others appear united; and different parties' leaders may be more or less successful at publicly conveying competence and leadership ability. (Moreover, the same party leader's valence image may fluctuate over time, as has notably been true for British politicians such as Margaret Thatcher, Tony Blair and Theresa May.) As a result, some parties may enjoy valence advantages compared to their opponents. Furthermore, unlike positional dimensions where parties are free to change their positions, parties have only limited abilities to 'strategize' over their valence images: they can strive to achieve and convey to the public an image of competence, honesty and unity, but if these efforts fail, parties cannot simply 'decide' to improve their valence images. Therefore, in the short term, political parties may be considered to occupy more or less fixed (positive or

negative) positions along valence dimensions of voter evaluation (but see Serra, 2010; Curini and Martelli, 2015).

One historical example of the electoral importance of valence issues is the British Conservative Party during the 25 years following World War II, when the party consistently enjoyed a positive public image (compared to its opponent the Labour Party) with respect to competence and leadership. During this period the Conservatives converged toward Labour's long-term leftist economic and social welfare orientation (for this reason the period has been labelled 'the Postwar Settlement', due to the unusual policy consensus between the parties), but largely based their electoral appeals on their positive valence-based image as a 'safe pair of hands' that could administer these policies more efficiently than Labour could. The combination of the Conservatives' 'us too' Left-Right positions and their positive valence image earned the party the derisive nickname 'the party that has no ideas but that knows how to govern'. This strategy proved electorally successful, as the Conservatives won three consecutive general elections between 1951 and 1959, primarily due to their superior valence image. In addition, the Conservatives' later electoral successes in winning four consecutive general elections between 1979 and 1992 – a period when the party had shifted sharply rightward on economic and social welfare policy – largely reflected the

Conservatives' even larger valence advantage arising from the Labour Party's widely publicized internal divisions and weak leadership across much of this period (Norton, 2000).

The consideration of the types of effects discussed above prompted scholars to incorporate valence dimensions into their spatial models. Feld and Grofman (1991) expanded the positional model to include voters' tendencies to accord incumbent parties or candidates a 'benefit of the doubt' that was unrelated to the incumbent's positional stances, and that was conceptually equivalent to an advantage on valence issues. The authors' approach posits that voters' valence- and position-based party evaluations can be meaningfully compared. Figure 12.9 illustrates this approach by incorporating voters' valence considerations into a spatial model that also includes the positional Left-Right dimension. Here we specify that citizens choose between parties **A** and **B** that are positioned at 3 and 7, respectively, along the scale, by comparing these parties' Left-Right positions *and* their valence images. Figure 12.9a displays a scenario where the parties have equal valence, i.e.,  $V(A/B)=0$ , so that all voters prefer the party with the more proximate Left-Right position. In this case a voter with a Left-Right ideal point located at 5 is indifferent between the parties, since this voter's position is equidistant between parties **A** and **B**, so that we label 5 the 'indifference point' for this scenario. All voters located to



**Figure 12.9** How valence affects voter choice in a model with one positional dimension

the left of 5 prefer Party A, while voters to the right of 5 prefer B.

Figure 12.9b displays a different scenario, where party A has a superior valence image. Specifically, citizens evaluate party A's valence advantage relative to party B as equivalent to two units of position along the 0–10 Left–Right scale, which we denote  $V(A/B)=+2$ . This implies that a voter will prefer party A to B *unless the voter's Left–Right ideal point is located at least two units nearer to party B than to A*.<sup>13</sup> In this example with  $V(A/B)=+2$ , a voter located at 6 on the positional scale is now indifferent between parties A and B – which are located at points 3 and 7, respectively – since this voter is located two units closer to B than to A on the Left–Right scale, a positional preference for B which exactly balances the voter's valence-based preference for A. All voters located to the left of 6 now prefer Party A, while those to the right of 6 prefer B.

Several spatial modeling studies explore how the introduction of valence dimensions affects parties' positional strategies (for example, Ansolabehere and Snyder, 2000; Serra, 2010).<sup>14</sup> A key insight from this literature is that valence-disadvantaged parties have incentives to diverge on position from their valence-advantaged rival(s). To grasp this strategy, note that if two parties converge on position then all voters rate the parties equally on position, and hence choose between the parties based entirely on valence

considerations. Hence if one party's valence substantially exceeds its competitor's, the valence-disadvantaged party must diverge from its opponent to win support. In this way the valence-disadvantaged party attracts voters whose ideal points are close to its position but far away from its opponent's position. For this reason, when voters' policy preferences are unimodally distributed (which, as noted above, is true in most Western democracies) valence-advantaged parties are motivated to position themselves near the center of the voter distribution, whereas valence-disadvantaged parties – particularly given proportional representation elections, where office-seeking parties seek to maximize seats and thus votes, even if they cannot win a popular plurality<sup>15</sup> – have centrifugal incentives to diverge from the centrist positions of their valence-advantaged rival(s). In two-party competition over one positional and one valence dimension, the valence-advantaged party can assure victory by locating at the median voter position (or even some distance from this position).

Figure 12.10a illustrates such a configuration for the strategic scenario pictured earlier in Figure 12.9b, where party A's valence advantage relative to party B is equivalent to two units of Left–Right position, i.e.,  $V(A/B)=+2$ , and where we assume that the median voter position is located at 5. Here we picture party A located at the median voter position, which forces party B to locate

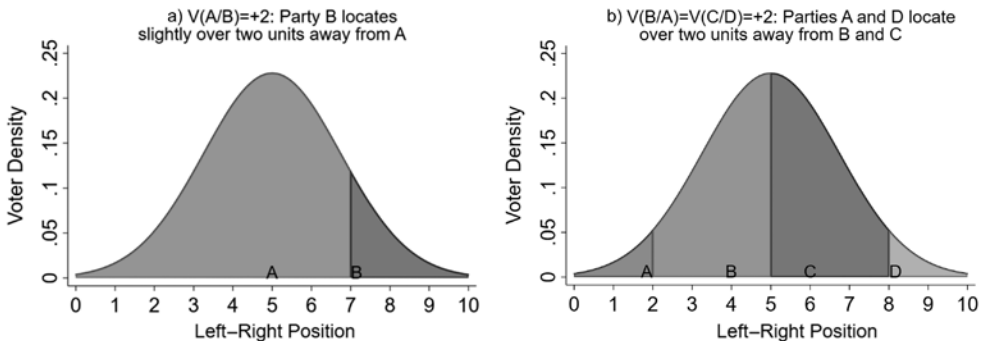


Figure 12.10 Party strategies in elections with one positional and one valence dimension



more than two units away from this position in order to attract any support at all; in this example, with **A** positioned at 5 and **B** positioned just to the right of 7, party **A** wins the election while attracting support from all voters located at or to the left of 7, while **B** wins support from voters located to the right of 7. Note, moreover, that in this example party **A** is assured of victory so long as its position is sufficiently moderate relative to the center of the voter distribution – in this case provided that **A** locates anywhere inside the Left–Right interval [3, 7] – and that regardless of **A**'s strategy, Party **B**'s optimal strategy will be to locate slightly more than two units away from **A**.

Figure 12.10b illustrates how valence considerations may play out in a four-party scenario involving parties **A**, **B**, **C**, **D**, where the two interior parties **B** and **C** each possess a two-unit valence advantage over their peripheral rival parties **A** and **D**, i.e.,  $V(B/A)=+2$   $V(C/D)=+2$ , and where we additionally assume that **B** and **C** have equal valences, i.e.,  $V(B/C)=0$ . With the valence-advantaged parties **B** and **C** located at the moderate positions 4 and 6, respectively, the peripheral parties **A** and **D** are no longer motivated to converge towards their ideological neighbors; instead they locate just over two units away from these rivals, near 2 and 8, respectively.<sup>16</sup>

## THE CONSEQUENCES OF POLICY-SEEKING PARTY MOTIVATIONS

To this point we have reviewed spatial models where parties single-mindedly seek political office. We have seen how, when we incorporate a valence dimension into the standard positional model, valence-disadvantaged parties may rationally present sharply noncentrist positions, away from their valence-advantaged opponents (and from the center of the voter distribution). While this logic illuminates the positional strategies of small, radical parties, there are many real-world

examples of large, valence-advantaged parties that also present sharply noncentrist positions – such as the British Conservatives under Margaret Thatcher and the Republicans under Ronald Reagan, discussed above. In fact, many contemporary democracies feature two large mainstream parties, one located on the center left and one located on the center right. This is true in two-party polities such as the United States, and for much of its history, the United Kingdom, as well as in many multiparty democracies, including Germany, the Netherlands, and Norway. Why might these large, mainstream parties take positions sharply different from each other and from the center of public opinion, particularly – in light of the median voter theorem – when they are the only parties present?

Donald Wittman (1973, 1977, 1983) suggested a possible solution to the above puzzle, which involved extending the model of party motivations. Wittman analyzed situations involving *policy-seeking* politicians who attach utilities to the policies that the winning party implements after the election. Wittman motivated this policy-seeking perspective by noting, first, that elected officials face pressures to implement the policies they promised during the election campaign, since to do otherwise would undermine the credibility of their promises in future elections. Second, Wittman observed that party elites – in common with rank-and-file voters – experience the ‘public good’ of government policy outputs. Wittman therefore analyzed the logic of party strategies when parties have preferences over the policies they are committed to implementing if they win office.

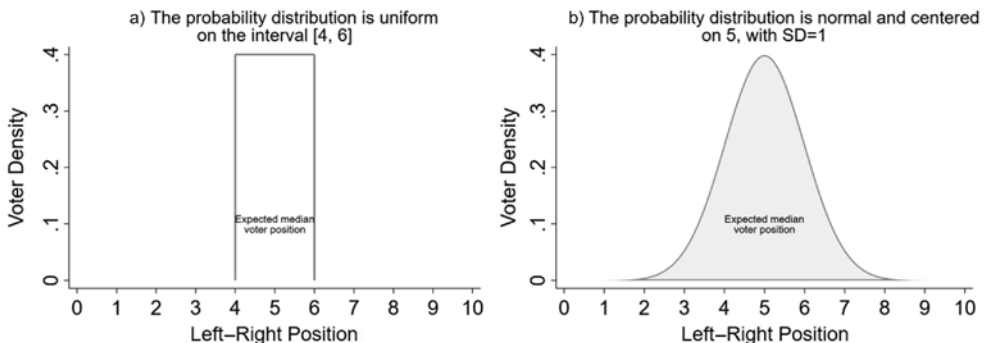
Spatial models with policy-seeking parties assume that each party, like each voter, has an ideal point which is the policy position it would prefer to implement. But this does not imply that a policy-seeking party – in attempting to optimize its policy expectations – should advocate its ideal point in elections. Party elites must still consider the electoral consequences of their policy promises, since

they must win office in order to implement these promises (and to prevent the implementation of disagreeable policies if a rival party wins). For a two-party election involving one positional dimension without valence, and where the rival parties' ideal points fall on opposite sides of the median voter's position, then, provided that party elites have perfect knowledge of this position, the positional spatial model implies that policy-seeking parties will converge towards the median voter's position – the same outcome as for office-seeking parties. To see this, note that when a party with sincere left-wing policy preferences relative to the median voter position competes against a party whose elites hold sincerely right-wing views, then if the left-wing party takes a position to the left of the median voter, the right-wing party need only choose a position to the right of the median – but nearer to that median voter – to win office and implement a policy which it prefers to its opponent's position. Since the same logic applies to the left-wing party's strategic reaction to any right-of-the-median position its opponent announces, it follows that the unique Nash equilibrium in policy-seeking party strategies is the median voter position.

But now let us assume, more realistically, that politicians *are not* certain of the median voter's location in advance of the election, where this uncertainty may reflect the limitations of public opinion polling or uncertainty

over voter turnout. Suppose, instead, that leaders of each party have a general idea of where the median voter should be located, but not a precise notion. For example, assuming a Left–Right scale from 0 (most left-wing) to 10 (most right-wing), party leaders might be quite confident that the median falls somewhere in the middle part of the scale, perhaps between 4.0 and 6.0, but be otherwise unsure of just where. In other words, they would represent their uncertainty about the median location by a uniform distribution between 4.0 and 6.0 (see Figure 12.11a). Or, perhaps more realistically, they might judge that the possible locations of the median voter follow a normal distribution, with, say, a mean of 5.0 and a standard deviation of 1.0 (Figure 12.11b).

We now analyze policy-seeking parties' strategies for this type of election-related uncertainty, using more technical mathematics than we have presented so far.<sup>17</sup> We will represent the party leaders' subjective notion of where the median voter may turn out to be located by a probability distribution with density function  $f(x)$  and cumulative distribution function  $F(x)$ .<sup>18</sup> For simplicity we assume that there are two parties, that this probability function on the median position is the same for both parties, and that voters are moved entirely by positional considerations. Note that the probability distribution of possible median voter positions is distinct from (and usually much narrower and peaked than) the underlying distribution of



**Figure 12.11** Examples of probability distributions over the median voter's position

voters themselves. We will see that, given uncertainty about the location of the median voter, the consequences for office-seeking and policy-seeking parties are different. To see this, first suppose that the two parties are both purely office-seeking. Let  $M_0$  denote the median of the subjective probability distribution  $f(x)$  that the party leaders estimate for the likely position of the median voter. (Note that  $M_0$  is a median of the voter medians, not the median of the (unknown) voter distribution.) See Figure 12.12, which depicts both the density of  $f(x)$  and its median  $M_0$ . Here we again picture the probability distribution  $f(x)$  on the median voter position as normal and centered on  $M_0=5$  with a one-unit standard deviation, as in Figure 12.11b.

Figure 12.12 also illustrates possible locations of each of two parties – one that prefers leftist policies that we will call **L**, located at  $L=4$ , and another that prefers right-wing policies that we label **R**, located at  $R=7$ . In this illustration, **L** is located nearer to  $M_0=5$  than is **R**. Note that party **L** will win the election if the actual median voter position, say  $m$ , turns out to be to the left of the midpoint  $(L+R)/2$  between the two party locations; in this example the midpoint is  $(4+7)/2=5.5$ . Party **R** will win if  $m$  turns out to be to the right of this midpoint. In fact the probability that **L** wins is equal to the cumulative probability from the extreme left up to the midpoint

$M_{LR}=(L+R)/2$ ; while the probability that **R** wins is 1 minus that cumulative probability, or in symbols,  $\Pr[L \text{ wins}] = F(M_{LR})$ , while  $\Pr[R \text{ wins}] = 1-F(M_{LR})$ . In this example the probability that **L** wins is about 0.69, and **R**'s election probability is about 0.31. Thus, unlike in the positional model with certainty over the median voter position, here the party with the more advantageous position (Party **L** in this example) is no longer certain to win election.

For a policy-seeking party, utility for a party, say **L**, is a sum of the party's valuation of the two possible outcomes of the election (either a win for **L** or a win for **R**), weighted by the respective probabilities of these outcomes. Therefore, assuming that **L** prefers leftist policies and **R** prefers rightist policies, utility of **L** for the outcome may be represented by

$$U_L = -L * F(M_{LR}) - R * [1 - F(M_{LR})]^{19}$$

and utility for **R** is given by

$$U_R = L * F(M_{LR}) + R * [1 - F(M_{LR})].$$

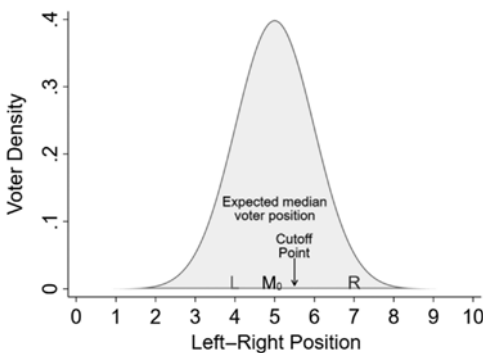
We are then able to show that a Nash equilibrium must be given by

$$R = \frac{1}{2f(M_0)}$$

and

$$L = -\frac{1}{2f(M_0)}.^{20}$$

This result tells us that at equilibrium the positions of policy-seeking parties experiencing uncertainty about the location of the median voter are separated by a distance that is related to the degree of spread (i.e., standard deviation) of the subjective median distribution, i.e., the separation at equilibrium increases with uncertainty about where the median voter may be located. For example, if the subjective notion of the median voter is normally distributed, then the equilibrium



**Figure 12.12** How party positioning affects parties' election prospects when there is uncertainty over the median voter position

separation is equal to approximately  $2.5\sigma$ , where  $\sigma$  is the standard deviation of the uncertainty distribution. Thus, on the 0–10 Left–Right scale, if  $\sigma=1$ , and  $M_0=5$  (as in Figure 12.12), then the policy-seeking parties **L** and **R** are separated by 2.5 units at equilibrium, i.e.,  $L=3.75$  and  $R=6.25$ .<sup>21</sup>

Adams and Merrill (2006) extend this model to a setting in which, in addition to the two major policy-seeking parties, there is a small, centrist third party that has no realistic chance of winning election. This model may apply to the British party system in recent decades, in which the Liberal Democrats ran on a platform that was often considered to lie between those of Labour on the left and the Conservatives on the right. By extending the arguments above, Adams and Merrill show that such a centrist party – if it itself prefers that a centrist policy be implemented – effectively shoots itself in the foot twice by entering the competition. This is because, first, the Nash equilibrium positions of the two major parties are even farther apart than would have been the case had the centrist party not competed, so that whichever is elected is in a position to implement an even more extreme policy, and second, the entry of the centrist party increases the likelihood that the major party that is farther from its preferred position is the one elected.<sup>22</sup>

Finally, we analyze policy-seeking parties' strategic incentives when voters are moved by valence as well as policy considerations, but the location of the median voter is assumed known with precision. In a number of spatial modeling studies, a key result that obtains for such models – with an important exception noted below – is that, in contrast to valence-advantaged parties' centripetal incentives in the office-seeking case, such policy-seeking parties typically have centrifugal incentives to announce non-centrist positions relative to the voter distribution (Londregan and Romer, 1993; Adams et al., 2005). To understand this dynamic, consider the case of positional spatial competition between a valence-advantaged

party **R** with sharply right-wing policy preferences, and party **L** with sincere policy preferences at or to the left of the median voter's position (known with certainty). In this scenario, party **R**'s valence advantage gives it leeway to diverge some distance to the right of the median voter position and still be assured of winning, with this degree of divergence increasing with the size of **R**'s valence advantage. Hence the unique Nash equilibrium in policy-seeking strategies is for **L** to locate at the median voter position while the valence-advantaged party **R** locates as near as is possible to its preferred right-wing position at the same time as, by leveraging its valence advantage, it still retains the median voter's support. Indeed, if **R**'s sincere policy preference is sufficiently moderate and/or its valence advantage sufficiently large, any configuration in which **R** locates at its preferred position is a Nash equilibrium.

The above intuition about the positional motivations of valence-advantaged, policy-seeking parties provides a plausible account of the empirical puzzle of the sharply non-centrist positional strategies of the British Conservatives under Margaret Thatcher, and the US Republicans under Ronald Reagan. Both parties benefited from valence advantages vis-à-vis their main competitors during much of the periods of these leaders' tenure in office – the Republicans' advantage was due largely to Reagan's image as a strong, charismatic leader; the Conservatives' advantage was because the Labour Party throughout the 1980s was plagued by public divisions and an image of weak leadership. Lending support to this policy-seeking perspective is the fact that both Reagan and Thatcher were widely viewed as 'conviction' politicians, who were unusually focused on pursuing their policy objectives. Londregan and Romer (1993) have delineated this spatial logic of valence-advantaged, policy-seeking parties in two-party elections, while Adams and Merrill (2009) extend this logic to multi-party elections.

## CONCLUSION

Beginning with Anthony Downs' pioneering work, research on the spatial model of elections has been extended from two-party to multiparty elections; from electorates whose voters are purely policy-focused to electorates that also weigh parties' character-based valence characteristics; from competition between office-seeking parties to elections where parties have policy motivations; and from competition between parties with complete information to elections where parties experience uncertainty about the distribution of the voters' ideal points. These extensions are intended to capture real-world election contexts, and to explain why actual political parties and candidates rarely converge to identical, centrist policies – the prediction associated with the basic Downsian model of two-party, one-dimensional, positional competition.

The extensions reviewed here by no means exhaust the variations on the basic Downsian model. In particular, a growing literature analyzes the implications of 'two-stage' candidate elections in which office-seeking candidates must first win a party primary election in order to advance to the general election (for example, Owen and Grofman, 2006; Serra, 2010). In addition, Adams et al. (2005) develop an approach that unifies the Downsian positional spatial model and the behavioral voting model associated with the University of Michigan (see Campbell et al., 1960), which emphasizes the importance of voter party identification as a long-term, affective orientation, and the authors show that parties may have electoral incentives to appeal on policy grounds to their pre-existing partisans. Since voter partisanship correlates strongly with policy preference in most real-world electorates, this implies that rival parties have electoral motivations to present dispersed policies, with each party taking positions that reflect their long-term partisans' beliefs. Finally, Curini (2018) integrates valence and positional concerns into a spatial model where parties strategize over

whether to raise valence-related corruption issues during election campaigns, showing that this strategic decision may depend on the parties' locations along positional dimensions of competition relative to their opponents' positions. Curini presents a variety of theoretical and empirical analyses (the latter derived from analyses of party programs, politicians' legislative speeches and social media data) to show how this approach illuminates the rise of negative campaigning in contemporary democracies, with particular emphasis on the strategies of emerging populist parties. These ongoing research agendas illustrate how the Downsian model can accommodate theoretically interesting, empirically realistic variations in real-world election contexts.

## Notes

- 1 We frame our discussion in terms of political parties, even though in some countries, citizens cast votes for individual candidates, not parties. An extensive literature documents that even in candidate-based electoral systems, citizens' vote choices are strongly influenced by the candidates' party affiliations (see, e.g., Dalton et al., 2011).
- 2 We realize that the assumption of common valence for a party among all voters may sometimes be unreasonable. The degree to which different voters value honesty or competence may vary, and voters may disagree in their assessments of the honesty or competence of a party or its leader (see, e.g., Zakharova and Warwick, 2014). Nevertheless, there is plausibly more voter agreement over what are desirable character-based valence attributes than over desirable Left-Right positions.
- 3 We note that some researchers apply the term 'valence' to parties' reputations for successfully addressing specific issues such as education, crime, and so on (see, e.g., Clarke et al., 2009; Green and Jennings, 2012, 2017).
- 4 Technically, this position is unique only if the number of voters is odd, but with a very large electorate it is essentially unique in any case. Furthermore, for a large electorate, any significant movement by either party can be expected to cross over the positions of some of the voters.
- 5 We note, however, that predicted party convergence breaks down under various extensions of the Downsian model, including that citizens

- abstain from voting if neither party offers a sufficiently attractive position (Adams and Merrill, 2003); that parties seek to deter entry by new parties (Palfrey, 1984); that political parties select their candidates through primary elections (Owen and Grofman, 2006); and many others. Grofman (2004) reviews these explanations.
- 6 As we discuss below, the arguable exception to this generalization is the British post-World War II party system, which prior to the 1980s was similar to the United States in that it featured two dominant political parties: the Conservatives and Labour.
  - 7 The figure illustrates a scenario in which the voters' ideal points are normally distributed, but the logic extends to any type of voter distribution.
  - 8 Of course, the Downsian model of party convergence applies to two-party competition. A similar logic, as we have seen, does not imply such convergence with three or more parties. Nevertheless, a Downsian expectation of centripetal tendency has been applied to the multiparty scenario.
  - 9 British party politics has subsequently diverged from a system of two dominant parties, as the Liberals (and in later periods the Liberal Democrats) emerged as a third competitive party. However, since the 2015 elections, the Liberal Democrats have become less competitive due to the rapid deterioration of their reputation for competence and integrity (Zur, 2019).
  - 10 These respondent Left–Right self-placements and party placements are based on the following survey questions: 'In politics people sometimes talk of left and right. Where would you place yourself on a scale from 0 to 10 where 0 means the left and 10 means the right? Where would you place [PARTY NAME] on this scale?'
  - 11 The figures display the mean placements of all parties that won at least 2% of the national vote. The respondents' self-placements and their party placements in these national election studies were based on the same questions reported in note 10 above.
  - 12 We note that alternative measures of parties' Left–Right positions, such as those based on political experts' party placements from the Chapel Hill Expert Survey (Bakker et al., 2015), or based on content analyses of parties' election manifestos (Volkens et al., 2018), support the same substantive conclusion.
  - 13 We assume here that as a party's position diverges from the voter's Left–Right position the voter's utility for the party declines at a constant rate, i.e., voters have linear loss functions. There are other utility loss functions we could consider, but these would complicate our discussion.
  - 14 In addition, Aragonés and Palfrey (2002) develop a spatial model with uncertainty about the location of the median voter, where one party may enjoy an (unspecified) advantage, which appears equivalent to the valence advantages we discuss here, with probabilistic divergence and uncertainty about the winner.
  - 15 Most western democracies feature some form of proportional, multi-member district system to select representatives to the national parliament. The major alternative to proportional representation is the plurality-based, single-member district voting system which is employed in France along with most of the English-speaking democracies.
  - 16 We note that the four-party configuration displayed in Figure 12.10b is not a Nash equilibrium since the interior parties **B** and **C**, who are equally matched with each other on valence, can each increase their support by unilaterally shifting position closer to the center of the voter distribution.
  - 17 Calvert (1985), in a setting with multiple policy dimensions, analyzes these issues and shows that candidate policy-seeking alone does not induce divergence; nor does candidate uncertainty about voter response alone (as long as weak assumptions are made), but both policy-seeking and uncertainty together prompt divergence, although small departures from the classic model lead to small levels of divergence.
  - 18 The cumulative probability  $F(x)$  (from  $-\infty$  to  $x$ ) is given by  $\int_{-\infty}^x f(t) dt$ .
  - 19 Here we assume that each party has a linear policy loss function, i.e., that their utilities for various positions decrease at a constant rate as the position diverges from the party's ideal point. The negative signs for the terms in the formula for  $U_L$  occur because **L** prefers a more negative policy while **R** prefers a more positive one. We have also omitted additive constants, which drop out when derivatives are taken.
  - 20 To see this, note that at a Nash equilibrium,
 
$$\partial U_L / \partial L = -F(M_{LR}) + (1/2)(R - L)f(M_{LR}) = 0,$$
 and similarly,
 
$$\partial U_R / \partial R = 1 - F(M_{LR}) - (1/2)(R - L)f(M_{LR}) = 0.$$
 Adding and subtracting these two equations, we obtain  $2F(M_{LR}) = 1$  and  $1 = (R - L)f(M_{LR})$ . From the first of these latter equations, we conclude that  $M_0 = M_{LR} = (L + R) / 2$ , and from the other equation, we infer that  $R - L = 1/f(M_0)$ , i.e., that
 
$$R = \frac{1}{2f(M_0)} \text{ and } L = -\frac{1}{2f(M_0)}. \text{ Q.E.D.}$$

- 21 This equilibrium configuration obtains provided that each party's sincerely preferred policy output is at least as extreme as its equilibrium position.
- 22 Merrill and Grofman (2019) consider a mirror image of this problem – namely, how should policy-seeking mainstream parties react when an extreme third party enters on the flank of one of them. In this setting, they determine conditions such that – just as in the face of entry of a centrist party – both mainstream parties should move further from the new entry.

## REFERENCES

- Abney, Ronni, James Adams, Michael Clark, Malcolm Easton, Lawrence Ezrow, Spyros Kosmidis and Anja Neundorf (2013). 'When does valence matter? Heightened valence effects for governing parties during election campaigns.' *Party Politics* 19(1): 61–82.
- Adams, James (2016). 'Competing for votes' in Jac C. Heckelman and Nicholas R. Miller, eds, *Elgar Handbook of Social Choice and Voting*. Cheltenham: Edward Elgar Publishing, pp. 201–217.
- Adams, James F. (2018). 'Spatial voting models of party competition in two dimensions' in Roger D. Congleton, Bernard N. Grofman, and Stefan Voigt, eds. *The Oxford Handbook of Public Choice*. Vol. 1. Oxford University Press, 2019, pp. 187–207.
- Adams, James, and Samuel Merrill III (2003). 'Voter turnout and candidate strategies in American elections.' *Journal of Politics* 65(1): 161–189.
- Adams, James, and Samuel Merrill III (2006). 'Why small, centrist third parties motivate policy divergence by major parties.' *American Political Science Review* 100(3): 403–417.
- Adams, James, and Samuel Merrill III (2009). 'Policy-seeking parties in a parliamentary democracy with proportional representation: a valence-uncertainty model.' *British Journal of Political Science* 39(3): 539–558.
- Adams, James, Jane Green and Caitlin Milazzo (2012). 'Has the British public depolarized along with political elites? An American perspective on British public opinion.' *Comparative Political Studies* 45(4): 507–530.
- Adams, James, Samuel Merrill III and Bernard Grofman (2005). *A Unified Theory of Party Competition: A Cross-National Analysis Integrating Spatial and Behavioral Factors*. Cambridge: Cambridge University Press.
- Ansolabehere, Stephen, and James Snyder (2000). 'Valence politics and equilibrium in spatial economic models.' *Public Choice* 103(3–4): 327–36.
- Aragones, Enriqueta, and Thomas R. Palfrey (2002). 'Mixed equilibrium in a Downsian model with a favored candidate.' *Journal of Economic Theory* 103(1): 131–161.
- Bakker, Ryan, Catherine de Vries, Erica Edwards, Liesbet Hooghe, Seth Jolly, Gary Marks, Jonathan Polk, Jan Rovny, Marco Steenbergen and Milada Anna Vachudova (2015). 'Measuring party positions in Europe: the Chapel Hill expert survey trend file, 1999–2010.' *Party Politics* 21(1): 143–152.
- Black, Duncan (1948). 'On the rationale of group decision-making.' *Journal of Political Economy* 56(1): 23–34.
- Bølstad, Jørgen, and Elias Dinas (2017). 'A categorization theory of spatial voting: how the center divides the political space.' *British Journal of Political Science* 47(4): 829–850.
- Campbell, Angus, Philip E. Converse, Warren E. Miller and Donald E. Stokes (1960). *The American Voter*. Chicago and London: The University of Chicago Press.
- Calvert, Randall L. (1985). 'Robustness of the multidimensional voting model: candidates, motivations, uncertainty, and convergence.' *American Journal of Political Science* 29(1), 69–95.
- Clark, Michael (2009). 'Valence and electoral outcomes in Western Europe, 1976–1998.' *Electoral Studies* 28(1): 111–122.
- Clarke, Harold D., David Sanders, Marianne C. Stewart and Paul F. Whiteley (2009). *Performance Politics and the British Voter*. Cambridge: Cambridge University Press.
- Cox, Gary W. (1990). 'Centripetal and centrifugal incentives in electoral systems.' *American Journal of Political Science* 34(4): 903–935.
- Curini, Luigi (2018). *Corruption, Ideology, and Populism: The Rise of Valence Political Campaigning*. London: Palgrave Macmillan.
- Curini, Luigi, and Paolo Martelli (2015). 'A case of valence competition in elections: Parties' emphasis on corruption in electoral manifestos.' *Party Politics* 21(5): 686–698.
- Dalton, Russell J. (2013). *Citizen Politics: Public Opinion and Political Parties in Advanced Industrial Democracies, 6th edition*. Thousand Oaks (CA): Congressional Quarterly Press.
- Dalton, Russell J., David M. Farrell and Ian McAllister (2011). *Political Parties and*

- Democratic Linkage: How Parties Organize Democracy*. Oxford: Oxford University Press.
- Dalton, Russell J., and Ian McAllister (2015). 'Random walk or planned excursion? Continuity and change in the left-right positions of political parties.' *Comparative Political Studies* 48(6): 759–787.
- Downs, Anthony (1957). *An Economic Theory of Democracy*. New York: Wiley.
- Eaton, B. Curtis, and Richard G. Lipsey (1975). 'The principle of minimum differentiation reconsidered: some new developments in the theory of spatial competition.' *The Review of Economic Studies* 4(1): 27–49.
- Feld, Scott L., and Bernard Grofman (1991). 'Incumbency advantage, voter loyalty and the benefit of the doubt.' *Journal of Theoretical Politics* 3(2): 115–137.
- Green, Jane, and Will Jennings (2012). 'Valence as macro-competence: an analysis of mood in party competence evaluations in Great Britain.' *British Journal of Political Science* 42(2): 311–343.
- Green, Jane, and Will Jennings (2017). *The Politics of Competence: Parties, Public Opinion and Voters*. Cambridge: Cambridge University Press.
- Grofman, Bernard (2004). 'Downs and two-party convergence.' *Annual Review of Political Science* 7: 25–46.
- Hotelling, Harold (1929). 'Stability in competition.' *Economic Journal* 39(153): 41–57.
- Kitschelt, Herbert (1994). *The Transformation of European Social Democracy*. New York: Cambridge University Press.
- Londregan, John and Thomas Romer (1993). 'Polarization, incumbency, and the personal vote' in William A. Barnett, Melvin Hinich and Norman Schofield (eds), *Political Economy: Institutions, Competition, and Representation*, New York: Cambridge University Press, pp. 355–377.
- Marks, Gary, Lisbet, Hooghe, Moira Nelson and Erica Edwards (2006). 'Party competition and European integration in east and west: different structure, same causality.' *Comparative Political Studies*, 39(2): 155–175.
- McCarty, Nolan, Keith T. Poole and Howard Rosenthal (2006). *Polarized America: The Dance of Political Ideology and Unequal Riches*. Cambridge, MA: MIT Press.
- Merrill, Samuel, and Bernard Grofman (2019). 'What are the effects of entry of new extremist parties on the policy platforms of mainstream parties?' *Journal of Theoretical Politics* 31(3): 453–473.
- Norton, Philip (2000). *The British Polity*, 4th edition. London: Pearson Education.
- Owen, Guillermo, and Bernard Grofman (2006). 'Two-stage electoral competition in two-party contests: persistent divergence of party positions.' *Social Choice and Welfare* 26(4): 547–569.
- Palfrey, Thomas R. (1984). 'Spatial equilibrium with entry.' *Review of Economic Studies* 51(1): 139–156.
- Powell, G. Bingham (2000). *Elections as Instruments of Democracy: Majoritarian and Proportional Visions*. New Haven, CT: Yale University Press.
- Serra, Gilles (2010). 'Polarization of what? A model of elections with endogenous valence.' *Journal of Politics* 72(2): 426–437.
- Stokes, Donald (1963). 'Spatial models of party competition.' *American Political Science Review* 57(2), 368–377.
- Stokes, Donald (1992). 'Valence politics' in Dennis Kavanagh, ed., *Electoral Politics*. Oxford: Clarendon Press, pp. 141–162.
- Volkens, Andrea, Pola Lehmann, Theres Mattheiß, Nicolas Merz, Sven Regel and Bernhard Weißels (2018). The Manifesto Data Collection. Manifesto Project (MRG/CMP/MARPOR). Version 2018a. Berlin: Wissenschaftszentrum Berlin für Sozialforschung (WZB). <https://doi.org/10.25522/manifesto.mpd.2018a>
- Wittman, Donald (1973). 'Parties as utility maximizers.' *American Political Science Review* 67(2): 490–498.
- Wittman, Donald (1977). 'Candidates with policy preferences: a dynamic model.' *Journal of Economic Theory* 14(1): 180–189.
- Wittman, Donald (1983). 'Candidate motivation: a synthesis of alternatives.' *American Political Science Review* 77(1): 142–157.
- Zakharova, Maria, and Paul V. Warwick. 2014. 'The sources of valence judgments: the role of policy distance and the structure of the left–right spectrum.' *Comparative Political Studies* 47(14): 2000–2025.
- Zur, Roi (2019). 'Stuck in the middle: ideology, valence and the electoral failures of centrist parties.' *British Journal of Political Science*, Online First: 1–18. doi:10.1017/S0007123419000231.



# New Directions in Veto Bargaining: Message Legislation, Virtue Signaling, and Electoral Accountability

Charles Cameron and Nathan Gibson

## INTRODUCTION

In the years since the creation of separation-of-powers (SOP) models – aimed first at courts,<sup>1</sup> then at Congress,<sup>2</sup> and finally at presidents<sup>3</sup> – much has changed though much remains the same. Needless to say, the constitutionally mandated architecture of the American government hasn't changed at all. This architecture separates the three branches and forces them to interact through a structured bargaining process of proposals and vetoes. On the other hand, the coalition structure of the political parties, the participants in politician selection and the media environment have all changed, arguably dramatically.<sup>4</sup> The causal linkages remain disputed but the net effects are striking and manifest to all: elite partisan polarization, political rancor, congressional stasis, aggressive presidential unilateralism, and puissant courts.<sup>5</sup> In the new American politics, policy outcomes are generally quite understandable using the classic SOP models, or so we assert. But

much of the action, the sound and fury of daily politics, is quite mysterious and clearly beyond the ambit of those simple frameworks. Examples include repeated fruitless attempts to pass doomed bills, hopeless vetoes, futile filibusters, lop-sided cloture votes, obviously doomed attempts at bicameral legislating, hostage-taking via government shut-downs, manifestly impossible impeachment attempts, ostentatiously illegal executive orders and more.

In this chapter we focus on the mysterious, and we offer some suggestions on how to make the murky more transparent.

We begin with a brief review of the classic separation-of-powers (SOP) models, focusing on the veto bargaining version but noting easy extensions to the filibuster. We emphasize the use of incomplete information models to study not just outcomes but process. We are terse because handy and more elaborate reviews are available elsewhere.<sup>6</sup> Then, we note the rise of several puzzling empirical phenomena. These include bizarre

vote margins on vetoed bills and during over-ride attempts; similarly weird vote margins for filibustered bills and during cloture votes; and the useless re-passage, many times, of virtually the same doomed legislation. (If space allowed, we would add more from the laundry list above.) We trace much of these phenomena to a single cause: *the desire of political agents to send credible signals to political principals about their dedication and ideological fealty, using the policymaking procedures of the SOP system*. In other words, they are variants on or consequences of what congressional scholar Frances Lee, in a seminal contribution, called ‘message legislation’ in the lawmaking context.<sup>7</sup> We dub this phenomenon ‘virtue signaling’. Virtue signaling is closely related to, complementary of, but distinct from, blame game politics.

With one exception – Groseclose and McCarty’s prescient explication of ‘blame game vetoes’ – the first-generation SOP models did not accommodate, and say nothing about, message-oriented manipulation of the SOP system’s policymaking procedures.<sup>8</sup> Instead, they assume serious policy-minded actors who pursue genuine policy goals by bargaining with one other in a straightforward and serious way. Even the blame game veto model, which takes a big step away from this paradigm, does not fully capture the new direction in American legislative politics. We assert, however, that if the SOP models are suitably modified, then new veto bargaining, pivotal politics and related models *can* make sense of the novel phenomena while retaining their broad accuracy about policy outcomes.

The trick (in our view) is to move beyond the first-generation framework by embedding the SOP games within what is now called an ‘accountability’ model of elections.<sup>9</sup> In other words: situate the SOP game in a larger model that features retrospective voting or similar action by political principals. The principals we have in mind are the high-information ‘base’ or ‘selectorate’, that is, the individuals who are critical in candidate

recruitment, fund-raising, participation in primaries, campaigning and turnout.<sup>10</sup> Without the enthusiastic support of these individuals, a member of Congress or president is in serious electoral peril. Furthermore, the selectorate will be enthusiastic only about politicians who, if circumstances permit, are willing to work hard to enact the base’s policy agenda. That programmatic agenda is, in contrast to the typically muddled and inchoate desires of less engaged citizens, usually quite definite in some particulars. Politicians’ seemingly bizarre SOP manipulations, such as fruitlessly repealing portions of the Affordable Care Act dozens of times in a legislatively hopeless configuration, can be seen as rational efforts to prove to their skeptical ‘boss’ that they are indeed the type who will bring home the policy bacon should circumstances permit in the future. And demanding such signaling is actually rational for a boss who is doubtful whether the agent possesses ‘true grit’.

To illustrate these points, we sketch a simple model that embeds a stripped-down veto bargaining game within a simple accountability model (we do not undertake a genuine formal analysis here; our discussion is merely illustrative). We hope these notes-to-a-theory suggest the potential for a new direction for SOP models.

We conclude with some observations about whether the sound and fury of phony legislating actually makes a substantive difference or is just meaningless political theater. Our simple new-style SOP model suggests it does make a difference.

## CLASSICAL VETO BARGAINING GAMES

The classical models feature bilateral bargaining between a policy proposer, Congress, or C, and a policy receiver, the President, P. Also making an appearance is the veto over-ride player, O. This player is defined as the legislator closest to the president for whom exactly

one third of the legislature has ideal points either lower or higher than the over-ride player's, depending on whether the president's ideal point  $p$  (defined momentarily) lies in the left or right portion of the policy space, respectively. In some versions another player, the filibuster pivot, appears as well. The filibuster pivot is defined similarly but only for the Senate and using 40 members (the threshold for cloture since 1975), most relevantly on the opposite side of the median from the President. The policy space is typically assumed to be one dimensional. So it is a policy evaluation space similar to the oft-used NOMINATE space in empirical studies of roll-call voting.<sup>11</sup> A critical point in the policy space is the current policy, the status quo, denoted  $q$ .

Each actor has a policy utility function defined over the policy space, with a well-identified most preferred policy, the ideal point. Call these ideal policies  $c$  and  $p$ , for Congress and President respectively, and that of veto over-ride player  $o$ . Policies increasing far from the ideal point have declining value. An example of such a utility function is the 'tent' utility function:

$$u(x, x_i) = -|x - x_i|$$

where  $x_i$  is player  $i$ 's most-preferred policy (e.g.,  $c$ ,  $p$ , and  $o$ ) and  $x$  is any policy in the policy space.

This simple apparatus was first developed to study elections and voting.<sup>12</sup> The SOP policymaking models take the apparatus in a somewhat different direction, however.

### ***The Engine: the One-shot Take-It-or-Leave-It (TILI) Bargaining Game***

The engine that makes the SOP models run is the celebrated one-shot take-it-or-leave-it (TILI) bargaining game first analyzed by Romer and Rosenthal (1978). Most of the SOP models, including veto bargaining, just make changes to this model, for example, by adding more moves, additional institutional

actors and incomplete information. In its simplest form, the sequence of play in TILI bargaining is:

- 1 C makes a proposal  $b$  (a 'bill') to change the status quo or reversion policy  $q$ .
- 2 P accepts or vetoes the offer. If P accepts the offer, the final policy outcome  $x$  is the bill  $b$ , and the game ends.
- 3 If P vetoes the offer, a vote on a motion to over-ride occurs. If O supports the motion, the bill is successful and again  $x = b$  is the new policy. If O does not support the motion, the bill fails and  $x = q$ , so the status quo remains the policy in effect.

Because the game features complete and perfect information, it is easily solved using backward induction, thereby incorporating the idea of forward-thinking strategically minded actors. The resulting subgame perfect equilibrium is unique, depending only on the configuration of ideal points and the location of the status quo. We will not go into any of the details since very clear expositions are readily available. However, several points are worth noting. The first three are substantive; the last two are theoretical.

First, the basic model reveals a *prominent advantage for Congress relative to the President*. The presidential veto acts as a check on congressional power, but Congress's ability to force an unamendable offer on a president who can only say 'yes' or 'no' (and who might not be able to make 'no' stick) gives a huge, constitutionally entrenched power advantage to Congress.

Second, given much policy disagreement between the legislature and the executive or across the parties, *moving the status quo usually requires supermajorities* in the legislature. Given the Constitution's veto override provisions and the Senate's privileging of the filibuster, this should hardly be a surprise. But it is a point of major historical importance – almost every piece of important legislation in the post-World War II era was enacted through supermajorities.<sup>13</sup> It also implies that the American rules of the policymaking game force narrow coalitions

of extremists to compromise if they are to accomplish anything legislatively. Moderates will see this as highly desirable normatively; passionate extremists will see it as a bug, not a feature, of American government.

Third, (and related to the second point), *often no policy movement is possible*: the status quo lies in the so-called gridlock region. In fact, the model and its variants supply the causal mechanisms behind the *status quo bias* so characteristic of American politics. We all know that status quo bias exists because there are so many choke points in the policy process. The models go beyond this cliché to show exactly how the choke points work to create policy gridlock.

Fourth, because the core model is so simple and easy to analyze, *the analysis is very extendable*. This is a lovely feature for the theoretically inclined. For example, one can add congressional committees with gate-keeping power;<sup>14</sup> filibusters and cloture votes;<sup>15</sup> a powerful Speaker of the House with gate-keeping power;<sup>16</sup> agencies that begin the game by setting a policy via regulation, so the model becomes a model of the administrative state in action;<sup>17</sup> presidents who move first via an executive order, so the model illustrates presidential unilateral action;<sup>18</sup> and more. With very simple tools requiring minimal mathematical ability, one can easily see how a great deal of national policymaking works.

The fifth point is subtle and deep and not easy to grasp on first acquaintance. In complete information models of the kind we have been discussing, policy typically moves quickly to its final resting place. There are no vetoes, over-rides, filibusters or cloture votes along the path of play; policy just adjusts. If no movement is possible, nothing happens at all. In this sense, the modern analysis of vetoes and filibusters is similar to modern analyses of wars, litigation and strikes. Complete information models of those phenomena predict changes in territorial boundaries, cross-litigant payments and wages. But they also predict no wars, no

strikes, no strikes. The reason is that the participants understand perfectly what ultimate adjustments will happen and therefore reach agreements that obviate uselessly destructive conflict. In order to get actual vetoes, filibusters, wars, trials, strikes, and so on, a model *requires a degree of incomplete information*. In other words, some actor must lack knowledge about an important variable, and this ignorance or uncertainty leads to 'mistakes' (more accurately, the rational calculations of the actors lead one or both to insist on obdurate actions that would not occur if everyone knew everything).

This fundamental point about human interactions is often met with skepticism: do you mean to say the horrific slaughter in the trenches of World War I (for instance) was caused by a lack of information, not nationalism, militarism, military technology, age-old hatreds, and so on and so on? Not exactly: nationalism and so on may have been *necessary* for the conflict, in the same way that policy disagreement is necessary for a veto. But nationalism was not *sufficient*. It took nationalism *plus* incomplete information to produce the tragic slaughter. Similarly, in SOP models, it takes policy disagreement *plus* incomplete information to produce a veto, a filibuster, an over-ride attempt, a cloture vote, a judicial strike-down of an executive order, a congressional reversal of a judicial policy, and so on.

There is a logical corollary: analysts who want to study not just policy outcomes but phenomena such as vetoes, filibusters, cloture votes and over-ride attempts need to use models that incorporate incomplete information.

### ***Bilateral Bargaining under Incomplete Information***

Early analysts of separation-of-powers politics moved to do just that. McCarty (1997), for example, studied how a president can use vetoes to build a reputation across different

policy arenas over time. This model affords one explanation for the well-known honeymoon effect in presidential–congressional relations (Congress, knowing the freshman president is hungry to build a reputation for toughness, is extremely accommodating – at least at first).

Cameron (2000) explored a model of sequential veto bargaining. Here, Congress and President go through multiple rounds of passing and vetoing the same bill, with Congress making concessions each time in an effort to produce an offer the President will accept, and the President vetoing and re-vetoing in a gamble that Congress will return with a better offer before bargaining breaks down. Perhaps somewhat surprisingly, some of the most consequential legislation of the 20th century emerged from this sequential bargaining process (e.g., welfare reform under Clinton).

Cameron (2000) also offered a very simple model of over-ride attempts. Here, in the face of uncertainty about who the critical veto over-ride player will be at the actual moment of the attempt, over-rides can occur, both successful and unsuccessful. Essentially the same model could be used to study filibusters and cloture voting.

In a particularly clever model, Matthews (1989) studied veto threats. Here, a veto threat is a little like a bid in a poker game: the President opens with a ‘bid’ (a veto threat), Congress may or may not adjust its next ‘bid’ (a bill) and then the President ‘calls’ or ‘folds’ by vetoing or accepting.<sup>19</sup> Cameron et al. (2000) take this model to data, which generally display the predicted empirical patterns.

All of these models feature bilateral bargaining between the President and Congress with uncertainty about one of the player’s preferences. In most cases, the uncertainty involves the President’s preferences, although in the simple veto over-ride model the uncertainty is about the preferences of the over-ride player. Models in which the unknown-preference President moves before

a move or counter-move by Congress are signaling games.<sup>20</sup> These games feature strategic reputation-building and require more sophisticated modes of analysis than the simple complete information models (one must model player beliefs simultaneously with player strategies, and the two must reinforce one another).

Many of the incomplete information bilateral bargaining models make rather precise empirical predictions about vetoes, overrides and so on. Data from the mid 20th century (or earlier) through to the 1980s or so strongly display the predicted patterns. As a result, this analytical endeavor has often been seen as a success for the empirical implications of theoretical models (EITM) movement in political science. Critically, however, some of the key predictions of the incomplete information bilateral bargaining models show signs of breaking down – a point we return to below.

### ***Bargaining before an Audience: Message Votes***

One of the early incomplete information models stands out from the others, because it is not a bilateral bargaining game. We refer to Groseclose and McCarty’s blame game veto model (2001). This model involves three intrinsically important players. Specifically, Congress and the President play a legislative game before an audience, a Voter. The President and Congress understand each other’s preferences perfectly, so there is no incomplete information at that point. But the Voter is somewhat uncertain about the President’s preferences; therein lies the critical incomplete information. The Voter’s uncertainty creates the opportunity for Congress to set up a policymaking sequence which, if observed by the Voter, will lead her to draw a relatively unfavorable inference about the President’s preferences (even knowing that Congress would like this to happen). And that is the whole point – not truly

legislating, but play-acting legislating in order to cast blame on the other side. Indeed, the veto-bait bill may fail miserably in enactment but still succeed as symbolic action.

The ideas in the Groseclose–McCarty model should resonate with contemporary scholars, for blame game vetoes are closely related to what Frances Lee has called ‘message votes’. According to Lee (2016: 143–4), message votes occur when

a party brings to the floor an attractive-sounding idea with the following characteristics: (1) its members support it; (2) the other party opposes it; and (3) it is not expected to become law. Former Senator Olympia Snowe offers a more detailed explanation: ‘much of what occurs in Congress today is what is often called “political messaging”. Rather than putting forward a plausible, realistic solution to a problem, members on both sides offer legislation that is designed to make the opposing side look bad on an issue and it is not intended to ever actually pass.’

The Groseclose–McCarty model works out the logic of ‘mak[ing] the opposing side look bad’ in the specific context of the presidential veto.<sup>21</sup>

An obvious question is, how frequently have blame game vetoes occurred? We take a look at some relevant data below. But Cameron (2000a) addressed this question over the 20th century, using an admittedly stringent set of criteria: the veto needed to be prominent, occur in the run-up to a presidential election, and led to a hopeless over-ride attempt (so the enactors should have known that serious legislating was off the table). The historical data on vetoes during the 20th century uncovers relatively few blame game vetoes, according to these criteria (see *ibid.*, table 5.1). Most vetoes did not look like this. To the extent that this is a fair test, the blame game model does not look like a general model of vetoes, at least over much of the 20th century. However, the data reveal that *some vetoes were clearly blame game vetoes*. An example was the Family and Medical Leave Act of 1991, passed by a Democratic Congress and presented to Republican

President George H. W. Bush immediately before the 1992 presidential election. Bush had publicly opposed the bill and his veto was entirely predictable. Given the vote margins, a successful over-ride was clearly doomed. So from a serious legislating perspective, the bill was futile. The Democrats nonetheless pressed ahead, and then used the failed bill as a signature electoral issue. Upon re-gaining the presidency, they quickly enacted family leave in 1993 and touted it as a flagship legislative accomplishment. Quintessential blame game politics!

The general phenomenon of blame game politics, presciently explored by Groseclose and McCarty in the specific context of veto bargaining, has now become routine, at least in the opinion of astute observers such as Lee and candid participants such as Snowe. In fact, a series of empirical anomalies in separation-of-power politics suggest the need for some fresh thinking.

## EMPIRICAL ANOMALIES

Recent years have seen congressional legislative behavior that is extremely difficult to reconcile with the classical SOP models. Let’s look at some of the empirical anomalies.

### ***What to Look for: Vote Margins at the Pivots and Policy Concessions***

The first question, though, is this: where should we look for legislative anomalies? The incomplete information bilateral bargaining models assume a degree of uncertainty about the preferences of a key player, but not a huge amount of incomplete information. This has important implications for vote margins at the pivots and for policy concessions in re-passed bills.

First, *vote margins at the critical pivots should be close*. To see the logic, suppose, for example, a bill is geared to beat a likely

presidential veto with the veto over-ride player as the critical pivot. Then the roll call margin on passage in both chambers should be about two-thirds. If it is much higher, the proposers have not been tough enough; they have conceded too much. If it is far shy of two-thirds then the bill is a sitting duck, doomed from day one, and the proposers are wasting their time. The margin for the over-ride attempt should also be about two-thirds. Now, suppose the president himself is the critical pivot (that is, the veto over-ride player is more extreme than the president). Then the passage margin may be lower than two-thirds but if the president does veto the bill, no over-ride attempt should follow, as the over-ride is hopeless. If an over-ride attempt did occur (anomalously), the vote margin would be well short of two-thirds. In short, unless the president is moderate relative to the over-ride pivot, passage margins for vetoed bills should be about two-third yeas and one-third nays, over-ride attempts should not occur for vetoed bills with narrow passage margins and actual over-ride margins should be about two-thirds yeas and one-third nays.

Similar ideas apply to filibusters. Suppose a bill is geared to beat a filibuster in the Senate. Then a bill that is likely to provoke a filibuster should pass the Senate with about 60 votes. If it passed with many more votes, the filibuster is pointless since cloture will be easy, hence no filibuster should occur (and the bill's proponents conceded too much to the opposition). If initially passed with a narrow majority, then cloture seems likely to fail and the bill should not have been passed in the first place – its authors should have conceded more, or just abandoned the effort. Similarly, actual cloture votes should show about 60 votes in favor of cloture. Lop-sided successful cloture votes should not occur because the filibusterers should have known they would fail; lop-sided failed cloture votes should not occur because the bill authors should have known the bill was a sitting duck and either conceded more or given up the cloture attempt.

A second anomaly can occur with re-passed, previously failed legislation: *no concessions*. (That is, for re-passage under the same configuration of players.) Under the sequential veto bargaining model, re-passage of vetoed bills can occur, but the re-passed bill should contain a compromise in the direction of the president, so either he will sign it or the veto over-ride player will support the bill. As a result, the cutting line between the yeas and nays in NOMINATE space should shift toward the president, and the aye margin should increase.<sup>22</sup> Similar logic applies to bills that die from a filibuster in the Senate: if re-passed, they should contain a compromise to the filibusterers so that either they will accept it or cloture will succeed. The same logic also applies to bills that are enacted by one chamber during split-chamber divided government, but then die in the other chamber (perhaps they are never taken up). If the first chamber re-passes the bill, it should contain concessions to the recalcitrant chamber. Cutting lines for the roll call vote in the enacting chamber should shift in the direction of the recalcitrant chamber and vote margins should increase.

In sum, the place to look for legislative anomalies are: (1) lop-sided supermajorities or, conversely, very narrow enactment votes for vetoed bills upon initial passage; (2) veto over-ride margins far from two-thirds in one or both chambers; (3) enactment votes for filibustered bills far from 60–40 in the Senate; (4) cloture vote margins far from 60–40; and (5) re-passed previously failed bills in the same legislative configuration that do not contain concessions from bill to bill.

So, how many legislative anomalies have occurred in recent decades? Has the rate of anomalies increased? Unfortunately, a comprehensive empirical analysis lies outside our writ here. However, we can present some simple data and mini-case studies that suggest anomalies now abound and have distinctive features.

### ***Veto Anomalies***

Table 13.1 presents some simple summary statistics on vetoes from 1975 to 2018. There were 167 vetoes in that period, with about half escaping an over-ride attempt. Of those that were challenged (90), about 69% were sustained (the over-ride attempt failed) while 31% succeeded. Under traditional veto bargaining models, we would expect that if a veto is challenged it should either succeed or fail by a narrow margin. Otherwise, either the president should not have vetoed it or Congress should not have challenged it. Hence, a 70% failure rate for over-ride attempts may raise an eyebrow; one might expect something closer to 50–50. In fact, Cameron (2000) reports a success rate of 45%, using earlier data (p. 56). Still, one needs to look more closely at actual vote margins to identify anomalies.

Table 13.2 takes a closer look at sustained vetoes, that is, failed over-ride attempts. It focuses on hopeless over-ride attempts. In the House, over half of the time that an over-ride attempt failed, it failed by at least 10% of the required votes (29 votes). In the Senate, some 10 of the 22 failed over-ride attempts failed by the comparable 10% margin (6 votes). Hence, the ‘hopeless over-ride’ rate among the failures was 60% in the House and 45% in

the Senate. Theory would predict something close to zero. It should also be noted that, of the 34 hopeless over-ride failures, six of these over-ride attempts failed in the Senate after success in the House (so the House success was immaterial), while the other four hopeless over-ride failures in the Senate occurred for vetoes where the House did not even attempt an over-ride (so they were truly hopeless failures). In sum, the number of hopeless over-ride attempts was not large but this phenomenon has become a notable feature of veto politics.

What type of bills did Congress typically try so hopelessly to over-ride? At least in recent cases, the bills were highly visible, highly contentious vehicles for partisan position-taking. They are similar to the bills involved in the frenetic, frenzied re-passage episodes discussed momentarily; in fact, some of them are the same bills. So, for example, bills repealing parts of ObamaCare and the Dodd–Frank financial legislation both generated vetoes and hopeless over-ride failures in the Republican Congresses facing President Obama. Hopeless over-ride failures during the Bush administration were generated by vetoed bills banning waterboarding and establishing a timeline for withdrawing troops from Iraq.

Some of the hopeless over-rides seem to follow the script of Groseclose and McCarty’s

**Table 13.1 Summary statistics on vetoes, 1975–2018**

Sustained <sup>1</sup>	62	37.1%
Over-ridden	28	16.8%
Unchallenged	77	46.1%
<i>Total vetoes</i>	<i>167</i>	<i>100%</i>

<sup>1</sup>This counts two vetoes that were over-ridden in one chamber but unchallenged in the other, technically leading to an outcome where the veto was challenged but not over-ridden. Accordingly, we classified these as sustained but exclude them from the following analysis of sustained votes.

**Table 13.2 Hopeless over-ride attempts, 1975–2018**

Sustained in House	40	Sustained in Senate	22
Failed by more than 10%	24	Failed by more than 10%	10
<i>Percent not close</i>	<i>60%</i>	<i>Percent not close</i>	<i>45%</i>



blame game vetoes. For instance, the waterboarding episode can be seen as an attempt by the Democrats to demonstrate to the public the inhumanity of the president and his administration. However, in some cases there are hints of another dynamic. Thus, reporting in *The Hill* noted: ‘Republicans say they are playing the long game with the [ACA] repeal vote, hoping it will give voters a glimpse of how they would govern if they win back the White House in November.’<sup>23</sup> We will return to this point below.

We have looked at hopeless over-ride attempts; what about hopeless vetoes? How frequently does the president get massively rolled after vetoing a bill? Given the hopeless quality of the veto, why did he veto it in the first place?

In the time period we study, the president occasionally vetoed a bill with massive support, so that an over-ride was virtually certain. Of the 28 over-ridden vetoes during this time period, Congress overrode nine of them by at least 10% in each chamber. Six of these massive rolls came during the first 12 years of the data (during the Ford and Reagan Administrations). Since then,

massive rolls of vetoes have occurred only about once per decade. Table 13.3 provides a brief overview of these vetoes.

At the time of writing, the most recent massive roll of a presidential veto involved President Obama’s veto of the Justice Against Sponsors of Terrorism Act (JASTA). This bill would have allowed private individuals to pursue legal action against foreign companies in US courts, primarily in response to the victims of the 9/11 terrorist attack. President Obama veto message cited foreign policy concerns.<sup>24</sup> President Bush’s lone massive roll came from his veto of the Water Resources Development Act of 2007. Bush claimed the bill was too pork-ridden to serve the nation’s interests.<sup>25</sup> Finally, a bill canceling Clinton’s line-item veto of military construction projects was also overridden by large margins.<sup>26</sup> As with Bush’s veto, the concerns behind the veto seem primarily centered on pork.<sup>27</sup>

In each of these examples, the president had genuine policy concerns, but the veto – a hopeless endeavor from the get-go – seems to have been undertaken partly or primarily for position-taking. Perhaps the president wanted

**Table 13.3 Massive rolls of presidential vetoes, 1975–2018**

<i>Bill Number</i>	<i>Bill Name</i>	<i>Date of Veto</i>	<i>House Vote</i>	<i>Senate Vote</i>	<i>Reason for Veto</i>
S.2040	Justice Against Sponsors of Terrorism Act	9/23/16	348–77	97–1	International concerns
H.R.1495	Water Resources Development Act of 2007	11/2/07	361–54	79–14	Too much pork
H.R.2631	Line Item Veto Cancellation	11/13/97	347–69	78–20	Too much pork
H.R.1	Water Quality Act	1/30/87	401–26	86–14	Too much spending and federal oversight
H.R.2409	Health Research Extension Act	11/8/85	380–32	89–7	Too much red tape and bureaucracy
H.R.6198	To amend the manufacturing clause of the Copyright Law	7/8/82	324–86	84–9	Free trade concerns
H.R.7102	Veterans’ Administration Health-Care Amendments	8/22/80	401–5	85–0	Spent money on VA physician bonuses instead of helping veterans
H.R.5901	Education Division and Related Agencies Appropriation Act	7/25/75	379–41	88–12	Fiscal irresponsibility
H.R.4222	National School Lunch and Child Nutrition Act	10/3/75	397–18	79–13	Fiscal irresponsibility/personal responsibility

to signal his frugality and good stewardship to a national audience (or, in Obama's case, an international one). Or, the president may have wanted to highlight Congress's fiscal imprudence, a sort of *reverse* blame game veto. In all three of these examples, both chambers of Congress were controlled by the other party.

### ***Filibuster Anomalies***

Discussion of the filibuster may seem somewhat odd in an essay on veto bargaining, but we argue that the anomalies are similar in both cases and likely to have a common origin. Therefore, let us quickly examine 'strange' patterns in filibusters, focusing on cloture vote margins.

Figure 13.1 displays vote margins in all cloture votes in the 111th to 115th Congresses (2009–18). Recall that the required quota for success was 60 votes in this period; in the figure, a margin of 0 corresponds to 60 votes for cloture. The thin vertical line shows the average margin in these Congresses: about 7.3 votes (in the 94th through 98th Congresses, the average margin was almost exactly 0). Two features stand out in the figure.

First, and most noticeable, is the very long and rather flat right-hand tail, that is, successful cloture votes. As shown, some cloture votes succeeded with absolutely spectacular margins, suggesting that the filibuster in question was a hopeless endeavor. Unfortunately, this inference is clouded by the changing vagaries of senatorial procedure. As explained by CRS experts: 'In recent times ... Senate leadership has increasingly made use of cloture as a normal tool for managing the flow of business on the floor, even when no evident filibuster has yet occurred.'<sup>28</sup> Thus, cloture is now used pre-emptively and as a device to restrict non-germane amendments. This change in procedure probably accounts for some of the huge positive margins in cloture voting. Some

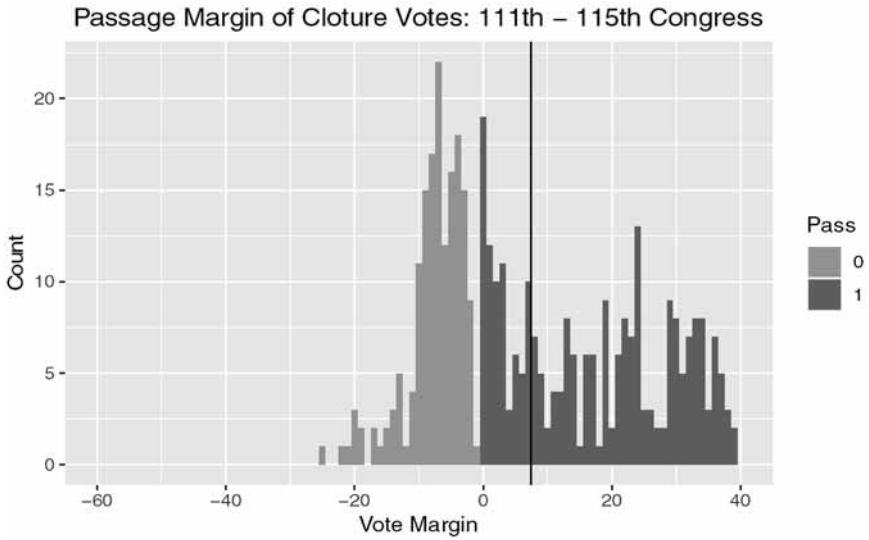
filibusters may have been fruitless efforts leading to a crushing cloture vote, but one cannot easily detect such filibusters using positive cloture margins alone.

Therefore, let us turn our focus to the left-hand tail: failed cloture motions. In the face of incomplete information about the filibuster pivot, one would expect some cloture motions to fail, but generally with margins close to zero. Yet one sees some eye-popping negative margins, some by 20 votes or more. Thus, some invincible filibusters provoked completely hopeless cloture attempts. Votes like this are hard to reconcile with classical SOP style models.

Have futile cloture efforts increased over time? Figure 13.2 addresses this question by examining the number of hopelessly failed cloture votes, votes failing by a 10% margin or more (that is, six votes or more). The time period is longer, from the 94th Congress to the 115th, in order to provide more of a historic contrast (the critical cloture margin was 60 votes over the entire period). As shown in the figure, there appears to be a jump in the number of big failures starting at the 104th Congress (1995–6). Using the benchmark of a 10% short-fall in votes, the average number of futile cloture votes was 10.1 in the 94th to 103rd Congresses; it was 23.5 in the 104th to 115th Congresses. Thus, the number of futile cloture votes doubled beginning with the 'Gingrich Revolution' Congress after the 1994 mid-term election. We note that the percentage of futile cloture votes (relative to all cloture votes) did not change much over this time period, though the number of such votes seemed to increase.

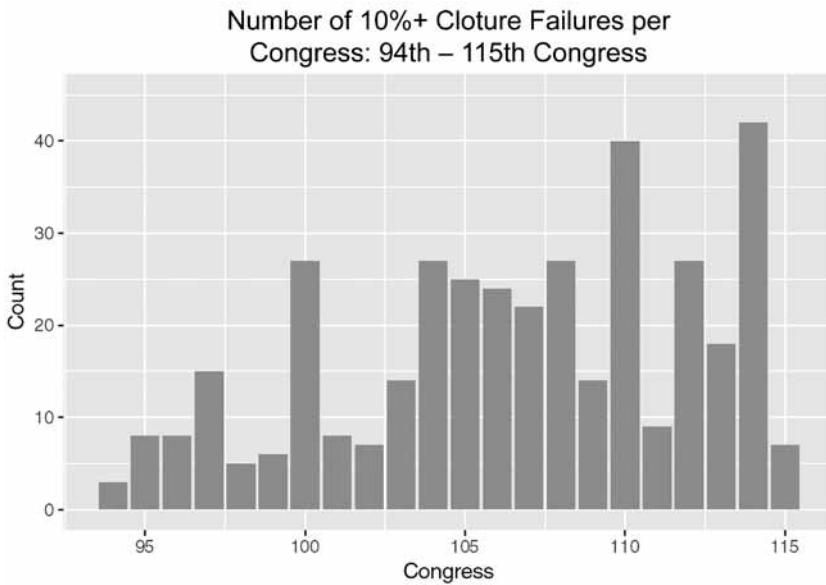
What were some examples of recent hopeless cloture votes? In the most recent period, many deal with border security, sanctuary cities, DACA and abortion – all highly visible and highly partisan issues.

We have just scratched the surface of this material but clearly some filibuster and cloture attempts look quite strange from a bilateral bargaining perspective.



**Figure 13.1** Passage margin of cloture votes, 2009–2018

*Note:* A vote margin of 0 corresponds to 60 votes for cloture. The right-hand tail of the figure captures successful cloture votes; the left-hand tail unsuccessful ones. Not every cloture motion resulted in a vote. The data exclude nominees considered under a pure majority confirmation rule.



**Figure 13.2** Futile cloture votes, 1975–2018

*Note:* Shown are counts of dramatically failed cloture votes by Congress, using the benchmark of a 10% short-fall in votes.

### ***Frenetic Failed Legislation***

One of the strangest recent phenomena in the new legislative politics is what can be called *frenetic failed legislation*. With frenetic failed legislation, one or both chambers of Congress repeatedly enact almost identical bills that all participants understand have no realistic chance of becoming law. And they do not, until the legislative configuration changes. In the traditional SOP frameworks, this spastic re-passage of doomed legislation makes about as much sense as repeatedly slamming oneself in the face with a baseball bat: it is a sign of madness. Yet, Congress has spent significant time and resources on such bills in recent Congresses. In fact, it has become a signature activity of contemporary legislative politics.

To be clear, frenetic failed legislation typically occurs under divided government, where one chamber (typically the House) passes and re-passes a bill (sometimes with minor variations) favored by the majority party in that chamber but opposed by the other chamber and/or the president. The status quo clearly falls within the gridlock interval. That is, the bill lacks the votes to overcome a filibuster or veto or both. In contrast to the sequential veto bargaining model, which envisions repeated passage of a succession of modified bills in a serious effort at policymaking, there is no effort at compromise. Instead, these repeated efforts are characterized by their intransigent and clearly infeasible nature. Let us look at a few examples from recent periods of divided government to illustrate.

The most famous example of frenetic failed legislation is of course the Republican efforts to ‘repeal and replace’ the Affordable Care Act. Recall that this landmark legislation was enacted by the 111th Congress after a historic donnybrook and signed into law by President Barack Obama in March 2010. The mid-term elections that November then saw the electorate administer a brutal drubbing to the Democrats, racking up some of the largest losses since the Great Depression – a

‘shellacking’, in President Obama’s memorable phrase. Critically, the Republican gained control of the House of Representatives, while the Democrats retained the Senate until the 2014 election, when the Republicans established a narrow majority. The Democrats held onto the presidency until the 2016 election.

The classic SOP models clearly indicate that the Republicans had no realistic prospect of repealing the ACA in the 112th, 113th and 114th Congresses. In the first two Congresses, the Democratic-controlled Senate would simply ignore House legislation. In the third Congress, Democratic filibusters or presidential vetoes would surely kill Republican bills. These were the transparently obvious outcomes predicted by the models, and that is what transpired.

Accordingly, using the SOP models, one might expect Republican legislators to focus on other legislation that might actually have a chance of enactment. Or, they might concentrate their efforts on congressional oversight, constituency service, fund-raising and just plain electioneering. Nonetheless, *The Washington Post* documented a total of 54 total or partial repeals of the ACA in the first four years of Republican control.<sup>29</sup> While these bills were far from identical, attacking the existing law from a plethora of angles, they all had the exact same chance of becoming law: zero.

The ACA wasn’t the only Obama-era statute that Republicans repeatedly attempted to repeal during this period. They also made several efforts to undo the Dodd–Frank regulations on the financial industry. For example, in 2013 alone, House Republicans passed H.R. 1256, H.R.992, H.R.2374 and H.R.1105, all of which were intended to repeal aspects of Dodd–Frank.<sup>30</sup> None of these bills were considered by the Democratic Senate.

It should be noted that Republicans held no monopoly on frenetic failed legislation. Democrats found themselves in a similar political configuration during the 109th Congress, when they had a House majority during the waning years of the Bush

administration. And they engaged in similar legislative behavior. In particular, House Democrats repeatedly attempted to restrict activities in the Iraq war, such as through requiring troop withdrawals. As noted in *CQ Weekly*: ‘In July, for example, the House passed a bill (HR 2956) sponsored by Armed Services Chairman Ike Skelton, D-Mo., that would have required troop withdrawals. But like about a half dozen other measures, it went no further.’<sup>31</sup> Furthermore, the accounts make clear that House Democrats were fully aware of the futility of their efforts:

After Republicans blocked an effort last week to require a withdrawal of U.S. troops from Iraq, Senate Democrats put the issue aside and are not expected to return to it until after the August recess. House Democrats, however, plan to do their part to keep the subject alive this week, with war-related votes possible during committee consideration of the fiscal 2008 Defense spending bill and on the floor.<sup>32</sup>

If the multitudinous ACA repeals are a sign of legislative madness, the malady, unlike much in Washington today, is refreshingly bipartisan.

But perhaps there is a method in the madness, a method outside the ambit of the classical SOP models.

### ***What Is Going on? Blame Game versus Virtue Signaling***

Our admittedly cursory review of recent empirical evidence suggests that much legislating continues to follow the script of the classical, incomplete information bilateral bargaining models. For example, in Figure 13.1 most cloture votes do fall near the 60-vote benchmark. As the same time, there appears to be a serious undercurrent of something else going on. What is it?

An obvious candidate is blame game politics. As suggested by Senator Snowe’s comment, a phenomenon like forcing a futile cloture vote in the face of an invincible filibuster may be an attempt by the chamber’s

majority to highlight the perfidy of the opposition: ‘Look, everyone! We would have this wonderful legislation but for the intransigence of these terrible people!’ So: throw the bums out!

At the same time, much of the weirdness seems somewhat distinct from pure blame game politics. For example, it may make sense to try and fail to pass a symbolically resonant bill once, in order to demonstrate that the fault for failure lies with the opposition. But why pass the same bill 60 times? How much more education in the vileness of the opposition does the public need, once the opposition is revealed to be bad via the first failure? If, as Senate Majority Leader Mitch McConnell liked to suggest, ‘There’s no education in the second kick of a mule’,<sup>33</sup> how much is there in the 40th, 50th or 60th? Similarly, even in clear blame game politics such as the veto of the Family and Medical Leave Act, part of the signaling was not just that President Bush was blocking family leave. The message was also, ‘we Democrats are really in favor of this idea and can be trusted to deliver if handed the keys to the kingdom’. In other words, the message sent to the audience is not just ‘the other side is horrible, so kick them out’ but also ‘our side is wonderful, so support us’. Virtue signaling seems as much at play as blame game.

Consequently, let’s briefly explore the politics of virtue signaling.

### **TOWARD A MODEL OF MESSAGE LEGISLATION: VIRTUE SIGNALING AND ACCOUNTABILITY**

Let’s consider a model of message legislation, legislation not intended for enactment but instead constructed *solely to send a message to outside observers*. Many obvious questions arise: who are the senders? Who are the receivers? What is the message? What gives the message meaning? What gives it credibility? Why is strategic information

transmission of this form advantageous to the parties? Many answers to these questions are possible and lead to different models. But let's sketch one set of answers, if only to suggest how to embed veto bargaining-style models of SOP policymaking within an accountability model of the electorate. We'll focus on the dramatic, frenetic failed legislation of the 'repeal and replace' variety.

First, let's assume the senders are members of a party that controls one chamber of Congress but does not control all the major veto points in the legislative process. So, the president may belong to the opposite party. Or, the other chamber may be controlled by the other party. Or, 'our' chamber may be the House while the other party controls the very constraining filibuster pivot in the Senate. Let's assume the status quo lies firmly in the gridlock region, so no enactment improving matters from the sender's policy perspective is actually possible.

Let's assume the receiver of the message is the sender's selectorate – the high-information, highly engaged portion of the party whose money, time and enthusiasm is vital for re-election. With the support of these hyper-engaged kingmakers, re-election is almost assured (the district is a safe one). But without it, the sender may well be 'primaried' and out of office. This approach is particularly compatible with the UCLA approach to parties, where parties are formed out of a coalition of policy-motivated groups which 'insist on the nomination of candidates with a demonstrated commitment to its program',<sup>34</sup> but can also fit with others in which the political marketplace is less than perfect.

Two broad classes of messages are possible. The first (as discussed above) is the *blame game message*: I will show you that the other side is terrible [so you should support me]. The second is the *virtue-signaling message*: I will show you that I am trustworthy, your faithful agent, one of you [so you should support me]. Let's consider the second class of models, since Groseclose and

McCarty already constructed an example of the former.

Virtue signaling requires the receiver (the selectorate) to have incomplete information *about the sender*, the incumbent legislator. This is in contrast to the blame game approach, where the incomplete information must be about the opposition (e.g., the opposition president or party). So, here, the selectorate is somewhat uncertain about the virtue of the incumbent representative. To make matters concrete, suppose there are two types of representatives: slackers (low virtue) and zealots (high virtue).<sup>35</sup> Slackers have no policy convictions but just value holding office. Zealots also value office but in addition they value policy, and value it similarly to the selectorate. From the viewpoint of the policy-minded selectorate, it doesn't make much difference which type holds office when policy is gridlocked. After all, no change is possible. But if policymaking becomes possible and is costly of time and effort, then it may make a great deal of difference who holds office. For on that happy day, the slacker won't do much work, but the zealot will toil like a Trojan in order to achieve the policy goal. Clearly, from the viewpoint of the policy-oriented selectorate, it will be much better to be represented by a zealot rather than a slacker on that future day.

How then can an incumbent zealot prove he is a zealot and worthy of re-election? A non-starter is, issue a raft of campaign promises. Any promise a zealot could make, a slacker could make as well. So, our model will not feature Downsian-style prospective campaign promises. Rather, it will incorporate V. O. Key-style retrospective voting. The selectorate will act in light of what has gone before, eliminating incumbents likely to be slackers and retaining those likely to be zealots. The point is to increase the chances of having a zealot incumbent when policy windows open in the future.<sup>36</sup>

Let's focus on one type of action the incumbent can undertake: frenetic failed legislation. So, pass, re-pass and continue

re-passing virtually the same bill in the face of an unbeatable veto, an invincible filibuster or just plain disregard from the opposite chamber. The resulting sequence of play is:

- 1 Nature selects the incumbent legislator's type (slacker or zealot), which is private information for the incumbent.
- 2 The incumbent engages in a futile legislative interaction with, say, the president, fruitlessly passing and re-passing the same bill with multiple vetoes and re-vetoes. Enactments are costly of time and effort that could profitably be spent elsewhere.
- 3 When either the president accepts a bill or the incumbent desists with fruitless legislating, players receive period 1 payoffs.
- 4 The voter then retains or fires the incumbent. If the voter fires the incumbent, nature selects the type of the new representative. Nature also selects a new president so that policy windows open.
- 5 The representative (either new or retained) engages in a legislative interaction with the new president.
- 6 Based on the outcome of the legislative interaction, players receive second period pay-offs.

Comparison of this sequence with that of the simple TILI game indicates a much more complex game. It features two periods, not one; incomplete information (held by the voter about the incumbent's preferences), not complete information; voter beliefs about the incumbent's preferences; costly signaling by the incumbent in period 1; retrospective voting by the voter; and, finally, serious policymaking in the second period. Still, as a costly signaling game, it is not hard to analyze using modern techniques.

We assert without proof that the virtue signaling game has two generic equilibria. In the first, a pooling equilibrium, both a slacker legislator and a zealot legislator behave the same way in period 1: they do nothing. And, in this 'incumbency advantage' equilibrium, the voter re-elects the incumbent despite the dearth of effort. Then, in the second period, a zealot legislator engages in fruitful legislating while a slacker does nothing. This

equilibrium is quite attractive for the incumbent politician regardless of type; after all, he doesn't have to do much policy work in period 1 and yet gets re-elected. But it is not so good for the voter, because when the policy window opens in the second period he may find himself saddled with a slacker as representative, resulting in a missed legislative opportunity.

The second, and more interesting, equilibrium is a separating equilibrium.<sup>37</sup> Here, in period 1 incumbent slackers and zealots behave in very different ways. The zealot engages in frantic frenetic failed legislating, fruitlessly passing and re-passing the same bill over and over and over. The slacker does nothing because imitating the furious action of the zealot would be too costly of effort. The zealot's policy-mindedness creates a wedge between him and the slacker that allows this separation to occur – but only at high levels of effort, hence the need to do a lot of futile policymaking. The voter then fires a revealed slacker and retains a revealed zealot. In the second period, when policy windows open, a zealot works hard to legislate and a slacker doesn't. This equilibrium is much worse for the legislator: a period 1 slacker gets fired, and a period 1 zealot must slave away at phony legislating in order to retain his job. But this equilibrium is much better for the voter because it boosts the chance of having a valuable zealot in place when policy windows open.

We have only sketched an analysis of message legislation and virtue signaling. But we hope we have at least suggested that the idea is worth pursuing, and that the politics of virtue signaling is distinct from but complementary to the politics of the blame game. Carefully elaborating the theory of virtue signaling may enable some parsing of the difference between the two and lead empirical work in new directions, for example, the effect of message votes on fund raising, primary challenges and citizen voting – all new directions for SOP-style models. In addition, further theoretical development might well

tackle the question: why the rise in message legislation? The new media environment, partisan polarization of elites, the rise of groups such as the Tea Party on the right and ‘the opposition’ on the left, partisan sorting geographically and across the parties, and the increase in competition to control the government are probably all implicated. But how exactly? In a related way, in the context of the virtue signaling model one might ask, across different issues, when should we expect the pooling equilibrium to prevail, and when the separating equilibrium (the difficult question of equilibrium selection)?

### **DOES BARGAINING BEFORE AN AUDIENCE MAKE A DIFFERENCE?**

We have suggested ways to modify classical SOP models, such as the veto bargaining models, in order to better capture the new American legislative politics. The new politics on which we have focused results from blame game politics but also (we suggest) from politicians’ virtue signaling to an attentive audience of ideological extremists. But is modifying the classical models to incorporate message legislation really worth the effort? After all, the classic SOP models more or less get it right with respect to policy outcomes: when they say policy windows are shut, relatively little is enacted. When they identify the key veto players, they are generally correct. And when they suggest the political evaluation needed for enactment – that is, the spatial position of viable legislation in something like NOMINATE space – usually they are close to the mark. So, one may well ask, does all the noisy action attendant on message legislation actually make a substantive difference? Or does the sound and fury signify nothing?

Our sketch model of message legislation and virtue signal suggests that the politics of bargaining before an audience does make a difference for outcomes, though a

fairly subtle one. The separating equilibrium in our proto-model involves considerable information transmission between the sending congressperson and the constituent. The constituent learns something about the congressperson and – critically – then uses the information in choosing either to support or oppose the incumbent. The result is an ideological filter applied to incumbents, resulting over time in greater retention of representatives who are zealous in pursuit of the electorate’s ideological goals. The result is not quite the same thing as ideological polarization *per se*. But because the key constituents who follow and respond to the political theater tend to be high-information ideologically consistent extremists, the net effect is to build a more extreme legislature over time. In essence, there is an enhanced feedback loop between incumbent position-taking and constituent response, leading to a legislature responsive to relatively extreme blocks within the electorate. Or so the model suggests. An obvious question is: is this actually happening?

There is another element, outside our sketch model, but of potential importance and concern. In the pooling equilibrium, zealous incumbents who face gridlock don’t undertake any policymaking effort since it won’t accomplish anything substantively and they will get re-elected anyway. But in fact, case studies show zealous policy-minded congressmen doing a lot of policy work during down periods. In particular, congressional policy entrepreneurs hone their legislative proposals and lay the foundation for future legislative coalitions. For instance, Senator Bill Bradley spent years working on tax reform before policy windows opened creating the opportunity for a big policy innovation.<sup>38</sup> Similarly, famed policy entrepreneur Representative Henry Waxman labored long and hard, often for years, to build carefully crafted bills well aimed at specific health policy problems.<sup>39</sup> The result was (arguably) high-quality bills ready to go, when the gridlock region narrowed and legislative opportunity presented itself.



In contrast, in the separating equilibrium, zealous legislators work extremely hard during the gridlock period, acting out a laborious pantomime of legislating – ‘repeal and replace’, for instance. All that effort devoted to phony legislating must come from somewhere. One obvious candidate is real legislating of the Bradley–Waxman variety: low-profile, under-the-radar preliminary work without which high quality enactments are impossible, or at least far less likely.<sup>40</sup> From this perspective, one consequence of the era of message legislation may be a reduction in the quality of actual enactments. In addition, the dearth of high-quality ready-to-go bills may suppress legislative productivity when policy windows open, again as suggested by the failure of the ACA repeal. Obviously, this possibility is speculative. But is it true? Is the quality of enactments down, is there a dearth of high-quality draft bills, and does legislation fail despite open windows because no one did the preliminary work of crafting a high-quality bill? These are troubling but compelling questions.

Changes in American politics create opportunities and challenges for empiricists and theorists both: for the first, to document what has happened, for the latter to explain it. Then, there is a challenge at the interface of theory and data: does the new theory really afford an understanding of the new patterns? Or has it missed the mark? In this chapter, we have reviewed some of the big changes in American legislative politics and offered a proposal on how to craft new theory for the new politics. Whether that new theory will be forthcoming, and whether it will prove successful in explaining the new politics, remains to be seen.

## Notes

- 1 Ferejohn and Shipan, 1990; Gely and Spiller, 1990.
- 2 Krehbiel, 1998; Brady and Volden, 1998; Wawro and Schickler, 2013.

- 3 Cameron, 2000; Groseclose and McCarty, 2001; Howell, 2003.
- 4 On the parties, see Bawn et al., 2012; Heaney et al., 2012; Levendusky, 2009. For an interesting elaboration and partial demurrer, see McCarty and Schickler, 2018. On new media and the new media environment, see Farrell, 2012 and Prior, 2013.
- 5 McCarty, forthcoming.
- 6 One of the authors is somewhat partial to Cameron and McCarty, 2004 and Cameron, 2009.
- 7 Lee, 2016.
- 8 We would be remiss not to note Gilmour, 1995, an analysis that in retrospect appears extremely perceptive.
- 9 Ashworth, 2012 is a succinct recent overview. Besley, 2006 is often seen as a touchstone, while Fearon, 1999 remains useful.
- 10 We borrow the concept of the selectorate from Bueno de Mesquita et al., 2005; see also Bawn et al., 2012.
- 11 In simple models, though, the space could be multi-dimensional so long as the players are assumed to be unitary actors.
- 12 For interested readers, a good introduction to the spatial theory of voting remains Enelow and Hinich, 1984; Duggan, 2005 provides a comprehensive recent survey.
- 13 Mayhew, 1991.
- 14 Krehbiel, 1998.
- 15 *Ibid.*, but see Wawro and Schickler, 2013 for a war-of-attrition approach to the filibuster.
- 16 Cox and McCubbins, 2007.
- 17 Ferejohn and Shipan, 1990.
- 18 Moe and Howell, 1999.
- 19 There are other ways to study veto threats that deserve attention: see Hassell and Kernell, 2016.
- 20 Banks, 1991.
- 21 Again, a forward-thinking precursor was Gilmour, 1995.
- 22 The cutting line between yeas and nays in the policy space occurs midway between the bill and the status quo. If the status quo remains the same and the bill is re-passed with a concession, the new cutting line should shift in the direction of the concession. The logic is explained in more detail in Cameron, 2000, which takes the test to actual data. Over most of the 20th century, one sees this pattern during sequential veto bargaining.
- 23 Weaver, 2016.
- 24 Obama, 2016.
- 25 Bush, 2007.
- 26 The Line Item Veto Act of 1996 conferred a line item veto on the president. The Supreme Court quickly struck down the legislation as unconstitutional, but in its brief life it allowed President Clinton to line-item veto portions of some bills.

- 27 Clinton, 1997.
- 28 Heitshusen and Beth, 2017: 1.
- 29 O'Keefe, 2014.
- 30 Weyl, 2014.
- 31 Donnelly, 2008: 41.
- 32 Donnelly and Graham-Silverman, 2007: 2186.
- 33 Bolton, 2013.
- 34 Bawn et al., 2012: 571.
- 35 This follows the nomenclature of Gailmard and Patty, 2007.
- 36 Ashworth, 2012 makes the interesting observation that rational retrospective voting is inherently prospective in intent, a point that is not completely transparent in early discussions like that of Key, 1966 or Fiorina, 1981. The logic is similar to punishing a child for bad behavior: the point is not to slake the parent's thirst for revenge, but to improve the child's conduct in the future.
- 37 The separating equilibrium requires technical conditions typical of costly signaling games (see Banks, 1991). Critically, the marginal cost of repeated failures must be smaller for zealots than for slackers.
- 38 Birnbaum and Murray, 1988.
- 39 Waxman, 2009.
- 40 The diversion of effort from low-profile bill crafting to high-profile messaging is an example of the perverse incentives often seen in multi-task principal-agent games: see Holmstrom and Milgrom, 1991.
- Bolton, Alexander. 2013. 'GOP's McConnell Promises No More Shutdowns over Obama-Care.' *The Hill*. <https://thehill.com/homenews/senate/329145-leader-mcconnell-no-more-shutdowns-over-obamacare>. Accessed June 2, 2019.
- Brady, David W., and Craig Volden. 1998. *Revolving Gridlock: Politics and Policy from Carter to Clinton*. Westview Press.
- Bueno de Mesquita, Bruce, Alastair Smith, James D. Morrow, and Randolph M. Siverson. 2005. *The Logic of Political Survival*. MIT Press.
- Bush, George W. 2007. 'President Bush Vetoes Water Resources Development Act of 2007.' The White House Office of the Press Secretary. <https://georgewbush-whitehouse.archives.gov/news/releases/2007/11/20071102-3.html>. Accessed January 28, 2019.
- Cameron, Charles M. 2000. *Veto Bargaining: Presidents and the Politics of Negative Power*. Cambridge University Press.
- Cameron, Charles M. 2000a. 'Bargaining and Presidential Power.' pp. 47–77 in Robert Y. Shapiro, Martha Joynt Kumar, and Lawrence R. Jacobs (eds), *Presidential Power: Forging the Presidency for the Twenty-first Century*. New York: Columbia University Press.
- Cameron, Charles M. and Nolan McCarty. 2004. 'Models of Vetoes and Veto Bargaining.' *Annual Review of Political Science* 7.1: 409–435.
- Cameron, Charles M., John S. Lapinski, and Charles R. Riemann. 2000. 'Testing Formal Theories of Political Rhetoric.' *Journal of Politics* 62.1: 187–205.
- Clinton, Bill. 1997. 'Veto of H.R. 2631.' U.S. Government Printing Office. <https://www.govinfo.gov/content/pkg/CDOC-105hdoc172/pdf/CDOC-105hdoc172.pdf>. Accessed January 28, 2019.
- Cox, Gary W., and Mathew D. McCubbins. 2007. *Legislative Leviathan: Party Government in the House*. Cambridge University Press.
- Donnelly, John M. '2007 Legislative Summary: Defense: Conduct of the Iraq War.' *CQ Weekly* (January 7, 2008): 41. <http://library.cqpress.com.ezproxy.princeton.edu/cqweekly/weeklyreport110-000002652090> (Accessed on December 1, 2018).
- Donnelly, John M., and Adam Graham-Silverman. 'Reid Shelves Troop Withdrawal.'

## REFERENCES

- Ashworth, Scott. 2012. 'Electoral Accountability: Recent Theoretical and Empirical Work.' *Annual Review of Political Science* 15.1: 183–201.
- Banks, Jeffrey. 1991. *Signaling Games in Political Science*. Routledge.
- Bawn, Kathleen, Martin Cohen, David Karol, Seth Masket, Hans Noel, and John Zaller. 2012. 'A Theory of Political Parties: Groups, Policy Demands and Nominations in American Politics.' *Perspectives on Politics* 10.3: 571–597.
- Besley, Timothy. 2006. *Principled Agents? The Political Economy of Good Government*. Oxford University Press.
- Birnbaum, Jeffrey H., and Alan S. Murray. 1988. *Showdown at Gucci Gulch: Lawmakers, Lobbyists, and the Unlikely Triumph of Tax Reform*. New York: Vintage.

- CQ Weekly* (July 23, 2007): 2186–2187. <http://library.cqpress.com.ezproxy.princeton.edu/cqweekly/weeklyreport110-000002555834> (Accessed on December 1, 2018).
- Duggan, John. 2005. 'A Survey of Equilibrium Analysis in Spatial Models of Elections.' Unpublished manuscript, University of Rochester.
- Enelow, James M., and Melvin J. Hinich. 1984. *The Spatial Theory of Voting: An Introduction*. Cambridge University Press Archive.
- Farrell, Henry. 2012. 'The Consequences of the Internet for Politics.' *Annual Review of Political Science* 15.1: 35–52.
- Fearon, James D. 1999. 'Electoral Accountability and the Control of Politicians: Selecting Good Types versus Sanctioning Poor Performance.' In Adam Przeworski, Susan Stokes, and Bernard Manin (eds), *Democracy, Accountability, and Representation*. Cambridge University Press.
- Ferejohn, John, and Charles Shipan. 1990. 'Congressional Influence on Bureaucracy.' *Journal of Law Economics & Organization* 6: 1–20.
- Fiorina, Morris P. 1981. *Retrospective Voting in American National Elections*. New Haven: Yale University Press.
- Gailmard, Sean, and John W. Patty. 2007. 'Slackers and Zealots: Civil Service, Policy Discretion, and Bureaucratic Expertise.' *American Journal of Political Science* 51.4: 873–889.
- Gely, Rafael, and Pablo T. Spiller. 1990. 'A Rational Choice Theory of Supreme Court Statutory Decisions with Applications to the "State Farm" and "Grove City" Cases.' *Journal of Law, Economics, & Organization* 6.2: 263–300.
- Gilmour, John B. 1995. *Strategic Disagreement: Stalemate in American Politics*. University of Pittsburgh Press.
- Groseclose, Tim, and Nolan McCarty. 2001. 'The Politics of Blame: Bargaining before an Audience.' *American Journal of Political Science* 45.1: 100–119.
- Hassell, Hans J. G., and Samuel Kernell. 2016. 'Veto Rhetoric and Legislative Riders.' *American Journal of Political Science* 60.4: 845–859.
- Heaney, Michael T., Seth E. Masket, Joanne M. Miller, and Dara Z. Strolovitch. 'Polarized Networks: The Organizational Affiliations of National Party Convention Delegates.' *American Behavioral Scientist* 56.12: 1654–1676.
- Heitshusen, Valerie, and Richard S. Beth. 2017. 'Filibusters and Cloture in the Senate.' Congressional Research Service. <https://www.senate.gov/CRSPubs/3d51be23-64f8-448e-aa14-10ef0f94b77e.pdf>.
- Holmstrom, Bengt, and Paul Milgrom. 1991. 'Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design.' *The Journal of Law, Economics, & Organization* 7: 24–52.
- Howell, William G. 2003. *Power Without Persuasion: The Politics of Direct Presidential Action*. Princeton University Press.
- Key, V. O. 1966. *The Responsible Electorate*. Belknap Press of Harvard University Press.
- Krehbiel, Keith. 1998. *Pivotal Politics: A Theory of US Lawmaking*. University of Chicago Press.
- Lee, Frances E. 2016. *Insecure Majorities: Congress and the Perpetual Campaign*. University of Chicago Press.
- Levendusky, Matthew. 2009. *The Partisan Sort: How Liberals Became Democrats and Conservatives Became Republicans*. University of Chicago Press.
- Matthews, Steven A. 1989. 'Veto Threats: Rhetoric in a Bargaining Game.' *The Quarterly Journal of Economics* 104.2: 347–369.
- Mayhew, David R. 1991. *Divided We Govern*. New Haven: Yale University Press.
- McCarty, Nolan M. 1997. 'Presidential Reputation and the Veto.' *Economics & Politics* 9.1: 1–26.
- McCarty, Nolan M. Forthcoming. *Political Polarization in America: What Everyone Needs to Know*. Oxford University Press.
- McCarty, Nolan M., and Eric Schickler. 2018. 'On the Theory of Parties.' *Annual Review of Political Science* 21.1: 175–193.
- Moe, Terry M., and William G. Howell. 1999. 'The Presidential Power of Unilateral Action.' *The Journal of Law, Economics, and Organization* 15.1: 132–179.
- Obama, Barack. 2016. 'Veto Message from the President – S.2040.' The White House Office of the Press Secretary. <https://obamawhitehouse.archives.gov/the-press-office/2016/09/23/veto-message-president-s2040>. Accessed January 27, 2019.

- O'Keefe, Ed. 2014. 'The House Has Voted 54 Times in Four Years on Obamacare. Here's the Full List.' *Washington Post*. March 21, 2014. <https://www.washingtonpost.com/news/the-fix/wp/2014/03/21/the-house-has-voted-54-times-in-four-years-on-obamacare-heres-the-full-list/>. Accessed June 2, 2019.
- Prior, Markus. 2013. 'Media and Political Polarization.' *Annual Review of Political Science* 16.1: 101–127.
- Romer, Thomas, and Howard Rosenthal. 1978. 'Political Resource Allocation, Controlled Agendas, and the Status Quo.' *Public Choice* 33.4: 27–43.
- Wawro, Gregory J., and Eric Schickler. 2013. *Filibuster: Obstruction and Lawmaking in the US Senate*. Princeton University Press.
- Waxman, Henry. 2009. *The Waxman Report: How Congress Really Works*. New York: Twelve.
- Weaver, Dustin. 2016. 'House Passes ObamaCare Repeal, Sending Measure to President.' *TheHill*. January 6, 2016. <https://thehill.com/policy/healthcare/264980-house-passes-obamacare-repeal-sending-measure-to-obamas-desk>.
- Weyl, Ben. 2014. '2013 Legislative Summary: Dodd-Frank Changes.' *CQ Weekly* (January 6, 2014): 34. <http://library.cqpress.com.ezproxy.princeton.edu/cqweekly/weeklyreport113-000004402465>.



# Models of Coalition Politics: Recent Developments and New Directions

Lanny W. Martin and Georg Vanberg

## INTRODUCTION

Coalition bargaining is at the heart of politics in most parliamentary democracies, particularly those operating under proportional representation electoral rules. Elections in such systems usually leave no single party in control of a majority of legislative seats. At the same time, governments must maintain the support of a parliamentary majority to survive votes of (no-)confidence and to pass legislation.<sup>1</sup> As a consequence, the formation of a government typically involves bargaining among multiple parties to produce a joint policy program and to distribute the various ‘spoils’ of office, such as cabinet ministries, among its members.<sup>2</sup> Given the prevalence of multiparty governments, it is not surprising that the study of coalition politics has been a central research program in political

science for several decades, which has led to the development of a rich tradition of formal, deductive models of coalition bargaining. Our aim in this chapter is to survey recent developments in the study of coalition politics, to provide a typology of approaches and to sketch potential avenues for further theoretical development.

To preview, we argue that recent developments can best be understood as the result of the arrival of the ‘new institutionalism’ in coalition politics. Early models of coalition formation and bargaining focused primarily on the size and ideological compatibility of political parties, but abstracted away from the institutional context within which coalitions operate (e.g., Riker, 1962; Axelrod, 1970; De Swaan, 1973). These theories were concerned with predicting the types of coalitions that would form, but did not speak directly to the more nuanced policy and distributional

aspects of coalition bargaining. Theoretical developments in the 1990s moved these latter questions to the fore, and in so doing greatly enriched theories of coalition politics by incorporating institutional features that structure the bargaining process.

Recent contributions to coalition theory can usefully be grouped into two overarching frameworks. One is a research tradition rooted in non-cooperative game theory. This tradition has tended to focus on the strategic bargaining process among potential coalition partners, and has relied on explicit assumptions about actor preferences and the bargaining protocols that structure their interaction. A second approach, which has been characterized by greater methodological heterogeneity, has shied away from making strong assumptions about bargaining protocols, focusing instead on the background constraints that limit the viability of coalition governments (such as, most obviously, the requirement to maintain majority legislative support as well as the support of all coalition members). As we discuss, and as one might expect, each approach has its advantages and limitations.

## THE NEW INSTITUTIONALISM IN COALITION POLITICS

Early coalition theories focused on characteristic features of potential coalitions that party elites might take into account as they bargain over government formation. One such feature is the *size* of a coalition: to maximize their share of distributional payoffs, party elites should not form coalitions that include members who are not necessary to reach the relevant ‘winning threshold’, that is, coalitions should be minimal winning (Gamson, 1961; Riker, 1962). Second, to the extent that coalitions must make policy, and as doing so is easier when parties are ideologically compatible, party elites should prefer coalitions that

are ideologically connected and compact (Axelrod, 1970; De Swaan, 1973). These considerations significantly narrow the set of coalitions that are attractive to party leaders, and generate powerful predictions of coalition formation outcomes. As Martin and Stevenson (2001) have shown, these predictions have considerable empirical support. But importantly, approaches focusing on the ideological compatibility of parties and the minimal winning status of a coalition have been largely silent on the content of the coalition bargain. What do coalitions agree to with respect to policy and the zero-sum perquisites that governance provides (such as cabinet portfolios)?

One of the major developments in coalition theory over the past 30 years has been the emergence of models that aim to provide a fuller picture of coalition politics by addressing both *which* kinds of coalitions are likely to form and *what* it is that coalition partners will agree to. There are, of course, multiple paths into this literature. For our purposes, a useful starting point is the game-theoretic approach proposed in a seminal paper by Baron and Ferejohn (1989), which focused on the way in which coalitions might allocate a fixed prize (such as cabinet ministries) among their members.<sup>3</sup> Their non-cooperative approach launched an influential modeling tradition in coalition politics that has been stated in its most general form by Snyder et al. (2005). The central logic of this tradition is to extend the famous Rubinstein bargaining model (Rubinstein, 1982) to a setting in which acceptance of a proposal requires the agreement of a winning coalition (instead of a single player). The essential features of the Baron–Ferejohn (hereafter, BF) approach consist of the following:

- a set of legislative parties, none of which controls a decisive number of seats;
- a legislative process that requires approval by a coalition of decisive size (typically, a majority) to accept a proposal;
- a fixed amount of a good that a proposal must divide among the parties;

- a random recognition rule that selects one party – the formateur – to make a proposal for a division of the good, which is then subjected to a legislative vote;
- if the proposal is approved, the implementation of the proposed division, and if it is not approved, a new random selection of a formateur.

While the BF model is a general model of legislative decision-making, BF explicitly suggested that their approach can be understood as a model of coalition formation in parliamentary systems, as well as the allocation of ministerial portfolios among coalition members. Their model (and extensions of it) have since occupied a central place in coalition theory. The key implication of the BF approach is that in equilibrium, the formateur (the party selected to form a government) will construct a minimal winning coalition by making an offer that attracts the agreement of the ‘cheapest’ set of partners available and allocates the lion’s share of the good to itself. A simple example can illustrate the intuition behind this result. Suppose there are three parties, with each party controlling  $\frac{1}{3}$  of the seats in the legislature. Further assume that parties are chosen as formateur with a probability that equals their seat share. Finally, assume that if a proposal fails, and bargaining moves to a new round (where, once again, each party is chosen as formateur with probability  $\frac{1}{3}$ ), parties discount their payoff by  $\delta \in (0, 1)$ . Suppose party A has been selected as formateur. Party A knows that if its proposal is rejected, all parties (being perfectly symmetrical) have the same *ex ante* expectation about their future payoff. Denote this continuation value (which each party expects to receive in the next period) by  $c$ . To secure a majority for its proposal, party A must thus offer one other party  $\delta c$  to secure its vote, keeping the remainder of  $1 - \delta c$  for itself. Given the symmetry of the parties, this implies a payoff of  $1 - \frac{\delta}{3}$  for the formateur and a payoff of  $\frac{\delta}{3}$  for its chosen partner. The (left-out) opposition party receives no payoff.<sup>4</sup>

Like earlier models, the BF approach yields the prediction that coalitions will be minimal winning. But the seminal contribution lies in characterizing the *distribution* of spoils among the coalition partners – with the key expectation that formateur parties are able to exploit their privileged position to secure a particularly favorable outcome at the expense of their partners. Applied to the formation of coalition governments in parliamentary systems, the expectation of this ‘formateur advantage’ has given rise to an extensive empirical literature that has examined the degree to which formateur parties are able to secure additional payoffs unavailable to ordinary coalition members.

Much of this literature has focused on one of the most salient and easily observable ‘zero-sum’ payoffs: the numerical allocation of cabinet ministries among coalition partners. The overarching conclusion that has emerged is that there is no formateur bonus in payoffs (see, e.g., Warwick and Druckman, 2006). Rather, these studies suggest that portfolios are distributed in a matter that is largely proportional to the seat contribution of parties to the coalition (with a small bonus for small parties) – a regularity frequently referred to as ‘Gamson’s Law’ that had already been established in much earlier work (Gamson, 1961; Browne and Franklin, 1973). Put differently, it appears empirically that formateur parties are not able to ‘extract’ portfolios beyond the number they would be expected to receive purely on the basis of their size – a finding that stands in obvious tension with the theoretically influential BF paradigm.

### **Critiques of BF**

This disjuncture between the empirical regularity of Gamson’s Law and the theoretical prediction of BF-style models has been the lynchpin of critical assessments of this modeling approach for understanding multiparty governments in parliamentary systems. The formateur advantage that is central to the BF

model (and to models that extend this approach, such as that of Snyder et al. (2005)) is rooted in two ‘institutional’ assumptions: (1) there is a strict order in which proposals can be made; and (2), once made, the proposal is subjected to an ‘up-or-down’ vote.<sup>5</sup> This is critical, because it implies that only one offer is on the table at any one time and that rejection of this offer is costly, since it implies delay and reversion to a new round of bargaining in which any individual party is uncertain about its position. The formateur can exploit this fact to make an offer that is minimally acceptable to its prospective partners, and thereby secure a favorable payoff for itself. Put differently, it is the rigid bargaining protocol which restricts who can make proposals, and how those proposals must be dealt with, that generates the privileged position of the formateur.

It is precisely this feature of the BF framework that has recently come under theoretical and empirical attack in a series of significant papers by Laver et al. (2011) and Cutler et al. (2016). While proponents of the BF approach argue that its bargaining protocol ‘is consistent with the empirical pattern found in the formation of coalition governments’ (Snyder et al., 2005: 982), these critics charge that applications of the BF bargaining model to coalition politics are problematic because the bargaining protocol – which drives the central results of the model – does not accurately characterize the context within which real-world coalition negotiations take place.<sup>6</sup> As Laver et al. (2011: 300) point out, coalition negotiations are not subject to a rigid, enforceable protocol that approximates the BF model:

*Lack of structure* is of the essence because negotiations over government formation involve the most experienced and sophisticated politicians playing for the highest possible stakes. Rather than assuming such people adhere to an unenforceable norm under which they make public offers in an exogenously choreographed sequence and engage in no backroom discussion, it is far more reasonable to assume that *nothing* can

prevent *any* politician from proposing *any* deal at *any* time during government formation.

Put differently, the BF approach illuminates the dynamics of a highly structured bargaining situation. But for that very reason, it is theoretically inappropriate for understanding coalition politics in parliamentary systems.<sup>7</sup> Even if a specific party is identified as a formateur *de jure*, bargaining in these contexts is fluid and open *de facto*: all party leaders are free to negotiate with anyone, to bargain with multiple parties in parallel, and all participants can float potential offers at any time – if need be, in smoke-filled backrooms. As a result, conclusions that rest on a rigid bargaining protocol do not travel well to these contexts.<sup>8</sup>

The conclusion that Laver et al. (2011: 296) draw is that models of coalition politics must rest ‘on premises that can plausibly be argued, on substantive grounds, to have empirical relevance’. Put differently, to the extent that particular modeling assumptions are critical to substantive conclusions that emerge from a model, these assumptions should accurately reflect the context within which party elites bargain. In their view, this can be accomplished within two distinct (though overlapping) paradigms.

One approach is that of non-cooperative game theory, which rests on explicit assumptions about player preferences, the strategies available to them and the structure and timing of their interaction. For Laver et al. (2011: 301), the key to progress within this paradigm is to ensure that ‘the local institutional detail that structures such models ... come[s] from genuinely binding institutional constraints in the substantive environment modeled, not from “institutions” that are, when all is said and done, just tools in the modeler’s box’.

A second avenue is to move away from models rooted in non-cooperative game theory and the necessity of fully characterizing the bargaining protocol. Such approaches can focus on overarching constraints on



the government formation process and constraints on the characteristics of coalition outcomes, while making no particular assumptions about the underlying bargaining process. Laver et al. (2011) find this approach particularly promising for two reasons. One is that such an approach seems to capture the notion that elite politicians, bargaining over the ultimate prize in politics, are unlikely to be constrained by fixed formal rules that govern the bargaining process. The second is that even if such rules exist, they would vary tremendously over time and place, making a focus on overarching constraints that are common across many bargaining environments more useful. We now turn to illustrative examples of each type of approach.

## ENRICHING NON-COOPERATIVE MODELS

As just highlighted, a central critique of the BF approach as a model of coalition formation in parliamentary systems is that its central result – the formateur bonus – derives from a bargaining structure that appears in tension with the process by which real-world coalitions form. Moreover, the theoretical prediction of the formateur bonus stands in clear tension to the empirical reality of Gamson's Law. A major focus of coalition theorists working within the tradition of non-cooperative game theory has therefore been to consider alternative bargaining models that can resolve this discrepancy between theory and empirics, and offer more substantively grounded accounts of coalition bargaining. Doing so has typically involved moving away from the alternating offer framework of BF. We briefly highlight several such approaches.

The seminal – and relatively early – contribution to these efforts was the 'demand competition' model of Morelli (1999). The Morelli model introduces two important variations to BF-style bargaining. The first

is that parties bargain not only over a divisible good (such as portfolios), but also over a common coalition policy position. The second is that bargaining does not proceed through a series of sequential take-it-or-leave-it offers, but rather a series of (sequential) *demands*: the head of state chooses a formateur, who in turn determines the order in which parties will make demands. When it is a party's turn, the party proposes a coalition policy, and makes a demand for a share of the private good. If, at some point, a winning coalition emerges because its members agree on policy, make compatible demands and have the backing of a legislative majority, the coalition takes office. If no coalition emerges, a new formateur is chosen and the game begins again (with a caretaker government taking office and leaving parties with no share of the divisible good after a set number of failed negotiations).

Critically, in the Morelli model *any* (winning) collection of parties that makes compatible demands can take office – *even if this coalition leaves out the initial formateur* (or other parties whose demands have been made). This bargaining protocol induces an implicit competition among the parties because any party's demand can be ignored – there is no power to make take-it-or-leave-it demands. (This feature also ensures that the order in which demands are made does not matter.) The result is that the formateur advantage is eliminated. Instead, 'the ex post equilibrium distribution of payoffs is proportional to the ex ante distribution of bargaining power (within the majority coalition)' (Morelli, 1999: 817).

To see the intuition, return to a three-party example. For simplicity, suppose that failure to agree on a coalition leads to the formation of a caretaker government after one round, and that parties only care about the divisible good.<sup>9</sup> Note that the bargaining power of the parties is equal: each party can form a majority coalition with any other party. What demand should the first party make? Suppose it demands a share  $p_1 < \frac{1}{2}$ . The second

party now faces a choice. It could accept  $1 - p_1$  and form a coalition with the first party. Or it could issue a demand for  $p_2 = p_1 - \varepsilon$ . Should it do so, Party 3 has three options. It could form a coalition with Party 1, form a coalition with Party 2 or let negotiations fail (leading to a caretaker government). Since  $1 - (p_1 - \varepsilon) > 1 - p_1 > 0$ , Party 3 chooses a coalition with Party 2. In words, Party 1's initial demand has backfired, because the other parties are able to ignore it. Demanding more than half of the divisible good leaves Party 1 empty-handed. What about demanding  $p_1 < \frac{1}{2}$ ? Party 2 would immediately accept such an offer, and form a coalition with Party 1. (To see this, note that demanding a larger share than  $1 - p_1$  would lead Party 3 to accept Party 1's offer, leaving Party 2 out of the coalition.) But Party 1 could have gotten a better deal by increasing its demand slightly. Thus, the equilibrium demand by Party 1 – which will be accepted immediately by Party 2 – is for half the divisible good. The implicit competition between the parties leads to an even division between the members of the coalition, given that all are (essentially) in the same bargaining position: each party can form a majority coalition with any other party. More generally, competition between potential coalition partners limits the ability of any party to extract 'extra' concessions – and because the degree of competition a party faces depends on its bargaining power (loosely speaking, the proportion of winning coalitions of which it is a part), the equilibrium distribution of payoffs is proportional to parties' bargaining strength.

The significance of the Morelli contribution does not lie only in proposing a model of coalition bargaining that reconciles the tension between theory and empirics by predicting a payoff distribution that approximates Gamson's Law (as long as seat distributions and bargaining power are reasonably highly correlated). The more important contribution is methodological. Staying within the paradigm of non-cooperative game theory – which requires a clear articulation of the

game form, including players' strategies and the timing of actions – the model captures our intuitions about 'real-world' coalition bargaining in the sense that the competitive logic of the model provides a close analogy to a free-form bargaining process in which any party is able to float a potential coalition deal, parties are able to ignore some potential partners and any group of parties can 'get together' to form a government. In other words, while there is, as in the BF approach, a strict order of play in the Morelli model, this order does not affect the central equilibrium result.<sup>10</sup> Instead, the equilibrium outcome is driven by what is, intuitively, the critical aspect of real-world bargaining: the extent to which a party is in a powerful position (because it is central to many winning coalitions) or in a weak spot (because it can easily be ignored by other parties in forming a winning coalition). In this sense, the Morelli model represents precisely the kind of non-cooperative approach that Laver et al. (2011) suggest.

Of course, Morelli's (1999) is not the only non-cooperative approach that provides an alternative to the BF tradition. While a full review of all these models is beyond the scope of this chapter, we briefly highlight two additional contributions here. Bassi (2013) proposes a model in which parties bargain over a divisible good in a process in which parties auction off the right to act as formateur. As in the Morelli model, in equilibrium, no formateur bonus emerges: the competition among parties for the role of formateur ensures that potential rents to securing this position are dissipated in equilibrium.<sup>11</sup> Second, building on Austen-Smith and Banks (1988), Baron and Diermeier (2001) offer a model that integrates voter decisions in an election with a subsequent government formation process among parties that care about policy and office-holding. A formateur selects a proto-coalition, which bargains efficiently over policy (using office benefits as side payments). Because bargaining over policy is assumed to be efficient, government policy

does not depend on the (BF-style) bargaining protocol in the model. But the distribution of office benefits and the identity of the proto-coalition does, exposing the model to the critiques identified above.<sup>12</sup>

## BEYOND NON-COOPERATIVE GAME THEORY

The models we have just reviewed illustrate approaches that expand on the BF approach within the paradigm of non-cooperative game theory. The Morelli (1999) model in particular captures the competitive nature of the open-ended and essentially unconstrained process in which potential coalition partners bargain, thus addressing the limitations of the rigid alternating offer protocol of BF. Nevertheless, it still employs its own rigid bargaining protocol. As we argued above, a second approach in coalition theory has been to move away from explicit assumptions about the structure of bargaining, and instead toward frameworks that focus on assumptions about background constraints on coalition bargaining. We review two examples of such approaches, beginning with the seminal ‘ministerial autonomy’ model proposed by Laver and Shepsle (1996).

### *Laver and Shepsle’s Ministerial Autonomy Model*

This model, which also occupies a central place in coalition theory, applies the logic of the structure-induced equilibrium approach developed by Shepsle (1979) in the context of US congressional committees to coalition governments in parliamentary systems. It rests on three key assumptions:

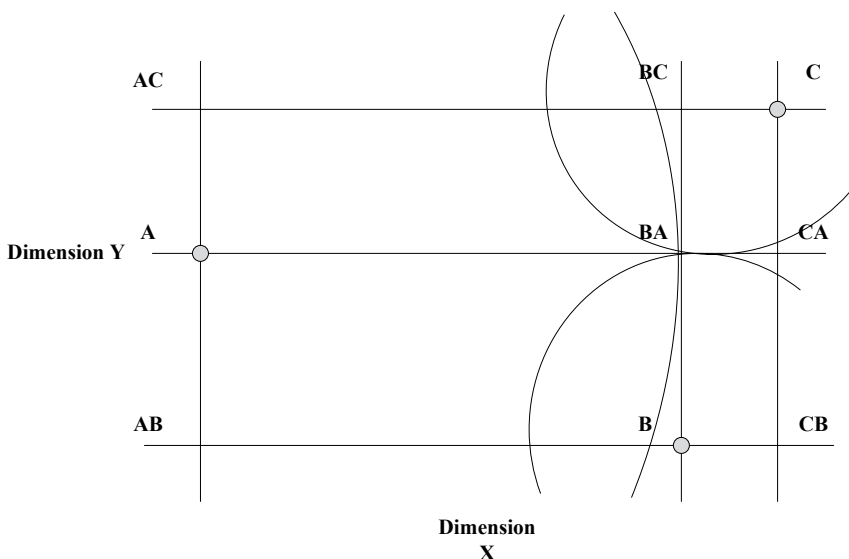
- coalition politics involves government parties agreeing on a government policy package in a multi-dimensional policy space;

- each ministry in a coalition government has jurisdiction over a single issue dimension;
- within each issue dimension, the party that controls the relevant portfolio enjoys autonomy to determine policy.

The consequence of these assumptions is to transform a collective choice problem with no (preference) equilibrium under majority rule (McKelvey, 1976) into one with well-defined properties. The combination of radically reducing the relevant policy space by associating each dimension with one ministry and assuming that each ministry has autonomy within its sphere serves to ensure that any allocation of portfolios among a specific set of parties implies a particular policy outcome. Specifically, in each jurisdiction, the ministerial party implements its preferred policy, and the collection of these policies across all jurisdictions constitutes the coalition’s policy program.

Put more provocatively, the Laver–Shepsle (hereafter, LS) model essentially assumes away the possibility of meaningful bargaining among coalition partners, and assumes instead that – given the central role of cabinet ministers in drafting and implementing policy within their jurisdiction – each party acts as dictator within its portfolios, implements its preferred policy and expects all other parties to do so. Consider the example in Figure 14.1, taken from Diermeier (2006), consisting of three parties in a two-dimensional policy space, that is, a policy space with two ‘ministries’. Assume that no party controls a majority of seats (implying that any pair of parties does). The ideal points of each party are given by the points *A*, *B* and *C*. In this scenario, a government consists either of a single-party government, which receives both portfolios, or a two-party coalition in which each party receives one portfolio.

Given the assumptions above and the assumption that party preferences are given by Euclidean distance, the two-dimensional policy space is reduced to the ‘lattice points’ implied by any particular allocation of



**Figure 14.1** Laver–Shepsle ministerial autonomy model

portfolios. For example, a coalition of parties *A* and *B*, in which party *A* controls the portfolio for dimension *Y* and party *B* controls the portfolio for dimension *X*, implies policy outcome *BA*. If the parties reverse the assignment of portfolios, the resulting coalition policy becomes *AB*. The nine possible government policy positions are indicated by the lattice points in the figure, and labeled according to the allocation of portfolios associated with each.

Which of these potential governments (characterized by a specific distribution of portfolios) is likely to emerge? To answer this question, a fully game-theoretic approach would require detailed specification of the bargaining process by which a government forms (and that specification would have to incorporate the LS background assumptions about policymaking). The structure-induced equilibrium approach takes a different tack. Rather than ask *how* potential coalition partners bargain toward the formation of a coalition, the question that Laver and Shepsle ask is which governments are viable or stable, in the sense that they cannot be replaced

by another feasible government. Given the constraints on formation of a government in parliamentary systems, this requires two conditions: there cannot be an alternative feasible government that is preferred to the current government by a *majority* of the legislature, and by *all* the members of the alternative government.

In Figure 14.1 (taken from Diermeier, 2006:167), the feasible alternative governments are indicated by the lattice points. Consider the government in which party *A* controls the portfolio for dimension *Y* and party *B* controls the portfolio for dimension *X* (with resulting policy *BA*). This government constitutes a structure-induced equilibrium in the sense that no feasible government (that is, none of the lattice points) lies within the win-set of *BA*; that is, no alternative feasible government is preferred by a majority to the existing coalition. In contrast, consider a coalition in which party *C* controls the *X*-ministry and party *A* controls the *Y*-ministry, with resulting coalition policy *CA*. This coalition is *not* viable: Parties *A* and *B* both prefer the government associated with policy outcome

BA, and could vote to replace the existing government.

The LS approach offers a number of important contributions. On the conceptual level, the model explicitly integrates a theory of coalition policymaking into a theory of coalition formation. A critical component of what makes a coalition viable or vulnerable in this approach is the expectation about the policy position the coalition will adopt, compared to those policy outcomes that would be produced by other feasible governments. While the logic that policy expectations feed into coalition formation is implicit in much of coalition theory, the LS approach was one of the first to explicitly connect these two features.<sup>13</sup>

A second contribution is related to the Laver et al. (2011) critique of the BF framework. The ministerial autonomy model does not focus on the bargaining process among potential coalition partners, and for this reason requires no assumptions about a rigid bargaining protocol. Rather, it rests on a twin foundation:

- 1 a bedrock constraint that any government in a parliamentary system must respect, namely, that there cannot be an alternative government that is preferred to the incumbent by all of its members and a parliamentary majority, and;
- 2 an assumption about the coalition policymaking process, namely, that ministerial parties are able to set policy within their jurisdictions autonomously.

Given this foundation, the model generates predictions about which coalitions are viable, the portfolio distribution associated with such governments and the resulting policy outcomes.

While the LS model constitutes a seminal and highly influential contribution, the approach is subject to at least three limitations. One is that viable cabinets (i.e., cabinets that constitute structure-induced equilibria) do not exist in all circumstances. Unlike models that rely on Nash equilibrium solution concepts (which always exist), the LS model will

thus not yield a solution in all circumstances (Austen-Smith and Banks, 1990). Perhaps more seriously, the LS approach is designed to identify a set of viable governments (i.e., governments that are stable in the sense that no alternative feasible government is preferred to it by a legislative majority, and by all its members). But the approach is silent on the choice *among* viable governments if more than one exists. In other words, Laver and Shepsle offer no theory of government selection in the face of multiple feasible alternatives. How to think about this limitation is not clear. On the one hand, theories (including game-theoretic approaches) that incorporate particular bargaining protocols do yield implications regarding government selection. On the other hand, if such selection is (as in the BF framework) driven by bargaining protocols that do not have a strong substantive grounding, an agnostic approach (such as the LS framework) may be preferable.

The final limitation, in our view, is the most serious. The core logic of the ministerial autonomy model is to identify viable governments by reducing the set of feasible alternatives. The key to doing so is to assume that ministers act as policy dictators within their jurisdiction. It is this assumption that implies a one-to-one mapping between a portfolio allocation and the (multi-dimensional) policy position adopted by a coalition government. Put in the context of the Laver et al. (2011) critique, the problem is not so much that Laver and Shepsle impose a rigid bargaining protocol. Rather, it is that they *replace* the bargaining process with strong and rigid assumptions about policy-making. Because the only incentive-compatible policy choices are for ministers to pursue their (unique) ideal policy within their jurisdiction, there are no opportunities for meaningful bargaining among parties. Policy positions that do not correspond to the lattice points associated with ministerial autonomy cannot be achieved. Just as the key results of the BF model are driven by its (rigid and, in the context of coalition formation, implausible)

bargaining protocol, so the key results of the LS model are driven by the strong assumption of ministerial autonomy.

As a result, the empirical relevance of this assumption becomes central: do ministers – and by extension, coalition parties – enjoy considerable discretion within the policy areas under their jurisdiction? Over the past decade, a number of scholars have argued that the answer depends on the institutional framework within which coalition governments make policy. For example, the presence of junior ministers who can ‘keep tabs’ on their coalition partners may allow parties to rein in the discretion of cabinet ministers (Thies, 2001). Similarly, legislative institutions (such as strong committee systems) that allow for effective scrutiny and amendment of ministerial proposals can effectively limit the ability of ministers to act as policy dictators (Martin and Vanberg, 2005, 2011). There is strong evidence that such institutions limit the discretion of ministerial parties, and allow for genuine compromise agreements (Martin and Vanberg, 2014a). With respect to legislative institutions, in particular, recent empirical work has shown that in environments where committee systems and legislative procedures allow for effective scrutiny of government ministers, real-world policy outcomes tend to reflect the preferences of the coalition *as a whole*; in contrast, where committee systems do not permit such scrutiny, policy tends to reflect the preferences of the party controlling the relevant ministry (Martin and Vanberg, 2019a). These findings suggest scope conditions on the LS approach. If parties are able to limit ministerial discretion, and to pursue genuine compromise policies, the radical reduction of the policy space that is required for the emergence of structure-induced equilibrium governments no longer holds. Thus, the LS approach is most applicable in contexts in which ministers enjoy considerable unilateral power, but is less informative in contexts in which coalition parties can reach – and enforce – compromise agreements.

### ***The Zero-intelligence Model of Government Formation***

The LS framework combines a specific policymaking assumption (ministerial discretion) with the premise that cabinet stability requires that a government cannot be beaten by a feasible alternative that is preferred by a majority and all of its members to identify a set of viable governments. A second approach is even more radical in pursuing Laver et al.’s (2011) call to move away from the rigid bargaining protocols of non-cooperative game theory. The ‘zero-intelligence’ model of government formation developed by Golder et al. (2012) aims to investigate how the outcomes of coalition bargaining are structured by two foundational constraints on government formation in parliamentary systems. These constraints are that (1) there is always a status quo government, and (2) to replace a status quo government, a prospective new government must be preferred to the status quo by a legislative majority and all of its members. As we just saw, this requirement also plays a prominent role in the LS approach. But unlike LS, who add a particular assumption about the policy process, Golder, Golder and Siegel (hereafter, GGS) aim to impose no additional features beyond the requirement that a government not be vulnerable to being replaced by an alternative. The outcomes of the coalition formation process that GGS seek to explain are the types of governments that form (minority, minimal winning and supermajority), the allocation of portfolios across coalition partners and delays in government formation.

To explore how the two foundational constraints shape these outcomes, GGS investigate a model in which parties compete in a two-dimensional policy space. Governments are characterized by a policy position, and a distribution of portfolios among the coalition partners. Parties care about the share of portfolios they receive and the divergence between government policy and their own position. The model proceeds as follows.

At the outset, parties are assigned policy positions in the policy space at random, and non-strategic voters vote for the party that is located closest to their ideal point. This step results in a ‘party system’ that is characterized by parties with a particular ideological position, and a vote/seat weight that is determined by their share of the vote. At this stage, the government formation process begins in the following way:

- 1 from among all coalitions that include itself, each party randomly proposes a government, along with a randomly assigned policy position and portfolio allocation for that government;
- 2 given the collection of all proposed governments, each party compares its utility from each proposed government to the utility it derives from the incumbent government;
- 3 proposed governments that are preferred to the incumbent by (a) all of their members and (b) a set of parties that jointly control a majority of seats are deemed ‘viable’, resulting in a set of governments that are capable of replacing the incumbent;
- 4 one of the viable governments is chosen at random to replace the incumbent;
- 5 if no government is viable, then the full process repeats, and if no government has formed after 100 iterations, the incumbent remains in office.

Note that this modeling set-up imposes a background constraint on government viability, in stages (2) and (3), but does not impose a rigid bargaining protocol. To investigate the implications of this model, GGS employ a Monte Carlo approach to simulate thousands of party systems and government formation attempts, recording information on the types of governments that emerge, the distribution of portfolios and delay in government formation.

Significantly, the aggregate distributions of these outcomes bear reasonable resemblance to empirically observed distributions. That is, the model predicts the formation of minority, minimal winning and supermajority governments, ‘approaching the real-world distribution of government types’ (Golder et al., 2012: 436). It also generates a close

approximation of Gamson’s Law for portfolio allocation, as well as delays in formation.

The only systematic assumption that the model imposes is that replacement governments must be preferred to the incumbent by a majority and all of their members; all other features of the model (including the selection of governments) are *random*. As a result, GGS conclude that central dynamics of government formation are driven by the underlying constraints that are imposed on the viability of governments in parliamentary systems, rather than being a product of forward-looking, strategic behavior by political elites. As they put it (Golder et al., 2012: 443):

Overall, the results from our model indicate that the two constitutional constraints that bind in any parliamentary government formation process are sufficient to approximate many of our most robust empirical regularities, independent of the strategic behavior of party leaders and their bargaining protocols. This suggests that structure, not behavior, may be the most important thing when it comes to explaining government formation outcomes.

The zero-intelligence model is a powerful illustration of an approach that focuses on underlying constraints, while making no assumptions about a particular bargaining protocol. And yet, it is precisely this feature that is also a weakness of the zero-intelligence model. In focusing on background constraints that viable governments must conform to, the GGS model (like the LS approach) identifies a *set* of viable governments. In the face of multiple viable candidates, the key question for a theory of government formation is *which* among the many alternatives will emerge. Put differently, what is required is a theory of government selection. As we saw, LS do not address this issue. In contrast, the GGS model does include a specific theory of government selection: namely, the government is chosen at random.

As Martin and Vanberg (2014b: 875) point out, this is equivalent to a radical behavioral

assumption about party elites, namely, that *they have no preferences among alternative viable governments*. Rather than comparing potential governments to each other in order to choose the most attractive option (and to bargain if there is disagreement), elites are content to compare potential governments to the incumbent, and then to choose at random. Given that government formation is a game in which highly skilled politicians compete for the ultimate prize in their professional careers, the implausibility of this behavioral assumption is clear: party elites obviously have preferences among potential coalitions.

Critically, the ‘random choice’ assumption is not innocuous: the results of the GGS model (in particular, the fact that the model produces aggregate distributions that resemble real-world outcomes) are largely driven by the assumption of random selection. To understand why, note that if party elites have preferences over potential governments, these preferences discriminate systematically against certain types of governments. For example, a preference for maximizing distributional benefits (such as office) will steer elites away from minority or oversized coalitions. The impact of this assumption can be seen most clearly by shifting attention from the *aggregate* distribution of government characteristics to the (much more relevant) predictive power of the model in specific formation opportunities. As Martin and Vanberg demonstrate, while the zero-intelligence model can generate empirically plausible aggregate distributions of coalition characteristics, its record in predicting which particular coalition will emerge in a given formation opportunity is very poor, and it is simply no match for models that take account of elite preferences among potential governments.

### ***Beyond Zero-intelligence***

As we have just argued, the problem posed by the selection of a particular government from a set of plausible alternatives is a

challenge for models that seek to move beyond the restrictive bargaining protocols of non-cooperative game theory. In this context, a recent paper by de Marchi and Laver (2019) offers an innovative approach. Like GGS, de Marchi and Laver depart from the non-cooperative game-theoretic paradigm by imposing no rigid bargaining protocol. They also employ behavioral assumptions that depart from fully rational, strategic behavior. The core of their model is to treat coalition formation as a problem in which parties must agree on a joint policy program that is based on the often extensive election manifestos and pledges on which parties have campaigned. Given the large number of issues involved, this represents a complex bargaining problem in a high-dimensional space.

De Marchi and Laver assume that parties treat each issue as a binary choice between favoring and opposing a particular change to the status quo, and that parties attach different salience to issues. The key assumption of the model is that, given this set-up, parties treat the problem of agreeing on a coalition policy platform as a logroll. The fact that the intensity of preferences varies across issues opens up the possibility that parties can trade concessions on issues they care less about for concessions on issues that are salient to them. Naturally, in a high-dimensional issue space, such logrolling behavior can be extremely complex. Thus, de Marchi and Laver impose a behavioral assumption: parties look for ‘simple’ logrolls on issues that can be traded directly, but forego more complicated exchanges. Issues on which the parties disagree but that cannot be resolved with a simple binary logroll are ‘tabled’ and left unresolved.

A particularly novel feature of the approach – and the one we want to highlight here – is how de Marchi and Laver confront the issue of government selection. Consistent with their critique of BF-style models, they reject the notion of a rigid bargaining protocol. Of course, as we saw in our discussion of LS and GGS, this leaves the critical question of



government selection. It is here that de Marchi and Laver introduce an important innovation. Because coalition bargaining takes place in an unconstrained ‘open outcry’ environment, they focus on the following logic: while the bargaining environment is unconstrained in its *procedures*, any government that emerges must have majority support. In this context, cabinets that are Condorcet winners (i.e., are majority-preferred to any alternative government) are privileged since they are not vulnerable to other proposals. Moreover, when a Condorcet winner exists, the government formation process is – intuitively – less complex and less vulnerable to continuous bargaining. A key question for de Marchi and Laver is thus the likelihood that, in any particular bargaining environment, a Condorcet winner exists.

To identify the conditions under which that is likely to be the case, for any given formation opportunity, de Marchi and Laver examine the policy vector associated with each possible coalition after all simple logrolls have been executed and remaining issues have been tabled. This generates the set of all potential cabinets and their associated policy platforms. Given these, it is possible to ask whether there exists a cabinet that constitutes a Condorcet winner. Significantly, de Marchi and Laver demonstrate computationally that in the logrolling framework, and given the behavioral assumptions of limited rationality, Condorcet winners are more frequent than one might naively assume. They also show that their likelihood depends predictably on model parameters. In particular, in legislatures defined by Laver and Benoit (2015) as ‘top-three’ systems (where any two of the three largest parties can form a majority coalition), a Condorcet winner is especially likely to exist.

The significance of this innovation is that it offers a middle way between the precision of game-theoretic models that is often achieved through bargaining protocols that are difficult to defend on substantive grounds, and the agnosticism of models that

feature no assumptions about the bargaining process. The de Marchi and Laver approach respects the open-outcry nature of the bargaining process while offering a systematic way to narrow down the problem of government selection that is anchored in the fact that governments must enjoy majority support.

## CONCLUSION: MOVING COALITION THEORY FORWARD

In most parliamentary systems, most of the time, no single party controls a majority of legislative seats. As a consequence, coalition governance is central, and scholars have showered considerable attention on understanding how these governments form, govern and dissolve. In this chapter, we have provided a short overview of a particular subset of this literature: formal models of coalition formation. These models are concerned with two central questions:

- 1 Which parties are likely to ‘join forces’ in a government?
- 2 How do the parties that participate in a coalition distribute the spoils of office among themselves, and what kind of policy programs do they agree on?

Early coalition scholarship concentrated primarily on the first question. But with the arrival of the ‘new institutionalism’, which explicitly incorporated the institutional context within which coalitions govern, scholars have increasingly focused on the second.

One useful way to characterize the types of models that have played a central role in these developments is to distinguish between game-theoretic approaches and those approaches that forego the rigid assumptions of a particular game form. The primary advantage of the former is that in specifying a particular bargaining protocol and focusing on equilibrium outcomes, these approaches yield specific characterizations

of government selection and the distribution of payoffs among coalition partners. The primary weakness of the approach is closely related: the particular equilibrium properties are often driven by the specific bargaining protocol assumed. In the context of real-world coalition politics, which is essentially an unconstrained ‘open-outcry’ environment, this is potentially troubling since the properties of equilibrium cabinets that depend on a bargaining protocol may not carry over to this environment.

One solution to the ‘protocol dependence’ of game-theoretic models is to move toward models that do not impose a particular bargaining structure. This is the aim of the second class of models we have considered. These models typically incorporate some ‘background constraints’ (such as the need to receive majority support), but they resist more specific assumptions regarding how parties bargain over coalitions. The strength of these models is that their results appear not to depend on strong assumptions about the bargaining process. But these approaches also face two limitations. The first is that – depending on the model – the absence of a rigid bargaining protocol may be more apparent than real. Thus, as we saw in the case of the Laver and Shepsle (1996) model and the Golder et al. (2012) zero-intelligence model, equally strong assumptions regarding the policy process (ministerial autonomy) or elite behavior (random choice among viable cabinets) drive the results. The second limitation is that these models typically provide no clear answer to the question of government selection. While non-cooperative models yield specific predictions regarding which coalition will emerge, the second class of models usually identifies a *set* of viable coalitions, but has less to say about *which* of these governments will be chosen.

This tension points to an important area for future theoretical development. Making progress on understanding government selection while not falling victim to the desire to ‘solve’ this problem through the imposition of

bargaining protocols that have little substantive grounding represents one of the key challenges. The de Marchi and Laver approach of focusing on the presence of a Condorcet winner, and investigating the conditions that make this more likely, represents an important recent development in this regard. But there is clearly room for additional work.

A second important area for future development is to enrich the nature of the ‘goods’ that parties bargain over. For reasons of analytic tractability, current models typically focus on the distribution of a divisible good (a ‘divide the dollar’ game) and/or the selection of a (multi-dimensional) policy position in a spatial model as the two relevant aspects of the coalition bargain. While the insights that emerge from such models are valuable in understanding key aspects of coalition bargaining (such as the importance of bargaining weights, as revealed in the Morelli (1999) model), they also leave unexplored significant aspects of coalition bargaining that become salient when one considers the heterogeneous qualities of the goods that parties bargain over, as well as more complex features of the bargaining environment.

A brief example can illustrate the point. A recent paper by Martin and Vanberg (2019b) examines the consequences for coalition bargaining of the fact that parties must bargain before an audience of supporters. The key assumption of their model is that supporters fall into two groups: sophisticated supporters, who are capable of evaluating all aspects of a coalition bargain, and unsophisticated supporters, who only focus on the easily observable features of coalition bargains, and evaluate these by simple heuristics. As Martin and Vanberg demonstrate, the consequences for coalition bargaining are significant: parties have strong incentives to strike bargains that meet the expectations of unsophisticated voters for those outcomes these voters observe, while using those aspects of the bargain that these supporters are unaware of as *side payments* to reconcile the coalition bargain with the

underlying bargaining strength of the parties. The result is that the distribution of payoffs that is apparent for more observable features of coalition bargains is systematically different than the distribution of payoffs for more complex aspects. In other words, what voters see is *not* necessarily what they get from the bargains struck by their parties. An empirical analysis of the numerical allocation of portfolios received by parties compared with the implied policy benefits they receive from that allocation strongly supports this conclusion.

The key point we wish to emphasize with this example is that there are at least two directions in which further theoretical development can aim. The first is to consider the more complex aspects of the bargaining environment within which parties act, including the effect of outside ‘audiences’ or the need to build reputations for subsequent bargaining rounds. The second is to move beyond the relatively simple characterizations of the goods that parties bargain over to a more nuanced, and substantively grounded, understanding of the heterogeneous characteristics of ‘goods’ (e.g., their observability to outside audiences). We suspect that both developments will require a closer collaboration between more theoretically oriented scholars and those with expertise in the empirical reality of everyday coalition governance.

## ACKNOWLEDGEMENTS

We wish to thank Scott de Marchi and Kristopher Ramsay for helpful discussions and comments on an earlier version of this chapter.

## Notes

1 Of course, such support does not require that the government itself controls a majority of parliamentary seats so long as it can rely on the implicit support of (some) opposition parties to provide legislative support to the government.

- 2 Naturally, bargaining among coalition partners continues over the lifetime of a government as parties make policy decisions. While the models we focus on in this chapter primarily concern the initial formation stage, scholars over the past several years have begun to explore the ongoing process of coalition governance (see, e.g., Strøm et al., 2008; Martin and Vanberg, 2011).
- 3 Natural alternative starting points would be a paper by Austen-Smith and Banks (1988), which presented an integrated model of an election and subsequent government formation, or the ‘ministerial autonomy model’ of Laver and Shepsle (1996). We review both of these contributions below.
- 4 To see this, note that if bargaining breaks down, and the process moves to the next period, the expected payoff for each party (and thus its continuation value) is given by  $c = \frac{1}{3}(1 - \delta c) + \frac{1}{3}\delta c + \frac{1}{3}0$ , which implies that  $c = 1/3$ . There is no incentive to provide any payoff to the opposition party since its vote is not necessary to pass the proposal, and giving it a positive share would only reduce the payoff of the formateur.
- 5 Baron and Ferejohn (1989: 1195) consider an extension that allows amendments to the proposer’s offer. As they show, this possibility reduces, but does not eliminate, the formateur advantage.
- 6 See also the original statement by Baron and Ferejohn (1989: 1194): ‘The bargaining over the allocation of ministries in a coalition government, then, corresponds closely to the model of a legislature operating under a closed rule: the party asked to form a government makes a proposal for allocating ministries (the benefits) among the parties, knowing that if its proposal does not receive a majority, another party (perhaps itself) will be asked (using the same recognition probabilities) to form a government.’
- 7 Of course, this leaves open the possibility that there are contexts that approximate the structure of BF-style models.
- 8 While the major thrust of the Laver et al. (2011) critique is theoretical, they also raise a central empirical measurement issue. An evaluation of a BF-style formateur bonus requires an *ex ante* measure of a party’s formateur status that is an indication of which party is charged, initially, with forming a government (if such a party exists). But as they demonstrate, coding of such a measure is difficult, as a result of which the measure of formateur status employed is actually an *ex post* measure – basically, an indicator of the party that ultimately secures the prime ministership in the government that forms.

- 9 These assumptions simplify the example, but do not affect the substantive conclusion. See Morelli (1999) for a general statement and proof.
- 10 There is some additional nuance here when parties have policy preferences, and the 'head of state' has some discretion in selecting the formateur. In this case, the equilibrium policy (but not the distribution of the divisible good) depends on the choice of formateur, and thus on the order of play. This aspect of the Morelli model is thus subject to the same Laver et al. (2011) critique as the BF approach.
- 11 There is a common logic between the Bassi and Morelli models in that it is competition among potential coalition partners/formateurs that eliminates the formateur advantage. That said, there is a significant difference in that the competition in the Morelli model – competition over inclusion in a cabinet, and the terms on which a party would join a coalition – appears substantively closer to 'real-world' coalition bargaining than competition for the role of formateur (whose formateur status is then firm).
- 12 Diermeier and Merlo (2000) generalize the Baron and Diermeier (2001) model, focusing on the dynamics of government turnover.
- 13 The Austen-Smith and Banks (1988) model, which was developed at roughly the same time as the LS approach, also has this feature.

## REFERENCES

- Austen-Smith, David and Jeffrey Banks. 1988. 'Elections, Coalitions, and Legislative Outcomes.' *American Political Science Review* 82(2):405–22.
- Austen-Smith, David and Jeffrey Banks. 1990. 'Stable Governments and the Allocation of Policy Portfolios.' *American Political Science Review* 84(3):891–906.
- Axelrod, Robert. 1970. *Conflict of Interest*. Chicago: Markham.
- Baron, David P. and Daniel Diermeier. 2001. 'Elections, Governments, and Parliaments in Proportional Representation Systems.' *The Quarterly Journal of Economics* 116(3):933–67.
- Baron, David P. and John A. Ferejohn. 1989. 'Bargaining in Legislatures.' *American Political Science Review* 83(4):1181–206.
- Bassi, Anna. 2013. 'A Model of Endogenous Government Formation.' *American Journal of Political Science* 57(4):777–93.
- Browne, Eric C. and Mark N. Franklin. 1973. 'Aspects of Coalition Payoffs in European Parliamentary Democracies.' *American Political Science Review* 67(2):453–69.
- Cutler, Josh, Scott de Marchi, Max Gallop, Florian M. Hollenbach, Michael Laver and Matthias Orłowski. 2016. 'Cabinet Formation and Portfolio Distribution in European Multi-party Systems.' *British Journal of Political Science* 46(1):31–43.
- de Marchi, Scott and Michael Laver. 2019. 'Government Formation as Logrolling in High-Dimensional Issue Spaces.' *Journal of Politics* (Forthcoming). <https://doi.org/10.1086/706462>
- de Swaan, Abram. 1973. *Coalition Theories and Cabinet Formations: A Study of Formal Theories of Coalition Formation Applied to Nine European Parliaments after 1918*. Amsterdam: Elsevier Scientific Publishing Co.
- Diermeier, Daniel. 2006. Coalition Government. In *The Oxford Handbook of Political Economy*, ed. Barry R. Weingast and Donald A. Wittman. Oxford: Oxford University Press pp. 162–79.
- Diermeier, Daniel and Merlo Antonio. 2000. 'Government Turnover in Parliamentary Democracies.' *Journal of Economic Theory* 94(1):46–79.
- Gamson, William A. 1961. 'A Theory of Coalition Formation.' *American Sociological Review* 26(3):373–82.
- Golder, Matt, Sona N. Golder and David A. Siegel. 2012. 'Modeling the Institutional Foundation of Parliamentary Government Formation.' *Journal of Politics* 74(2):427–45.
- Laver, Michael and Kenneth A. Shepsle. 1996. *Making and Breaking Governments: Cabinets and Legislatures in Parliamentary Democracies*. Cambridge: Cambridge University Press.
- Laver, Michael and Kenneth Benoit. 2015. 'The Basic Arithmetic of Legislative Decisions.' *American Journal of Political Science* 59(2):275–91.
- Laver, Michael, Scott de Marchi and Hande Mutlu. 2011. 'Negotiation in Legislatures over Government Formation.' *Public Choice* 147(3–4):285–304.
- Martin, Lanny W. and Randolph T. Stevenson. 2001. 'Government Formation in Parliamentary

- Democracies.' *American Journal of Political Science* 45(1):33–50.
- Martin, Lanny W. and Georg Vanberg. 2005. 'Coalition Policymaking and Legislative Review.' *American Political Science Review* 99(1):93–106.
- Martin, Lanny W. and Georg Vanberg. 2011. *Parliaments and Coalitions: The Role of Legislative Institutions in Multiparty Governance*. Oxford: Oxford University Press.
- Martin, Lanny W. and Georg Vanberg. 2014a. 'Parties and Policymaking in Multiparty Governments: The Legislative Median, Ministerial Autonomy, and the Coalition Compromise.' *American Journal of Political Science* 58(4):979–96.
- Martin, Lanny W. and Georg Vanberg. 2014b. 'A Step in the Wrong Direction: An Appraisal of the Zero-Intelligence Model of Government Formation.' *Journal of Politics* 76(4):873–79.
- Martin, Lanny W. and Georg Vanberg. 2019a. 'Coalition Government, Legislative Institutions, and Public Policy in Parliamentary Democracies.' *American Journal of Political Science* (Forthcoming).
- Martin, Lanny W. and Georg Vanberg. 2019b. 'What You See Is Not Always What You Get: Bargaining before an Audience Under Multiparty Government.' Unpublished Manuscript, Bocconi University.
- McKelvey, Richard D. 1976. 'Intransitivities in Multidimensional Voting Models and Some Implications for Agenda Control.' *Journal of Economic Theory* 12(3):472–82.
- Morelli, Massimo. 1999. 'Demand Competition and Policy Compromise in Legislative Bargaining.' *American Political Science Review* 93(4):809–20.
- Riker, William. 1962. *The Theory of Political Coalitions*. New Haven: Yale University Press.
- Rubinstein, Ariel. 1982. 'Perfect Equilibrium in a Bargaining Model.' *Econometrica* 50(1):97–109.
- Shepsle, Kenneth A. 1979. 'Institutional Arrangements and Equilibrium in Multidimensional Voting Models.' *American Journal of Political Science* 23(1):23–57.
- Snyder, James M., Jr., Michael M. Ting and Stephen Ansolabehere. 2005. 'Legislative Bargaining under Weighted Voting.' *American Economic Review* 95(4):981–1004.
- Strøm, Kaare, Wolfgang Müller and Torbjörn Bergman, eds. 2008. *Cabinets and Coalition Bargaining: The Democratic Life Cycle in Western Europe*. Oxford: Oxford University Press.
- Thies, Michael F. 2001. 'Keeping Tabs on Partners: The Logic of Delegation in Coalition Governments.' *American Journal of Political Science* 45(3):580–98.
- Warwick, Paul V. and James N. Druckman. 2006. 'The Portfolio Allocation Paradox: An Investigation into the Nature of a Very Strong but Puzzling Relationship.' *European Journal of Political Research* 45(4):635–65.



# Models of Interstate Conflict

James D. Morrow and Jessica S. Sun

The application of formal models to the study of international conflict has a long history, from the days of models of arms races based on systems of differential equations and the invocation of two-by-two strategic form games to the widespread use of non-cooperative game theory over the last three decades. Rather than simply survey the range of models commonly employed in the literature these days, this essay seeks to aid the researcher who is constructing a formal model in the key decisions that any modeling project requires. Our target reader is one who does not have great experience using these methods, and so can benefit from a guide to thinking of how different types of models focus attention on certain aspects of a situation at the expense of others. We survey a range of different models used to study international conflict with an eye to the strengths and limits of each of these classes of models. We hope that this survey will help readers make deliberate choices when they develop their models.

Formal modeling as a research enterprise seeks to identify a key causal process, simplify it down to its essence and then analyze that process to reveal its central logic. Any real situation has many possible causal processes active at any time. Including all of them renders analysis intractable: which one predominates? Do some cancel out others? Instead of presenting and considering all possible causal processes, the modeler focuses on one in the hope that in isolation we can understand what drives that process. With that deeper understanding of one causal process, we may be able to determine its observable consequences and how it might interact with other plausible causal processes (Powell, 1999: chapter 1).

The design of a model is critical to the enterprise. Elements of the situation that are not essential to the causal process being studied must be pared away. The actors must be defined, their choices and how they interact to produce outcomes specified, their preferences over those outcomes stated, and the information each holds when they choose, fixed.

These design choices force the modeler to consider the essence of the problem they wish to explore. In practice, we find the design process to be iterative, beginning with a first model, solving it, and then thinking about what parts should be cut away in favor of others. In the end, a good model lays out the logic of the causal process clearly, enabling the reader to grasp its essential logic.

Formal models are particularly useful tools for studying international conflict. The internal political process of governments making foreign policy is often private and only recoverable when it has become history and viewed as safe to reveal. Although international crises and war are public events, some key communications between governments in crisis are private. Given the complexity of some conflict events, important details were not collected when those events occurred and so are lost to history. While many impressive efforts have been made to collect data on international conflict, the resulting data sets often reflect only the parts of the behavior that can be easily captured. These difficulties reduce the value of a purely inductive approach to the study of international conflict and make models an attractive research tool to explore those areas that lie unobserved.

Faced with the inability to measure many important aspects of conflict behavior, theorists of international conflict have posited a wide range of arguments for why violent conflicts occur. Formal models could help us specify the logic of those arguments carefully and fully to understand the essence of these purported causal processes. We can also play out the observable consequences of a model to think about how what we can observe might allow us to judge which causal processes produce those patterns of behavior. Even if we do not use models as guides to empirical patterns, there is still value in working through the logic of the theories of international conflict to understand the essence of the causal processes advanced by those theories (Bueno de Mesquita and Morrow, 1999).

We begin by discussing models of complete information, because they were a natural

outgrowth of key questions in international conflict. Such models have important limits, however. The development of games of incomplete information opened up questions of perception and strategic communication to formal analysis that complete information assumed away. Repeated games could analyze questions of reciprocity in conflict and whether agreements could be enforced through reciprocity. Although games of complete information could address issues of commitment – whether parties would follow through on their threats and promises – stochastic games offered a superior technology for modeling commitment issues. These models could be married to models of domestic politics to study the interplay of international and domestic politics. Other classes of games, notably games of timing and Blotto games, have received less use but should be in the range of modeling tools to study international conflict.

## COMPLETE INFORMATION

Deterrence, both nuclear and conventional, was a central issue during the Cold War. Could the threat of military action dissuade the Soviet Union from expanding its influence through coercive means? An expected utility calculation lies at the heart of deterrence (e.g. Ellsberg, 1961; Russett, 1963): is the possibility of a military response sufficient to convince the other side to live with the status quo? That possibility encompasses both the chance that the deterring party will not follow through on its threat and the evaluation of the consequences of military conflict in comparison to the attractiveness of the status quo.

Deterrence was more than just a judgment by the challenging state, that is, the party to be deterred. The party seeking to deter also had judgments of its own to make. If deterrence failed, it would have to decide whether to carry out its deterrent threat by going to war. Before the latent threat by the challenging state became immediate and pressing,

the prospective deterring state had to judge how to communicate its deterrent threat through words and actions. These decisions could also be represented by expected utility calculations, but now all three decisions would be interrelated. Working backwards, whether the deterrent state will carry out its threat influences the willingness of the party being deterred to challenge the status quo. In turn, the latter's inclination to challenge the status quo affects the nature and character of the deterrent threat before any challenge is made. How the deterrent threat is made, and then backed up by actions, changes the challenger's perception about whether the deterring state will carry out its deterrent threat. In short, deterrence involves a series of interrelated decisions between the two parties, some of which may never occur. Sequence is essential to the strategic interaction of deterrence, though previous work obscured that fact by examining one of these decisions in isolation, typically the judgment of the challenger.

Sequence and interdependence of decisions are two hallmarks of games in the extensive form. These games require the analyst to be explicit about the sequence of moves and how they produce outcomes. This discipline clarifies aspects of the problem that are often left vague in non-formal analyses of deterrence. If the game was played under complete and perfect information, backwards induction would produce the *subgame perfect equilibrium* of the game (assuming no player is indifferent at any move). The character of the resulting equilibrium coincides with many of the intuitions about deterrence in the non-formal literature, such as the importance of the credibility of the threat to effective deterrence. But that equilibrium would also create some enigmas. If the equilibrium produced behavior where the deterrent threat is made but fails to deter the challenger, one might wonder why an ineffective deterrent threat would be made at all. If that situation ended in violent conflict, both sides must have preferred war to the status quo. When conflict occurs in a game of complete information,

the parties prefer conflict to the alternative outcomes; it is Pareto optimal. Further, both sides could anticipate the outcome of the crisis from the beginning because both fully understand the other's incentives in the game. Why engage in a crisis, in that case? The deterring state should just go to war or yield the stakes without making a threat that it is unwilling to carry out. But a majority of historical crises do not end in war. Thus, sequential games can yield results that appear perverse.

Models of complete information could be used to examine when commitment problems give rise to conflict. The sequence of moves would include the possibility of one or both sides reneging on a deal made earlier in the game. Backwards induction would then render an efficient bargain unattainable because at least one side preferred conflict to entering a deal that will be broken later to its disadvantage. Although such models can capture a commitment problem, the field has gravitated to stochastic games (to be discussed later in this chapter) as the preferred modeling technique for commitment problems. The extensive form games under complete information effectively assume that one party will have the incentive to break the agreement later, while the opportunity and incentive to break the agreement are not certain in any given period in the stochastic game technology, even though the conditions that lead to breakdown are very likely to occur at some point in the future.

An alternative modeling strategy that could produce a positive probability of conflict under complete information would be to assume imperfect information, that is, simultaneous moves. The classic example here is the two-by-two game of chicken with three equilibria: two in pure strategies where the players avoid conflict, and a mixed strategy equilibrium with positive probability of conflict. Models in this line (cf. Morrow, 1994: 180–6) could represent situations of the mutual fear of surprise attack. Because the risk of conflict arises in mixed strategy equilibria, models that rely on imperfect information to produce a positive probability



of conflict are open to critiques of such equilibria, such as their lack of dynamic stability.

Models of complete and perfect information are an appropriate and useful tool for studying questions of distribution and welfare, even when they do not produce a positive probability of conflict. For example, we might wish to study how different domestic politics shift the bargain between two parties, and thereby provide a bargaining advantage to one of them. The international status quo is the result of a string of past bargains from prior efforts to change the status quo. Understanding what advantages different parties have in the bargaining allows us to characterize how the status quo should reflect those advantages. Fearon (2018) examines the welfare effects of interstate competition under anarchy, where the parties arm both in the shadow of war and political competition short of war, and where military capabilities influence the outcome of both. Parties then commit resources to the military to deter war, protect their future stream of income and gain in the political competition short of war. Because these military allocations vary with the parameters of the model, we can see when they are larger, producing a greater loss to the inefficiency of military expenditures. The model then predicts patterns within military expenditures across countries and time. While the point of the model is not to predict the occurrence of war, war breaks out only under conditions where deterrence is impossible.

## REPEATED GAMES

States and their leaders often carefully cultivate reputation, valuing the ability to communicate through repeated truth-telling or to induce compliance by threatening punishment. Unlike negotiating an agreement over a single issue, establishing reputation requires repeated interaction over time. Therefore, studying a phenomenon such as reputation requires a model of an iterative process.

Repeated games offer a means to study recurring interactions, representing the ongoing relationship between states as the repetition of a *stage game* – typically a simple, simultaneous game. This repetition enables players to condition their strategies on the history of play and to threaten to play an inefficient (equilibrium) strategy to punish undesirable behavior. Therefore, counterintuitive, and often cooperative, outcomes can be sustained in repeated games when players are able to threaten to play *minmax strategies* – strategies that minimize the maximum payoff of their opponents – should cooperation break down.

Repeated games can be used to represent many types of substantive processes where relations between states depend on past behavior. They can be used to explain normatively desirable, and not obviously rational, outcomes like diplomacy. For example, in the context of a repeated interaction between countries, states that build a reputation for honesty can improve their chance of settling issues they care about in their own favor, while those that bluff, relying too often on implausible claims, can be punished by being ignored (Sartori, 2002). In this way, repeated games allow the researcher to explore questions of reciprocity and identify conditions under which such cooperative outcomes are sustainable. Alternatively, repetition can explain inefficient outcomes such as the onset of interstate war. If war is viewed as a costly process, fighting may be optimal if refusing to fight would induce an undesirable settlement where the selection of that settlement was enforced in future plays of the game (Slantchev, 2003).

A downside of repeated games is the multiplicity of equilibria that are feasible, which gives rise to a well-known set of results known as folk theorems. All Nash equilibria of the stage game can be supported by folk theorems, as well as many strategies that generate feasible payoffs that combine different outcomes of the stage game. Therefore, while a number of substantively meaningful outcomes are possible in repeated games, a large

set of less interpretable results may also hold in equilibrium. In many cases, scholars confront this problem by selecting an equilibrium that corresponds with their substantive argument (Sartori, 2002) and emphasize the conditions under which this equilibrium holds. While this approach may suffice to answer a number of questions about reputation building, in many cases this practice of selecting equilibria is unsatisfying. Alternatively, the selection of a particular equilibrium from the wide range of them hinges on the *common conjecture* – the assumption that all players share an understanding about which equilibrium they are playing (Aumann and Brandenburger, 1995). Such common understandings of how players should act could arise from a number of sources. Morrow (2014) argues that international law embodies such understandings that allow the parties to form strategic expectations about how they will act in a given situation. This approach turns the focus of the analysis away from the incentives of the players within the game and towards the shared understanding about how they play. Because the mathematical tools of equilibrium theory center on the calculation of mutual best replies, there are few tools available to model where the common conjecture comes from. We explore an alternative method for analyzing interactions with an infinite horizon – stochastic games – that avoids the multiplicity of equilibria under the folk theorems.

## STOCHASTIC GAMES

An important way in which states manage relations with each other, and use their power to avoid or initiate conflict, is through making threats or promises. This form of strategic communication, however, is plagued by a pervasive problem – it is difficult for states to commit to agreements. Even in instances where governments take concrete steps, like making concessions, when their preferences may change over time, states that want to

make promises remain unable to guarantee they will keep them. This is a classic commitment problem. In a complete information environment, as mentioned above, commitment problems limit possible agreements to only those that are reached in subgame perfect equilibrium. A more flexible way to represent issues of commitment is through a stochastic game. In stochastic games, some underlying state of the world can change, and with it the incentives of the players. In this way, stochastic games capture the endogenous process of the interaction between states over time. Stochastic games rely on *Markov strategies* and *Markov perfect equilibria* (a refinement of subgame perfection). Markov strategies condition play on the state of the world (rather than the history of the game until that point, as in a repeated game) where the state of the world provides some kind of payoff-relevant information. Strategies that comprise a Markov perfect equilibrium are both subgame perfect and require players to play the same strategies from any history that leads to a given state.

Commitment problems arise when one (or both) side(s) of a potential conflict has incentives that shift over time. They may enter a bargain wanting to honor it, but when their incentives change, they no longer wish to. In anticipation of one side reneging in the future, the other may refuse any bargain now, even in situations where both sides are better off than if they were fighting. For this reason, commitment problems are commonly given as an explanation for inefficient outcomes, like the continuation of conflict (Fearon, 2004), coups (Acemoglu and Robinson, 2001a,b) and appeasement (Powell, 1996b). The commitment problems and the inefficient results they generate in these examples can be represented by a common mechanism (Powell, 2004b). In each of these cases, the amount of benefits that each of two players would need to be assured of obtaining in an efficient equilibrium exceeds the total amount of benefits available to be divided. Therefore, it is impossible to satisfy both players' claims on

the total value of benefits at the same time. Further, as one player loses power while the other gains it, the bargaining surplus cannot be sufficiently large to satisfy one of the players. Therefore, large and rapid changes in relative power between bargaining parties imply that commitment problems will lead to only inefficient outcomes.

Stochastic games are particularly useful for examining dynamic problems where information is not the primary concern. All of the illustrative stochastic games mentioned above have complete information. As their name indicates, stochastic games do require some randomness that can generate shifts in the state of the world, but the distribution of this uncertainty is typically common knowledge for all players. In addition, stochastic games improve on repeated games in many ways by allowing the modeler to avoid the problem of equilibrium selection. As Powell (2004b) demonstrates, all stochastic models with certain characteristics generate inefficient equilibria, namely a state where one of the parties can guarantee itself more by shifting to an inefficient path for the long run than it would receive otherwise, even if the other side offers all available surplus from an efficient path in the short run. This is a significant improvement over arguments built on repeated games that suggest inefficient outcomes must be the result of dominated inefficient equilibria. Lastly, the pervasiveness of commitment problems in interstate relations makes the stochastic game a particularly useful tool for examining the long-term relations between states. Those who wish to develop an explanation for conflict based in a commitment problem can use the condition in Powell (2004b) to ensure that efficient settlements will be undermined by that commitment problem.

## INCOMPLETE INFORMATION

Another source of conflict is differences in information between states. That is, how does conflict change when one state knows

something, perhaps about its own preferences or tolerance for conflict, that another state does not. Games of incomplete information allow scholars to approach these questions in a way that provides a disciplined framework for thinking about differences in information and beliefs about another party's actions.

For many students of international relations, the most familiar setting in which incomplete information plays a crucial role is crisis bargaining. Before initiating conflict, states attempt to bargain to find some negotiated settlement short of war. However, when information is incomplete, and states have incentives to misrepresent the information they have and their adversary lacks, conflict may occur even though it is inefficient. In many cases, the relevant information relates to a privately known cost for conflict or level of resolve. This is the key insight of Fearon (1995) and has been expanded upon in a number of ways, which we will cover in this section. We begin, however, by exploring how incomplete information changes the modeling enterprise.

An incomplete information environment is one in which at least one player has private information – known to themselves but not to other players – about their preferences or their actions at the time they make strategic decisions in the game. Information, for the other player(s), is incomplete because some key feature of the strategic environment is unknown to them. The canonical bargaining model of war incorporates incomplete information of this form because the costs for war of each player is privately known. Players with incomplete information, who have some uncertainty about the preferences of their opponent(s), must make choices based on beliefs about what the other players will do.

Importantly, incomplete information is distinct from the information imperfections that arise in some extensive form games, such as those with simultaneous moves. When information is imperfect, at the outset of the game all players know the same thing, but one or more players might get private information

throughout the game. Incomplete information implies asymmetry in information between players that arises before the game is played.

There are two ways in which information can be incomplete, and each lends itself to a different type of model. First, players can make private decisions or take private actions. This is often thought of as some exertion of effort that is unobserved by the other player(s) in the game. Such incomplete information problems are typically referred to as *moral hazard* problems, represented by an interaction between a principal and an agent. Moral hazard models are common in economics and political science but are often used to study interactions between individuals rather than states. Second, players can have private information about payoffs. In the context of international conflict, this could mean players have some private level of resolve, value for winning or cost for conflict, as is the case in Arena and Wolford (2012); Fearon (1997); Morrow (1989); Powell (1996a); Slantchev (2005); and Trager (2010). This type of information is relevant to the strategic decision at hand, and the player possessing this information cannot be compelled to truthfully reveal this component of his or her payoff unless presented with a set of incentives, designed by other player(s), that make it beneficial to do so. Incomplete information problems of this form are often referred to as *adverse selection*, and lend themselves to models of screening and signaling; notable models of this form include Morrow (1992); Powell (1988); Schultz (1998); Slantchev (2010); Tarar and Leventoglu (2009); and Trager (2011).

Moving to a world with incomplete information requires a different solution concept to characterize equilibria to this type of game, and a way to represent the particular aspect of a player's payoff or action that is unknown. The information about a player that is unknown is represented by a *type*, which stands for all the possible values of that player's private information. For example, if one player has a private cost for conflict that is either high or low, that player has

two types: one high-cost type and one low-cost type (for an example with continuous types, see Fearon, 1997). There are multiple substantive interpretations for types. Bueno de Mesquita and Lalman (1990) use types to represent the magnitude of domestic political opposition to the use of force against another state. In Powell (1996b), a declining state does not know a rising state's type, where the type represents the rising state's resolve.

Because a player's type affects his or her strategy, the uninformed player would like the informed player to reveal this private information. In some cases, the informed player also finds it optimal to communicate her private information. Games where the player with private information takes some action that demonstrates her type before the strategic interaction between the informed and uninformed player are referred to as *signaling*. When some action that reveals the informed player's type is taken after the strategic interaction with the uninformed player, the uninformed player is said to be *screening*. The key distinction between these two types of models is whether the player with the private information acts, leading to signaling, or responds to an act of the uninformed party, leading to screening. In an interstate conflict context, if a leader wants to show an adversary he is resolved, he may signal his resolution by staging military exercises. In contrast, when bargaining with an adversary of unknown strength, a leader may want to choose the proposed division of resources so that by accepting, the adversary reveals whether she is a strong or a weak type.

For static games of incomplete information where the players move before any private information can be revealed, the appropriate solution concept is Bayesian equilibrium. A Bayesian equilibrium is a Nash equilibrium in which players update their beliefs according to Bayes's Rule. Therefore, a Bayesian equilibrium is a strategy profile and beliefs, for each player, about the types of the other player(s) that maximize the expected payoff for every player, given the other players' strategies and beliefs. For sequential games,

we refine this solution concept to *perfect Bayesian equilibrium*. A perfect Bayesian equilibrium is defined by a set of strategies for each type of each player and a set of beliefs or conjectures such that the chosen actions maximize conditional expected utility based on the beliefs about other players' types.

Returning to the crisis bargaining example, in the canonical crisis bargaining model articulated in Fearon (1995), there always exists a set of negotiated settlements which both potential belligerents would prefer to war. However, incomplete information about conflict-relevant factors such as military capability or resolve – and in particular, incentives to misrepresent that information in order to attain a more favorable bargain – lead to the breakdown of negotiation into war. Entering an incomplete information environment, then, is precisely what allows the modeler to develop a rationalist explanation for why wars occur when they are inefficient *ex post*.

The role information plays in leading to bargaining failure is reinforced by the take-it-or-leave-it structure of the bargaining protocol. This particular way of representing bargaining is well suited to modeling informational mechanisms (as opposed to, for example, an alternating offers protocol) because it allows the modeler to focus attention away from the dynamics of the negotiation. Further, a take-it-or-leave-it protocol has two desirable properties that align substantively with an intuitive notion of crisis bargaining. First, while the breakdown of bargaining into conflict is inefficient, there is a positive probability of bargaining failure, and thus war can arise rationally. Second, the state making the offer faces a risk–return tradeoff, which in part generates the incentive for the second state to misrepresent its cost for conflict. The offering country can give up more and reduce the odds of that offer being rejected, but that necessarily requires accepting a smaller share of the divided benefits. For the second country, misrepresenting its cost of conflict is preferred if doing so will lead to a greater share of the benefits – specifically

because the offering country then thinks it needs to offer more to avoid conflict.

Incomplete information generates a rich set of questions that can be addressed with a number of different models to treat aspects of interstate conflict ranging from the choice to go to war (e.g. Fearon, 1995) to the reaction of domestic constituents to interstate conflict (e.g. Schultz, 1998). In this section, we cover models built on the workhorse take-it-or-leave-it bargaining model, which leverage incomplete information in three different ways: models of mechanism design, games of bargaining during war and models that test the bounds of common knowledge. A number of other incomplete information models, including signaling games, are covered elsewhere in this volume.

### ***Mechanism Design***

The language of incentives is often used to explain why states choose to engage in interstate conflict. However, the precise role of these incentives, and what exactly they say about states' motivations and concerns, is often unclear. Models of mechanism design formally enumerate states' incentives and ask under what circumstances it is sufficiently beneficial for states to take some action. Moreover, mechanism design models demonstrate how an uninformed state can induce an opponent to reveal important information by taking advantage of the opponent's incentives.

While game theoretic models are designed to make predictions about strategic behavior taking the rules of the game as given, mechanism design concerns how the incentives of the situation limit what is possible under any set of rules for the game. A mechanism is a mapping from the private information of the players into outcomes. In a standard mechanism design setting, the mechanism designer proposes a mechanism that seeks to elicit truthful revelation of information from the players, each player reports its private information to the designer and then the designer

applies the mechanism to the revealed information to determine the outcome. Each player agrees to play the game and tell the truth about her private information if a set of conditions on incentives are met; these incentives reflect equilibrium strategies in some Bayesian game. An important result in mechanism design, referred to as the Revelation Principle, ensures that any set of outcomes that can be produced under a Bayesian Nash equilibrium of the game can also be produced under a mechanism which induces truthful revelation of information to the mechanism designer. More specifically, universal honesty is a Bayesian Nash equilibrium if and only if the choice of mechanism satisfies necessary and sufficient conditions on incentives (Myerson, 1979, 1983). Thus, a mechanism design approach shows the limits of what can be achieved under a set of incentives independent of the specific game being played.

Mechanism design models require that equilibrium strategies satisfy two constraints on incentives. The first is *individual rationality*. Often referred to as a participation constraint, this condition requires that the value of an action must be sufficiently high such that it is better than taking no action. The value of a player's best outside option is referred to as her reservation value, and satisfying individual rationality requires that the benefit from playing a particular strategy must equal or exceed this value.<sup>1</sup> Second, a strategy must satisfy *incentive compatibility*, meaning the strategy must be optimal given the expected costs and benefits of playing that strategy.<sup>2</sup>

Because incentive compatibility represents a condition that must be met for a strategy to be optimal, such constraints arise independently from any game form. Banks (1990) takes a 'game free' approach to understanding the role of incentive compatibility in crisis bargaining. His results therefore hold for any equilibrium of any game with incomplete information, and solely rely on incentive compatibility conditions. Considering a generic bargaining scenario in which war is possible and one side has private information

regarding its cost for war, he identifies principles of the outcomes that must be true in any game that contains those incentives. The probability of war and the expected benefits from reaching a negotiated agreement that avoids war are both increasing in the expected benefit of conflict for the informed player. More resolute types of states – that is, those with private information that increases the value of war for them – are more likely to fight interstate wars but derive higher benefits from settlement if war is averted. Importantly, this finding is only the result of requiring that states act in a manner that is incentive compatible and individually rational. Therefore, these results help scholars of interstate conflict better understand the decision-making environment of crisis bargaining, rather than decision making within the context of a particular model specification.

While the results derived from employing incentive compatibility conditions alone provide valuable insight about the likelihood of war, combining this type of result with a formalization of key features of the information environment can provide significant, general insights about the onset of interstate conflict. Fey and Ramsay (2007) use this approach to question the logic of the mutual optimism explanation for war. Mutual optimism implies that both states on the brink of war believe they are more likely to win (or, more precisely, that the sum of their subjective probabilities of victory is greater than one). This would require that war is simultaneously incentive compatible for both states, given their beliefs about their relative probability of victory. Fey and Ramsay (2007) consider a generalized negotiation procedure because, relying on the Revelation Principle, any equilibrium for any game chosen to represent the negotiation procedure can be represented by the state of the world (a distribution over the probability that one side wins a war) and the outcomes of this negotiation protocol. In this setting, they show there is no equilibrium under which war occurs due to mutual optimism, extending to cases where the states involved can be considered

only boundedly rational.<sup>3</sup> This is precisely because, as long as war's occurrence is self-evident to all players, war cannot be incentive compatible for both rational, Bayesian players at the same time.

Because the conditions under which a mechanism is optimal are not based on a particular game form, the mechanism design approach is a highly general way to examine interstate war in an incomplete information environment. As Banks (1990) and Fey and Ramsay (2007) demonstrate, models that rest on questions of mechanism design are particularly well suited to examine when interstate bargaining is individually rational and war is incentive compatible, that is, what the necessary conditions are for states to be willing to bargain over contested issues, and how high the expected benefits from war must be for war to be observed in equilibrium.

### ***Bargaining during War***

The types of incomplete information models discussed so far offer each actor only one opportunity to update their beliefs about the state of the world through a one-shot interaction such as bargaining. However, a more realistic representation of both the bargaining process leading up to war and the process of fighting wars may be to represent war as a costly process in which states have many opportunities to incrementally update beliefs about their opponents. Models that focus on learning are particularly well suited to answering questions about how war unfolds when states disagree about the distribution of power between them, rather than their private cost for war (Slantchev, 2003; Powell, 2004a).<sup>4</sup> In addition, models in which players learn over time can be used to assess the effects of shifts in power between two states. In these models, states attempt to resolve uncertainty about the relative decline and rise of their power, and its likely effect on the outcome of war (Powell, 2012; Wolford, Reiter and Carrubba, 2011).

A general result generated by models of learning through fighting is that the uncertain state makes offers intended to screen, to gain information about the informed state's type. Types with higher cost for conflict will settle sooner and accept smaller distributions of the prize than low-cost types. Thus, if war progresses as a series of battles with bargaining rounds between them, the uninformed state can develop an increasingly refined view of the informed state's privately held cost, based on the value of offers the informed state is willing to accept. This screening dynamic is captured in Powell (2004a), and arises regardless of whether the source of uncertainty is the cost of fighting or the underlying distribution of power between belligerents. However, the source of incomplete information does change the uninformed state's ability to screen when parties can make rapid offers before or between battles. When there is uncertainty over the cost of fighting, a process of rapidly made offers can avert war, whereas when uncertainty is over the underlying distribution of power, the states must fight a battle in order for the uninformed state to screen. This is because fighting reveals a different kind of information in this rapid bargaining environment. When uncertainty is over costs, each type distinguishes itself by accepting agreements it finds preferable to war. Therefore, the uninformed state can make offers such that every type agrees and war is avoided. However, when uncertainty is over the distribution of power, multiple types can have the same cost for conflict, and the uninformed state is unable to distinguish between these types without fighting.

Costly process models effectively represent learning by enabling players to update multiple times, refining their beliefs about the true state of the world. These types of model are therefore useful for thinking about how the specification of the bargaining environment shapes results. Relaxing some of the standard assumptions of the bargaining model, such as allowing for rapid offers, as in Powell (2004a), or incorporating an inability to commit to the negotiated distribution of

the prize, as in Wolford, Reiter and Carrubba (2011), generate equilibria in which war is optimal even though it is still destructive; these are qualitatively different than the equilibria in standard one-shot bargaining models.

### **Common Knowledge**

Most games of incomplete information make an assumption about the beliefs of players before the game is played, referred to as the *common priors* assumption. Common in this case implies something that every player knows, and that everyone knows that every player knows, and this mutual understanding continues infinitely (Aumann, 1976). Beliefs held before the introduction of new information are referred to as 'prior' or initial beliefs, represented by subjective probabilities of the occurrence of a given set of events. In the incomplete information context, common priors provide elements of the game that are common knowledge and are used by players to inform beliefs about types. In most models in which information is incomplete, it is this common prior that players update using Bayes's Rule.

It is not strictly necessary (though it may be technically and epistemologically desirable) to make the common priors assumption. Non-common priors convey differences in beliefs in cases where there is no difference in information between players. In Smith and Stam (2004), non-common or heterogeneous priors represent differing theories of war between bargaining states. For example, in Smith and Stam's model, conceptual differences about the role of a new technology on the battlefield or the importance of unit cohesion among troops generate divergence in beliefs about the likely outcome of war, despite the fact that both parties observe the same information, namely the outcome of battles. Smith and Stam argue that diverging from the common priors assumption better represents empirical realities. Moreover, they contend that relaxing the common priors assumption provides an opportunity to ask

questions about where beliefs come from, and how that influences decision making (Smith and Stam, 2006).

Modelers interested fundamentally in the role of information and learning, not in understanding the effects of actors having fundamentally different views of the world, are likely better off adopting the common priors assumption. It is possible to represent different world-views and produce divergence in beliefs under the common priors assumption, and thus this approach is arguably more suitable for anyone interested in questions of how knowledge and information relate to the onset of war (Fey and Ramsay, 2006).

### **MODELS LINKED TO DOMESTIC POLITICS**

Links between international conflict and domestic politics have been an important research topic for several decades now, and models are a primary research strategy for these arguments. The range of topics addressed include effects of international conflict on leader tenure in democracies and autocracies (Bueno de Mesquita et al., 2003), how those incentives influence how leaders act in crises (Smith, 1998) and how a legislature can influence foreign policy (Morrow, 1991). These models link an international game to a domestic politics game to examine how the incentives in one arena of politics affects what politicians do in the other. The key is that conflict decisions are made in part for their domestic political consequences, and domestic politics responds to the outcomes of international conflicts. We exclude those models that allude to the effects of domestic politics on international conflict but do not model the domestic politics directly (e.g. Fearon, 1994; Schultz, 1998).

Many of these models examine democratic politics because there are a wide range of accepted models to represent different aspects of democratic politics, such as elections and



legislation. Those models of domestic politics seek to show how democratic processes influence policy outcomes, which makes them suitable to link to international policies and outcomes. The mechanism linking policy to politics is often retrospective voting, giving national leaders reasons to produce successful outcomes to conflicts. Ideological models of elections, such as spatial models, are less common because foreign policy in the US has not been viewed as an ideological issue for decades.

Autocratic politics poses a challenge for this modeling strategy because of the nature of models of autocratic politics. While a range of models have been developed over the last two decades to analyze autocratic politics, they tend to focus on whether the autocrat can sustain himself (sorry, but dictators are almost always men) in power, not on the incentives of the system to produce specific policies (see Gehlbach et al., 2016 for a survey of these models). Models of autocratic politics then are quiet about why autocrats pursue the policies they pursue in office, beyond the simple desire to maintain power (e.g. Svobik, 2012). The prominent exception here is selectorate theory (Buono de Mesquita, 2003), but the application of that theory and model to international conflict requires a translation of the ends of international conflict into the dichotomy of public goods versus private benefits, where autocrats tend to produce private benefits over public goods because they answer to a small winning coalition. This general result focuses our attention on the distribution of rents as key to the maintenance of the leader's support coalition, which is common in other models that focus on the credibility problem inherent in the provision of private benefits in autocracies.

## **GAMES OF TIMING AND WARS OF ATTRITION**

Games of timing have a simple strategic structure. Each player can stop the game any time after it starts, with payoffs determined

by when the game stopped and which side stopped it. Their strategies are simply a probability distribution over stopping times. There are games of timing played in discrete time, as well as continuous time.

Wars of attrition are a particular type of game of timing where two sides vie for a prize. Once the game starts, both sides accumulate costs – the attrition of the war. The contest ends when one side yields the prize to the other. Both sides may be willing to continue the war with added costs in the hope that the other side will yield before they do. These games generalize the Dollar Auction (Shubik, 1971; O'Neill, 1986) where both sides pay their last bid in the auction. The side with the low bid may increase its bid in an effort to secure the item even to the point where the winning bid is greater than the value of the object. Because the Dollar Auction shares some features with international crises, namely that the side that backs down suffers some loss beyond the stakes, wars of attrition have been used to model international crises. Fearon (1994) modifies the continuous-time war of attrition by having only the loser, not the winner, pay its cost accumulated over the length of the game, which he famously labeled as an audience cost. Wars of attrition can be used as a model of negotiations at an impasse where the key step is which side makes a big concession to end the impasse. There is some evidence that wars of attrition represent the distribution of the duration of strikes – which have a similar strategic structure to wars – better than other models of bargaining that allow for haggling over the terms of the bargain (Kennan and Wilson, 1990).

Discrete wars of attrition have been used to model war, where each round is considered an individual battle (Smith, 1998). The two sides fight over a finite string of positions on a line; the endpoints represent the complete defeat of one side or the other, and victory in a battle pushes the outcome one position closer to the defeat of the side that lost that battle. Both sides suffer costs from each battle as well, and can

choose to quit to avoid further losses if they think their chance of winning enough battles in the future to win the war is not large enough to justify the additional costs. In equilibrium, each side has a breakpoint – a position close to their total defeat where they quit if the war reaches that position. Morrow (2014: chapter 3) provides a complete solution to this model of war and uses it as part of a model of the laws of war. These discrete wars of attrition differ from the models of bargaining while learning in two ways: one, they are played under complete information; two, the parties cannot divide the stakes.

## **BLOTTO GAMES**

Blotto games, one of the oldest types of game, concern the strategic allocation of resources in a competitive situation. Canonically, two players simultaneously distribute fixed resources, such as identical military units, across a number of locations which can be thought as battles to be fought. The side that allocates the most units to a location wins that battle, and both sides seek to win as many battles as they can. If all battles are valued equally by both sides, the resulting game is zero-sum. If a Blotto game has an equilibrium, it is in mixed strategies. If one side played a pure strategy, the other would allocate no units to some locations – conceding them – and exactly enough units to win all the other battles, doing so to maximize the number of battles it wins. But then the first side's strategy is no longer a best reply – the side playing the pure strategy has an incentive to deviate to win more battles itself. Mixed strategies produce uncertainty about how much strength each side is required to commit to a given location in order to win the battle there, which can make these mixed strategies mutual best replies. The calculation of those mixed strategies can be complicated. See Golman and Page (2009) for generalized Blotto games and

more detail on the nature of equilibria in these games.

When the same setup is played sequentially, the nature of play changes. The leading player must allocate her resources first, and the trailing player can then target the weakest locations. In equilibrium, the leading player divides its units across the locations to make them equally attractive targets, thereby leveling their values. The trailing player then mixes among the leveled locations because they are all equally attractive targets and to prevent the leading player from anticipating which locations will be targeted. Powell (2007a,b) applies the sequential Blotto setup to the problem of protecting targets against terror attacks. The government allocates its resources across targets to protect them, and then the terror group selects which targets to attack.

Blotto games offer a way to model conflicts that could be placed within a wider model of why states choose conflict. The mixed strategies that commonly are in equilibrium induce probabilities of victory, which in turn matter in both sides' continuation values for war in the prewar moves. We see this as unexploited terrain for models of conflict and why it occurs.

## **CONCLUSION**

Formal modelers have developed a rich toolbox for examining the logic of causal processes in detail. Because the discipline of modeling requires researchers to focus on one causal process in isolation from others, they should make conscious decisions about the process they wish to analyze. We have outlined how different types of models embody broad answers to questions about why international conflict occurs. Further, we have emphasized a couple key questions about the process of interest whose answers suggest some types of models are more suitable than others. The modeler must consider

who the relevant actors are (two states, many states, domestic constituencies), how much the players know about each other, whether they are motivated by short-term incentives or long-term considerations, and if are they preparing for war (or avoiding it) or in the midst of it. The decision as to the appropriate model ultimately rests with the researcher herself.

## Notes

- 1 For adverse selection problems, this is sometimes referred to as the self-selection constraint.
- 2 For a more detailed technical treatment of the basics of mechanism design, see Börgers (2015).
- 3 Bounded rationality implies that players suffer from some type of cognitive bias that impacts their strategies.
- 4 For a treatment of similar questions in a complete information environment, see Leventoglu and Slantchev (2007). Ramsay (2008) offers an empirical analysis of similar questions and tests some of the implications of models of war as a costly process.

## REFERENCES

- Acemoglu, Daron, and James A. Robinson. 2001a. Inefficient Redistribution. *The American Political Science Review* 95 (3): 649–661.
- Acemoglu, Daron, and James A. Robinson. 2001b. A Theory of Political Transitions. *The American Economic Review* 91 (4): 938–963.
- Arena, Philip, and Scott Wolford. 2012. Arms, Intelligence, and War. *International Studies Quarterly* 56 (2): 351–365.
- Aumann, Robert J. 1976. Agreeing to Disagree. *The Annals of Statistics* 4 (6): 1236–1239.
- Aumann, Robert, and Adam Brandenburger. 1995. Epistemic Conditions for Nash Equilibrium. *Econometrica* 63 (5): 1161–1180.
- Banks, Jeffrey S. 1990. Equilibrium Behavior in Crisis Bargaining Games. *American Journal of Political Science* 34 (3): 599–614.
- Börgers, Tilman. 2015. *An Introduction to the Theory of Mechanism Design*. New York: Oxford University Press.
- Bueno de Mesquita, Bruce, and David Lalman. 1990. Domestic Opposition and Foreign War. *The American Political Science Review* 84 (3): 747–765.
- Bueno de Mesquita, Bruce, and James D. Morrow. 1999. Sorting Through the Wealth of Notions. *International Security* 24 (2): 56–73.
- Bueno de Mesquita, Bruce, Alastair Smith, Randolph M. Siverson, and James D. Morrow. 2003. *The Logic of Political Survival*. Cambridge, MA: MIT Press.
- Ellsberg, Daniel. 1961. The Crude Analysis of Strategy Choices. *The American Economic Review* 51 (2): 472–478.
- Fearon, James D. 1994. Domestic Political Audiences and the Escalation of International Disputes. *The American Political Science Review* 88 (3): 577–592.
- Fearon, James D. 1995. Rationalist Explanations for War. *International Organization* 49 (3): 379–414.
- Fearon, James D. 1997. Signaling Foreign Policy Interests: Tying Hands versus Sinking Costs. *The Journal of Conflict Resolution* 41 (1): 68–90.
- Fearon, James D. 2004. Why Do Some Civil Wars Last so Much Longer than Others? *Journal of Peace Research* 41 (3): 275–301.
- Fearon, James D. 2018. Cooperation, Conflict, and the Costs of Anarchy. *International Organization* 72 (3): 523–559.
- Fey, Mark, and Kristopher W. Ramsay. 2006. The Common Priors Assumption: A Comment on ‘Bargaining and the Nature of War’. *The Journal of Conflict Resolution* 50 (4): 607–613.
- Fey, Mark, and Kristopher W. Ramsay. 2007. Mutual Optimism and War. *American Journal of Political Science* 51 (4): 738–754.
- Gehlbach, Scott, Konstantin Sonin, and Milan W. Svobik. 2016. Formal Models of Nondemocratic Politics. *Annual Review of Political Science* 19 (1): 565–584.
- Golman, Russell, and Scott E. Page. 2009. General Blotto: Games of Allocative Strategic Mismatch. *Public Choice* 138 (3/4): 279–299.
- Kennan, John, and Robert Wilson. 1990. Can Strategic Bargaining Models Explain Collective Bargaining Data? *The American Economic Review* 80 (2): 405–409.
- Leventoglu, Bahar, and Branislav L. Slantchev. 2007. The Armed Peace: A Punctuated

- Equilibrium Theory of War. *American Journal of Political Science* 51 (4): 755–771.
- Morrow, James D. 1989. Capabilities, Uncertainty, and Resolve: A Limited Information Model of Crisis Bargaining. *American Journal of Political Science* 33 (4): 941–972.
- Morrow, James D. 1991. Electoral and Congressional Incentives and Arms Control. *The Journal of Conflict Resolution* 35 (2): 245–265.
- Morrow, James D. 1992. Signaling Difficulties with Linkage in Crisis Bargaining. *International Studies Quarterly* 36 (2): 153–172.
- Morrow, James D. 1994. *Game Theory for Political Scientists*. Princeton, NJ: Princeton University Press.
- Morrow, James D. 2014. *Order within Anarchy: The Laws of War as an International Institution*. New York: Cambridge University Press.
- Myerson, Roger B. 1979. Incentive Compatibility and the Bargaining Problem. *Econometrica* 47 (1): 61–73.
- Myerson, Roger B. 1983. Mechanism Design by an Informed Principal. *Econometrica* 51 (6): 1767–1797.
- O'Neill, Barry. 1986. International Escalation and the Dollar Auction. *The Journal of Conflict Resolution* 30 (1): 33–50.
- Powell, Robert. 1988. Nuclear Brinkmanship with Two-Sided Incomplete Information. *The American Political Science Review* 82 (1): 155–178.
- Powell, Robert. 1996a. Stability and the Distribution of Power. *World Politics* 48 (2): 239–267.
- Powell, Robert. 1996b. Uncertainty, Shifting Power, and Appeasement. *The American Political Science Review* 90 (4): 749–764.
- Powell, Robert. 1999. *In the Shadow of Power: States and Strategies in International Politics*. Princeton, NJ: Princeton University Press.
- Powell, Robert. 2004a. Bargaining and Learning While Fighting. *American Journal of Political Science* 48 (2): 344–361.
- Powell, Robert. 2004b. The Inefficient Use of Power: Costly Conflict with Complete Information. *The American Political Science Review* 98 (2): 231–241.
- Powell, Robert. 2007a. Allocating Defensive Resources with Private Information about Vulnerability. *The American Political Science Review* 101 (4): 799–809.
- Powell, Robert. 2007b. Defending against Terrorist Attacks with Limited Resources. *The American Political Science Review* 101 (3): 527–541.
- Powell, Robert. 2012. Persistent Fighting and Shifting Power. *American Journal of Political Science* 56 (3): 620–637.
- Ramsay, Kristopher W. 2008. Settling It on the Field: Battlefield Events and War Termination. *The Journal of Conflict Resolution* 52 (6): 850–879.
- Russett, Bruce M. 1963. The Calculus of Deterrence. *The Journal of Conflict Resolution* 7 (2): 97–109.
- Sartori, Anne E. 2002. The Might of the Pen: A Reputational Theory of Communication in International Disputes. *International Organization* 56 (1): 121–149.
- Schultz, Kenneth A. 1998. Domestic Opposition and Signaling in International Crises. *The American Political Science Review* 92 (4): 829–844.
- Shubik, Martin. 1971. The Dollar Auction Game: A Paradox in Noncooperative Behavior and Escalation. *The Journal of Conflict Resolution* 15 (1): 109–111.
- Slantchev, Branislav L. 2003. The Power to Hurt: Costly Conflict with Completely Informed States. *The American Political Science Review* 97 (1): 123–133.
- Slantchev, Branislav L. 2005. Military Coercion in Interstate Crises. *The American Political Science Review* 99 (4): 533–547.
- Slantchev, Branislav L. 2010. Feigning Weakness. *International Organization* 64 (3): 357–388.
- Smith, Alastair. 1998. International Crises and Domestic Politics. *The American Political Science Review* 92 (3): 623–638.
- Smith, Alastair, and Allan C. Stam. 2004. Bargaining and the Nature of War. *The Journal of Conflict Resolution* 48 (6): 783–813.
- Smith, Alastair, and Allan C. Stam. 2006. Divergent Beliefs in 'Bargaining and the Nature of War': A Reply to Fey and Ramsay. *The Journal of Conflict Resolution* 50 (4): 614–618.
- Svolik, Milan W. 2012. *The Politics of Authoritarian Rule*. New York: Cambridge University Press.

- Tarar, Ahmer, and Bahar Leventoğlu. 2009. Public Commitment in Crisis Bargaining. *International Studies Quarterly* 53 (3): 817–839.
- Trager, Robert F. 2010. Diplomatic Calculus in Anarchy: How Communication Matters. *The American Political Science Review* 104 (2): 347–368.
- Trager, Robert F. 2011. Multidimensional Diplomacy. *International Organization* 65 (3): 469–506.
- Wolford, Scott, Dan Reiter, and Clifford J. Carrubba. 2011. Information, Commitment, and War. *The Journal of Conflict Resolution* 55 (4): 556–579.



# Models of the Judiciary

Deborah Beim

## PREFACE

This chapter provides an overview of how courts work. It focuses on American federal courts, and takes the perspective of contemporary analytical theory. In other words, it reviews game-theoretic models of American courts. The chapter's structure follows the path of a case through the courts. These paths are long (see the literary description of how long in *Jarndyce and Jarndyce* as told by Charles Dickens (1853)). As a result, the chapter necessarily skims over each step, and unfortunately leaves aside many relevant and important models, papers and treatises. My hope is that the chapter, targeted at scholars of other institutions, will provide a cursory overview of how game theorists think about courts. If the chapter is successful, game theorists who think about other ideas and other branches of government will feel equipped to join the study of judicial politics.

## INTRODUCTION: HOW COURTS RESEMBLE AND DIFFER FROM LEGISLATURES

Courts, like legislatures, are (often multi-member) institutions that 'write laws'. Political scientists have a strong theoretical grasp of how legislatures work. We can infer from that theory a bit about how courts work. But theory about legislatures is of limited use in modeling judicial behavior, because courts also differ from legislatures.

Courts differ (and are structured differently) from legislatures. A court is an institution that resolves disputes. Common-law courts, like those we have in the United States, write opinions that describe how similar disputes should be resolved in the future. They write laws *ex post*, rather than *ex ante* as legislatures do. They are often hierarchical. They do not have complete control over their agenda (or docket, in legal parlance). Their decisions are bound by precedent. And, finally, they are inherently adversarial.

It is now broadly acknowledged that courts are policy-makers (see Dahl, 1957). The consensus around this fact has many positive consequences for scholarship on the judiciary: it welcomes political scientists to study the third branch of American government, and brings to bear a *realpolitik* perspective.

But focusing too exclusively on this function is mistaken.

Fundamentally, courts resolve disputes. American courts can make policy only by deciding cases – that is, in constitutional terms, by resolving cases and controversies. In other words, although courts are policy-makers, courts (unlike legislatures) do not *only* write laws. Nearly any opinion issued by an American court is accompanied by a simultaneous resolution of a dispute.

Legislatures typically set their own agenda – they choose which issues require their attention. They then write and issue laws that govern future behavior. In principle, a piece of legislation could be passed at one day and govern behavior indefinitely (though problems such as legislative decay and drift can change how a law is actually applied – not to mention that future legislatures can change laws when and as they see fit). Some models of judicial politics conceive of courts as legislatures. This class of models is often referred to as residing in a ‘policy space’, in contrast to a ‘case space’. Such a distinction arises from the recognition that legislatures issue policy (via legislation) while courts resolve cases. But as this distinction suggests, there are two fundamental differences between judicial policy-making and legislative policy-making.

First, cases are ‘vehicles’ for policy-making. The specific facts of a case may matter. The fact that the court must resolve a dispute while issuing its policy may introduce strategic complexities into bargaining and policy-making. Second, by construction, common-law policy-making is slow, iterative and inductive. I discuss each of these differences as we follow the path of a case through the courts.

The ways in which judicial institutions differ from legislative institutions have consequences for judicial behavior. In this chapter I plan to discuss some of these institutional details, highlighting both those that have been well explored by contemporary formal models and those that I think deserve more attention. I focus almost exclusively on models written by political scientists in the field of judicial politics. I omit models from the economic analysis of law, excepting a few papers sufficiently relevant to the political science subfield of judicial politics to warrant inclusion.

Two good recent review essays cover related material. Cameron and Kornhauser’s review essay of 2017 focuses on the Supreme Court and answers two specific questions: how to model what the Supreme Court does (statutory interpretation, administrative law and constitutional review) and how to model what Supreme Court justices want. Kastellec’s review essay of the same year explores the judicial *hierarchy*, including game theoretic models. In contrast, my chapter tries to give a quicker overview of all aspects of a case’s trajectory through American courts.

### ***Background on the Court System to Which Litigants Go When They Dispute***

In principle, people who find themselves in disputes can resolve those differences without recourse to a judiciary. This system – ‘order without law’ (Ellickson, 1994; Milgrom et al., 1990) – relies on communication and behavioral norms between participants in a society. But most states provide courts for this purpose.

In the American system, the courts that are state-provided are adversarial. In other words, truth-seeking is facilitated by two advocates each arguing their own position, and an adjudicator – a judge – evaluating the strength of each’s argument.

### ***How Does the Adversarial Nature of the Judicial Process Affect its Work?***

Dewatripont and Tirole (1999) show that under most conditions an adversarial system leads to discovery of the truth more often, and at less expense, compared to an inquisitorial system. In inquisitorial systems, common in the French legal tradition, the state pays an inquisitor to discover all the relevant facts. In contrast, in the American adversarial system, each of two lawyers discovers and presents the facts that are favorable to her client's position. Priest and Klein (1984) argue that the adversarial nature of trials means plaintiffs should win about 50% of their cases. That is to say, adversarialism affects which cases enter the courts and affects truth-seeking when in court.

### **LEARNING ABOUT LAW AND STARE DECISIS**

The common-law system works by judges reasoning by analogy from previously decided precedents to new fact patterns. There is a small amount of work on reasoning by analogy – mostly one-dimensional spatial models where nearer points are more similar than further points (examples include Gilboa and Schmeidler, 1995, Callander and Clark, 2017, Fox and Vanberg, 2014, and Baker and Mezzetti, 2012, among others).

But there are very few applied models about reasoning by analogy in common-law courts. A fully fleshed out model would incorporate work on incomplete contracts and work on unawareness.

Judicial opinions are 'stickier' than legislative bills. This status quo bias is known as precedent or *stare decisis*. *Stare decisis* creates complex situations for foresighted judges who know their opinions will be applied to future, unknown circumstances. It also creates strategic complexities for judges who are

bound to follow what their predecessors have done, even when their ideologies differ.

Bueno de Mesquita and Stephenson (2002) explore the rational foundations for the institution of *stare decisis*. They argue that *stare decisis* makes doctrinal communication easier: by relying on a string of past cases, judges can communicate with greater accuracy and ensure greater compliance with doctrine. In a game-theoretic sense, the judge faces a mean-variance trade-off. Judges may value this increase in precision so highly that they are willing to sacrifice some ideological accuracy.

Other papers consider complexities surrounding the implications of *stare decisis* for doctrinal development. For example, Gennaioli and Shleifer (2007) consider the consequences of *stare decisis* in a multi-dimensional world – specifically, how a judge will craft policy knowing that policy will bind future judges but knowing that future judges may craft permanent policies on related dimensions. In another example, Callander and Clark (2017) use a Brownian motion model to understand how *stare decisis* influences similar cases without influencing dissimilar cases.

### **THE CASE SPACE INTRODUCED**

American judge-made law is made by deciding individual cases. This raises complexities in the decision-making process which have been well explored by formal theorists. The technology of the 'case space' allowed scholars to explore this formally (see Lax, 2011 for a review and discussion of the case space).

Kornhauser (1992) introduced the idea that a legal case is a point in a fact space, where each dimension corresponds to a legal question of fact. In nearly all political science models, cases are represented as points on a one-dimensional fact space. Judges *dispose of* cases by choosing between two



actions – sometimes understood as a liberal versus a conservative decision, or a decision for the plaintiff versus the defense, or a decision for the prosecution versus the defendant. All judges dispose of cases, irrespective of where they sit in the judicial hierarchy and irrespective of whether they are ‘making law’ or not. Judges’ preferences are defined by *indifference points*, where an indifference point is a case at which the judge is indifferent between his two choices. *Doctrine*, then, is a rule for future resolution of disputes: in a one-dimensional case space, a doctrine is a cutpoint that describes which plaintiffs should win and which defendants should win. Some judges have the capacity to set doctrine in addition to issuing a disposition.

This is the crux of what is unique about judicial behavior as compared to legislative behavior: the resolution of one dispute, accompanied by a statement predicting how similar disputes would be resolved in the future, is how courts make law. Legislatures make law in the abstract, without the elements of reasoning by analogy, *stare decisis* and resolution of individual disputes that typifies judicial decision-making.

## CURRENT AND FUTURE RESEARCH WITHIN THE CASE SPACE

The case-space architecture allows scholars to understand the attributes of judicial decision-making that are distinct from abstract legislative policy-making. For example, the case space makes concurring and dissenting opinions sensible. Concurring and dissenting opinions, in case-space terms, represent judicial calls (declarations) for alternative cutpoints. A *dissent* disagrees with the disposition for a particular case, because the judge argues for a different cutpoint – particularly, a cutpoint resulting in a decision for what became the losing party in a case. A *conurrence* agrees with the disposition for a case, but would have used a

different cutpoint, although one resulting in a decision for the ultimately prevailing party. In simple spaces, doctrine can be simple. For example, minors cannot be executed (*Roper v Simmons* 543 U.S. 551). Formally, adjudication under this doctrine can be modeled in simple steps: is the punishment execution? Is the defendant a minor?

The case space also allows modeling of complex doctrines. For example, consider (in somewhat simplified terms) the question of which statutes limiting individual rights pass constitutional muster. The standard according to which courts evaluate a statute varies according to various factors, including the category of individuals whose rights the statute limits. For the most part, courts evaluate only whether there is a rational basis for the challenged statute. But where the law potentially infringes on certain special rights, the court scrutinizes said law more strictly. This idea is known as ‘strict scrutiny’. Famously, infringing on a person’s rights on account of his race is strictly scrutinized. Infringing on a person’s rights on account of her gender is not *strictly* scrutinized but is also not afforded as much leeway as would be typical. This is called ‘intermediate scrutiny’.

The case space allows the formal modeling of the adjudication of a statute challenged as unconstitutionally rights-infringing: is a statute challenged? What rights does it infringe? The complexity is in an additional step, which depends on the answer to the second: is there a rational basis for the statute *or* does it pass intermediate *or* strict scrutiny? In other words, the case space can consider which rights were infringed and thus map to scrutiny levels; the scrutiny then maps to which infringements are nevertheless permissible and which fail to meet the differential standards.

A doctrine is analogous to a cutpoint (or cutline or cutplane) strategy for dispositions in most game-theoretic settings. As a result, there is perfect agreement between dispositions and doctrine. As a practical matter, however, few if any models have exploited

the case space's opportunity to simultaneously study dispositions and rule-making. There has been little to no exploration of whether a judge would choose his own strategy as his articulated doctrine – whether his chosen doctrine would be identical to the equilibrium cutpoint he chooses. This is an important opportunity for future research.

### **OTHER OPPORTUNITIES FOR FUTURE RESEARCH WITHIN THE CASE SPACE**

We have an early understanding of how case facts affect doctrine, but our understanding is incomplete. We do not have a comprehensive understanding of how a court should select the best case facts (or 'vehicle') to achieve its most preferred policy. Exceptions in this area include Bustos and Jacobi's 'Judicial Choice among Cases for Certiorari', forthcoming in the *Supreme Court Economic Review*. (I return to case selection below.)

The case-space architecture allows scholars to focus on the judicial resolution of *individual disputes*, behavior that is uniquely judicial (as compared to legislative). But the case-space model is not entirely distinct from the policy model, or, as readers of analytical political science may think of it, the standard spatial model (such as Downs, 1957). The standard spatial model of policy-making under uncertainty is very powerful and explanatory. Consider the seminal model used by Crawford and Sobel (1982) and Gilligan and Krehbiel (1987, 1990). (The policy model has more often been used to describe legislative rather than judicial policy-making.)

The policy and case-space models assume both that policy-makers (whether, e.g., legislators or judges) have preferences about outcomes and that the policy-makers cannot directly control outcomes. Rather, the policy-makers control policy. Policy influences outcomes, but so do random shocks outside of the control of policy-makers (or

others). Scholars model these assumptions as  $x=p+w$ , where  $x$  is the outcome dependent on policy  $p$  and external factors  $w$ . A policy-maker who does not know the realization of  $w$ , and is asked to choose between two policies  $p_1$  and  $p_2$ , faces a quandary identical to that of a judge evaluating the appropriate outcome for a case whose facts he does not perfectly know (as is often assumed in models of the judicial hierarchy, such as in Cameron et al. (2000) and subsequent papers).

Despite this (and other) apparent similarities, the case-space and policy models differ fundamentally. The policy model focuses on the judicial (or legislative) choice of cutpoint. The case-space models focus on the judicial choice of an outcome *for an individual case*.

### **AMERICAN COURTS CANNOT UNILATERALLY SET THEIR OWN AGENDA**

American courts – including the United States Supreme Court – cannot set their own agenda unilaterally. Most trial and appellate courts hear (virtually) all disputes litigants bring to them. Courts with a discretionary docket (such as the United States Supreme Court) nonetheless are constrained to choose between (among) disputes that litigants bring to them. In other words, discretion over its docket allows the United States Supreme Court (for example) to decline to hear certain cases, but does not allow it to select from all possible disputes, or even all disputes actively being litigated. This restriction of cases is an important corollary of the judicial focus on individual disputes discussed in the previous section.

Despite courts' lack of complete autonomy, judicial power to set the court agenda has powerful effects on judicial behavior and, appropriately, has been the subject of serious scholarship. Courts can affect who brings disputes to them in a couple of ways: by allowing more parties or suits into courts (e.g., by

changing formal rules of access or by allowing new kinds of suits), by enticing new suits into court (e.g., by creating plaintiff-friendly doctrine), and by shifting which disputes are brought into court (e.g., by excluding certain groups or by introducing legal institutions to new kinds of disputes or new legal institutions altogether).

I discuss below a few considerations with implications for the development of doctrine.

In some instances, judges may want to *attract more* (attract fewer or discourage) cases. All judges – including trial judges – can (try to) affect this. On attracting more cases, a set of models, led largely by Dan Klerman, considers how judges attract cases at all. Consider the well-known practice of ‘forum shopping’: a plaintiff chooses the venue in which to bring his case. Klerman evaluates the opportunity this creates for ‘forum selling’ – the opportunity for judges or courts to attract plaintiffs to their courthouses. In Klerman’s argument, judges and courts are incentivized to create plaintiff-friendly doctrine to increase their caseloads. This was considered in Simpson (2017) with regard to pre-modern England, and in the contemporary context in Klerman and Reilly (2016), which discusses forum selling by the Eastern District of Texas in patent cases.

Of course, the obverse also can be true as a theoretical matter (and is true in some cases as an empirical matter): courts may want to *discourage* cases in order to reduce case load (or for some other motivation). (See for example Posner, 1993.) There is an opportunity for further scholarship in modeling the circumstances under which one or more judges (and at which levels in the judicial hierarchy) may want to increase or decrease either total caseload or type of case. There is also room for scholarship on how the strategies which judges employ to attract (discourage) cases interacts with the strategies used by litigants, particularly repeat litigants, to ensure their disputes are heard in their preferred fora (e.g., by shifting certain other cases to yet other courts, simply by currying

favor with the court by following judicial signals on caseload volume, by inviting judicial decision-making on caseload-affecting factors in a zeroth step of a game, etc.: see, e.g., Baird 2007, presenting an informal argument, and Paraweswaran, 2018, discussed elsewhere in this chapter).

### **SHAPING THE DOCKET: CASE SELECTION (DISTINCT FROM ATTRACTING (OR DISCOURAGING) CASE VOLUME)**

In addition to attracting (or discouraging) cases altogether, certain American courts enjoy perhaps a more famous power to shape their dockets: case selection. The ability to shape the docket is especially pertinent for the Supreme Court, which hears a small proportion of the cases brought to it: around 7,000–8,000 new cases are filed each year, of which the Supreme Court resolves fewer than 100 (see [www.supremecourt.gov](http://www.supremecourt.gov)). Studies of *certiorari* – the Supreme Court’s decision to grant review – are many.

As outlined by H. W. Perry in his seminal book *Deciding to Decide* (1991), justices’ decisions to grant or deny *certiorari* might depend on two kinds of possible considerations: either ‘extralegal’ considerations focusing on *outcome* preferences (as distinct both from the *policy* preferences discussed in an earlier section and from preferences about law qua law), or ‘legal’ considerations about which legal *issues* need attention, or both.

In line with these possible considerations, I discuss two strategic considerations that go into case selection and shaping the docket: which cases will produce *outcomes* that policy-oriented justices seek, and which cases will produce high-quality *law* that legal-oriented justices seek? (Of course, at least as a theoretical matter, any individual justice may be both policy- and legal-oriented, simultaneously or with respect to different parts of the (potential) docket.)

First, thinking about extralegal considerations, judges may each have a preferred policy, and may care about building a winning judicial coalition around their preferred policy. And some cases may be better vehicles than others from the perspective of bargaining – some may offer more leverage for achieving a desired political goal. This induces preferences over cases and bargaining over case selection. (On the relationship of this possibility to the case-space architecture discussed, see Parameswaran, Cameron and Kornhauser, 2019; Lax and Rader, 2015.) Second, thinking about preferences on law qua law, some cases may be better vehicles than others from the perspective of learning and foresighted doctrinal development.

Of particular interest is one way in which some cases may be better vehicles: they may be more informative. This way is of particular interest if judges can learn. By ‘judges can learn’, several related claims can be meant. First, the Supreme Court must learn from lower courts. This idea is particularly relevant to the US Supreme Court because American federal courts (like most court systems) are hierarchical. So the Supreme Court needs to *learn*, in essence, what is going on in the lower courts, both in terms of whether lower courts are applying Supreme Court instructions faithfully and in terms of the practical effects of law previously announced (by the courts or by Congress).

It is important to note here that much of the existing literature on the Supreme Court assumes things are known. This assumption permeates scholarship on multiple aspects of Supreme Court behavior. So, for example, most formal models of the judicial hierarchy are principal-agent theories that assume that the Supreme Court knows what it wants in any given case and simply struggles to achieve it due to large caseloads and insubordination. Similarly, in the context of interactions among justices (judges), the focus is on bargaining. But a judge (or other player, for that matter) can only bargain rationally once she knows her preferences, the preferences of

her colleagues and the expected outcomes of their collective decisions.

The assumption that ‘things are known’ ignores that courts are structured not only as a hierarchy but also as a ‘knowledge hierarchy’ (see, for example, Garicano, 2000). Knowledge hierarchies are organizations in which subordinates deal with easier questions and superiors with harder ones. Those harder questions can be promoted by the subordinate or by others. In courts, for example, a lower court judge may ‘certify’ a case to another court, or a litigant may appeal. On this account, as applied to federal courts, the job of the district court is to resolve large swathes of ‘easy cases’. The correct resolution of difficult questions is not the proper function of the district court (or its equivalent in any other functionally structured knowledge hierarchy).

Thinking about courts in this way changes likely possibilities about judicial preferences, skillsets, and so on. Conversely, failing to appreciate the knowledge-hierarchy construction of courts leads to misperceptions such as focusing on interactions between individual courts rather than seeing the Supreme Court as supervising a collective.

There is ample game-theoretical work exploring the downsides of such a structure (i.e., standard principal–agent costs). These models tend to assume the Supreme Court has in mind an ideal disposition for every case, and seeks to ensure that lower courts – agents – enact that preference. The seminal example is Cameron et al. (2000). Kastellec (2007) and Beim et al. (2014) introduce a multi-member lower court to estimate the consequences of dissent for monitoring. Lax (2003) introduces a multi-member Supreme Court to estimate the consequences of the Rule of Four and the benefits of extremity. Clark and Carrubba (2012) and Carrubba and Clark (2012) introduce written lower court opinions into the structure.

There remain two main ideas missing from the compliance literature. First, we do not yet have a firm grasp of how the Supreme Court

monitors lower courts as a collective unit, since most models consider one higher court and one lower court. But overseeing a group of courts is fundamentally different from overseeing just one.

Two papers consider how the Court's limited resources – i.e., costly review – affect how it learns from lower courts. Clark and Kastellec (2013) describe how the Court can learn by watching lower courts' decisions. Lower courts sequentially make independent decisions about which of two doctrines is best. Sometimes they will disagree, creating an intercircuit split. Clark and Kastellec describe how long the Court should allow this process to continue, in an optimal-stopping sense: the Court must trade an intolerance of legal discord in the lower courts against the desire to learn more by allowing more courts to weigh in. Beim (2017) considers how the Court can learn from decisions that are made simultaneously, and focuses on which cases are most informative – which vehicles are best for doctrinal development. The Court can observe some elements of a lower court's decision, such as what the case is about and what the outcome was, even without reviewing it. Other elements, such as the reasoning behind the outcome, may require further review to be fully understood. But some decisions are more ambiguous than others. In particular, an unbiased judge can make a moderate decision either because of strong evidence on both sides or because of weak evidence on both sides. As a result, these 'moderate' decisions are more likely to be informative upon review. A third take considers how lower courts can evade review by competing with one another to be the least non-compliant (Cameron, 1993).

Second, we have a weaker understanding of the benefits of this hierarchical structure. The benefits likely accrue primarily from the structure of the judiciary as a 'knowledge hierarchy' (cf, e.g., Garicano and Hubbard, 2007). In particular, higher courts can save time that can instead be used to focus on challenging issues, and can adopt good doctrine instead of writing it themselves.

There are *some* models on this, discussed elsewhere in this chapter. For example, see Clark and Kastellec (2013) and Beim (2017) on the benefits of learning from percolation through various lower courts (see also Lindquist and Klein (2006) and Beim and Rader (2019) for empirical studies of the same). Cameron and Kornhauser (2006) consider the informational benefits of litigant appeals – if litigants are savvy, they will only appeal cases that were incorrectly decided. Carrubba and Clark (2012) and Clark and Carrubba (2012) look at how lower courts can buy decisions that higher courts dislike by writing very high-quality doctrine which the higher court can adopt.

If case facts impact bargaining among justices, then justices will have induced preferences about which cases are most favorable to review.

Some models illustrate the importance of case selection, but do not go so far as to model said strategy. For example, Carrubba et al. (2012) present a model of Supreme Court decision-making in which the case facts divide the court into a majority and a minority. The Supreme Court's policy decision is then set at the median of the majority coalition. As a result, each justice has severe incentives to bargain over the best possible case. Although this is not discussed in that paper, it is an immediate by-product of the strategic environment studied in the paper.

The desire to decide cases correctly and build good law is, of course, also a central element of judicial utility. This is especially true in a common-law system, as the fundamental innovation of the common law is that courts answer hard questions that could not have been properly dealt with *ex ante*. Common-law courts are more efficient institutions for developing 'good' commercial law than are legislatures, because they deal with hard questions *ex post* (see Posner, 1973).

These preferences can come into play during case selection, as well. If case facts are differentially informative, or have differential downstream consequences, then judges may

have preferences about which case is most worth their time. I discuss three examples.

First, Beim et al. (2017) consider cases that cause follow-up questions to arise, and argue that a court may delay deciding an easy case if it will raise difficult follow-up questions. Second, Baker and Mezzetti (2012) develop a model in which doctrine is built incrementally. Judges evaluate whether, in expectation, a case will be sufficiently informative for legal development to warrant the time necessary to resolve it. In their model, judges ‘gap fill’ as time progresses, issuing opinions based on the information they learn in the case they are deciding at the time. Judges need not adhere perfectly to judicial minimalism – they can issue dictatory opinions if they choose – and so there is sometimes legal inconsistency over time (see also Badawi and Baker, 2015). Relatedly, Parameswaran (2018) presents a model in which a legal rule is uncovered over time. But in that paper, judicial decisions can affect the flow of future cases.

### **JUDICIAL DECISION (THE JUDGE DECIDES OR THE JUDGES DECIDE THE CASE)**

Early work in judicial politics focused on the judge’s decision. How does a judge decide a case? Three informal models were developed in political science as summaries for how judges might behave.

The first of these, the ‘legal model’, supposed that judges followed the law and did what it said.<sup>1</sup> This is seemingly straightforward and obvious, but legal realists – the theoreticians who are the intellectual forefathers of the contemporary game-theoretic analysis of judicial decision-making – often pointed out that deciding a case correctly is rarely straightforward. Karl Llewellyn described this evocatively in his 1934 lecture ‘The Constitution as an Institution’: ‘Man (even though he learned square roots in high school) finds more than one right answer hard to conceive of.’

The second, the ‘attitudinal model’, developed by Segal and Spaeth (1993, 2002), focused on Supreme Court justices and argued that they governed their decisions by their personal preferences. That argument posits that when Supreme Court justices are making decisions on the merits, their decisions are governed solely by their wish for whom will win. ‘Simply put, Renhquist votes the way he does because he is extremely conservative; Marshall voted the way he did because he was extremely liberal’ (Segal and Spaeth, 2002: 86).

A third model, the ‘strategic model’, embraced the strategic elements of rational-choice analysis and claimed that judges and justices moderate their decisions in anticipation of other actors. (Ferejohn and Shipan, 1990 is a seminal example of work using this model.) The ‘strategic model’ is the closest to a game-theoretic tradition, since it stems from an interest in strategy and anticipating the actions of others. Because this is the oldest branch of judicial politics, there are excellent review essays discussing formal models of how judges decide cases. See Epstein and Jacobi (2010) and Cameron and Kornhauser (2017a, 2017b).

### ***Judicial Decision: the Opinion***

The judicial opinion is the closest thing to a legislative bill that the judiciary produces. As a result, there is ample game-theoretic material on the content of judicial opinions, especially on the Supreme Court. There are many legislative-like models about bargaining over opinion content. See, among others, Hammond et al. (2005), Carrubba et al. (2012) and Lax and Cameron 2007.

There are also interesting papers on author selection and opinion assignment on the Court of Appeals, where the opportunities of strategy are yet more complex. Two excellent examples include Farhang et al. (2015), which considers the interplay of gender and ideology in determining strategic author

selection, and Hazelton et al. (2016), which considers the interplay of publication and dissent.

Once a decision is issued, often accompanied by a precedential written opinion, the game begins again. Actors in the jurisdiction governed by the relevant court change their behavior in anticipation of the court's future actions. But no written opinion is perfectly complete, and so new ideas and new problems arise. Disputes follow, and some of those disputes land in court.

## Note

- 1 The 'legal model' is a term coined primarily by proponents of other models; there is no singular document that presents the argument. See George and Epstein (1992) for a good summary of the arguments.

## REFERENCES

- Badawi, Adam B. and Scott Baker. Appellate Lawmaking in a Judicial Hierarchy. *The Journal of Law and Economics*, 2015, 58(1): 139–172.
- Baird, Vanessa A. 2007. *Answering the Call of the Court: How Justices and Litigants Set the Supreme Court Agenda*. University of Virginia Press.
- Baker, Scott and Claudio Mezzetti. A Theory of Rational Jurisprudence. *Journal of Political Economy*, 2012, 120: 513–551.
- Beim, Deborah. Learning in the Judicial Hierarchy. *The Journal of Politics*, 2017, 79(2): 591–604.
- Beim, Deborah and Kelly Rader. Legal Uniformity in American Courts. *Journal of Empirical Legal Studies*, 2019, doi:10.1111/jels.12224
- Beim, Deborah, Tom S. Clark and John W. Patty. Why Do Courts Delay? *Journal of Law and Courts*, 2017, 5(2): 199–241.
- Beim, Deborah, Alex Hirsch and John Kastellec. Whistleblowing and Compliance in the Judicial Hierarchy. *American Journal of Political Science*, 2014, 58(4): 904–918.
- Bueno de Mesquita, Ethan and Matthew Stephenson. Informative Precedent and Intrajudicial Communication. *American Political Science Review*, 2002, 96(4): 755–766.
- Bustos, Álvaro and Tonja Jacobi. Judicial Choice among Cases for Certiorari. *Supreme Court Economic Review*, forthcoming.
- Callander, Steven and Tom S. Clark. Precedent and Doctrine in a Complicated World. *American Political Science Review*, 2017, 111(1): 184–203.
- Cameron, Charles M. New Avenues for Modeling Judicial Politics. Paper prepared for delivery at the Wallis Institute 2nd Annual Conference on *The Political Economy of Public Law*, University of Rochester, NY, October 15–16. Mimeo, 1993.
- Cameron, Charles M., and Lewis A. Kornhauser. Appeals mechanisms, litigant selection, and the structure of judicial hierarchies, in Jon Bond, Roy Flemming, and James Rogers (eds), *Institutional Games and the U.S. Supreme Court*. University of Virginia Press, 2006.
- Cameron, Charles M., and Lewis A. Kornhauser. What do courts do? How to model judicial actions. *Model Courts: Positive Political Theory and Judicial Institutions*, Chapter 2 (working paper), 2017.
- Cameron, Charles M. and Lewis A. Kornhauser. What do judges want? How to model judicial preferences, *Model Courts: Positive Political Theory and Judicial Institutions*, Chapter 3 (working paper), 2017.
- Cameron, Charles M., Jeffrey A. Segal and Donald Songer. Strategic Auditing in a Political Hierarchy: An Informational Model of the Supreme Court's Certiorari Decisions. *American Political Science Review*, 2000, 94(1): 101–116.
- Carrubba, Clifford J. and Tom S. Clark. Rule Creation in a Political Hierarchy. *American Political Science Review*, 2012, 106(3): 622–643.
- Carrubba, Cliff, Barry Friedman, Andrew A. Martin and Georg Vanberg. Who Controls the Content of Supreme Court Opinions? *American Journal of Political Science*, 2012, 56(2): 400–412.
- Clark, Tom S. and Clifford J. Carrubba. A Theory of Opinion Writing in a Judicial Hierarchy. *Journal of Politics*, 2012, 74(2): 584–603.
- Clark, Tom S. and Jonathan P. Kastellec. The Supreme Court and Percolation in the Lower

- Courts: An Optimal Stopping Model. *The Journal of Politics*, 2013, 75(1): 150–168.
- Crawford, Vincent P., and Joel Sobel. Strategic Information Transmission. *Econometrica*, 1982, 50(6): 1431–1451.
- Dahl, Robert A. Decision-making in a Democracy: The Supreme Court as a National Policy-maker. *Journal of Public Law*, 1957, 6: 279–295.
- Dewatripont, Mathias and Jean Tirole. Advocates. *Journal of Political Economy*, 1999, 107(1): 1–39.
- Dickens, Charles. 1853. *Bleak House*. New York: Hurd and Houghton.
- Downs, Anthony. An Economic Theory of Political Action in a Democracy. *Journal of Political Economy*, 1957, 65(2): 135–150.
- Ellickson, Robert C. 1994. *Order Without Law: How Neighbors Settle Disputes*. Harvard University Press.
- Epstein, Lee and Tonja Jacobi. The Strategic Analysis of Judicial Decisions. *Annual Review of Law and Social Science*, 2010, 6: 341–358.
- Farhang, Sean, Jonathan P. Kastellec and Gregory J. Wawro. The Politics of Opinion Assignment and Authorship on the US Court of Appeals: Evidence from Sexual Harassment Cases. *The Journal of Legal Studies*, 2015, 44(S1): S59–S85.
- Ferejohn, John and Chuck Shipan. Congressional Influence on Bureaucracy. *Journal of Law, Economics, and Organization*, 1990, 6: 1–21.
- Fox, Justin and Georg Vanberg. Narrow versus Broad Judicial Decisions. *Journal of Theoretical Politics*, 2014, 26(3): 355–383.
- Garicano, Luis. Hierarchies and the Organization of Knowledge in Production. *Journal of Political Economy*, 2000, 108(5): 874–904.
- Garicano, Luis and Thomas N. Hubbard. *The Return to Knowledge Hierarchies*. No. w12815. National Bureau of Economic Research, 2007.
- Gennaioli, Nicola, and Andrei Shleifer. The Evolution of Common Law. *Journal of Political Economy*, 2007, 115(1): 43–68.
- George, Tracey E. and Lee Epstein. On the Nature of Supreme Court Decision Making. *American Political Science Review*, 1992, 86(2): 323–337.
- Gilboa, Itzhak and David Schmeidler. Case-based Decision Theory. *The Quarterly Journal of Economics*, 1995, 110(3): 605–639.
- Gilligan, Thomas W. and Keith Krehbiel. Collective Decisionmaking and Standing Committees: An Informational Rationale for Restrictive Amendment Procedures. *The Journal of Law, Economics, and Organization*, 1987, 3(2), 287–335.
- Gilligan, Thomas W. and Keith Krehbiel. Organization of Informative Committees by a Rational Legislature. *American Journal of Political Science*, 1990, 34(2): 531–564.
- Hammond, Thomas H., Chris W. Bonneau and Reginald S. Sheehan. 2005. *Strategic Behavior and Policy Choice on the U.S. Supreme Court*. Stanford University Press.
- Hazelton, Morgan, Rachael K. Hinkle, See Seon Jeon. Sound the Alarm? Judicial Decisions Regarding Publication and Dissent. *American Politics Research*, 2016, 44(4): 649–681.
- Kastellec, John. Panel Composition and Judicial Compliance on the U.S. Courts of Appeals. *Journal of Law, Economics, & Organization*, 2007, 23(2): 421–441.
- Kastellec, Jonathan P. 2017. The Judicial Hierarchy: A Review Essay. Oxford Research Encyclopedia, Politics [online]. doi: 10.1093/acrefore/9780190228637.013.99.
- Klerman, Dan and Greg Reilly. Forum Selling. *Southern California Law Review*, 2016, 89(2): 241–316.
- Kornhauser, Lewis A. Modeling Collegial Courts. II. Legal Doctrine. *Journal of Law, Economics, & Organization*, 8(2): 441–470.
- Lax, Jeffrey R. Certiorari and Compliance in the Judicial Hierarchy: Discretion, Reputation and the Rule of Four. *Journal of Theoretical Politics*, 2003, 15(1): 61–86.
- Lax, Jeffrey R. The New Judicial Politics of Legal Doctrine. *Annual Review of Political Science*, 2011, 14(June): 131–157.
- Lax, Jeffrey R. and Charles M. Cameron. Bargaining and Opinion Assignment on the US Supreme Court. *The Journal of Law, Economics, and Organization*, 2007, 23(2): 276–302.
- Lax, Jeffrey R. and Kelly Rader. Bargaining Power in the Supreme Court: Evidence from Opinion Assignment and Vote Switching. *The Journal of Politics*, 2015, 77(3): 648–663.
- Lindquist, Stefanie A. and David E. Klein. The Influence of Jurisprudential Considerations



- on Supreme Court Decision Making: A Study of Conflict Cases. *Law and Society Review*, 2006, 40(1): 135–162.
- Llewellyn, Karl N. The Constitution as an Institution. *Oregon Law Review*, 1934, 14: 108.
- Milgrom, P. R., D. C. North and B. R. Weingast. The Role of Institutions in the Revival of Trade: The Law Merchant, Private Judges, and the Champagne Fairs. *Economics and Politics*, 1990, 2(1): 1–23.
- Parameswaran, Giri. Endogenous Cases and the Evolution of the Common Law. *The RAND Journal of Economics*, 2018, 49(4): 791–818.
- Parameswaran, Giri, Charles M. Cameron and Lewis A. Kornhauser. *Bargaining and Strategic Voting on Appellate Courts* (Working Paper), 2019.
- Perry, H. W. Jr. 1991. *Deciding to Decide*. Harvard University Press.
- Posner, Richard A. *Economic Analysis of Law*, Little Brown and Company, 1973.
- Priest, George L. and Benjamin Klein. The Selection of Disputes for Litigation. *The Journal of Legal Studies*, 1984, 13(1): 1–55.
- Posner, Richard A. What Do Judges and Justices Maximize? (The Same Thing Everybody Else Does). *Supreme Court Economic Review*, 1993, 3: 1–41.
- Segal, Jeffrey A. and Harold J. Spaeth. 1993. *The Supreme Court and the Attitudinal Model*. Cambridge University Press.
- Segal, Jeffrey A. and Harold J. Spaeth. 2002. *The Supreme Court and the Attitudinal Model Revisited*. Cambridge University Press.
- Simpson, Hannah. *Access to Justice in Revenue-Seeking Legal Institutions* (working paper), 2017. Available at <https://static1.squarespace.com/static/5668949cdc5cb47474730d69/t/59b3e2aecd0f685d6d2f0b9f/1504961200546/Access+to+Justice+August.pdf> (Accessed on 4 January, 2020).



# Wrestling with Complexity in Computational Social Science: Theory, Estimation and Representation

Scott de Marchi and Brandon M. Stewart

## INTRODUCTION

Computation is increasingly a feature of most social scientists' research. Unsurprisingly, given the breadth of social science research, the term 'computational social science' has come to mean different things to different researchers and the challenges posed by this increased reliance on computation are not well understood. In this chapter, we hope to show that computation is useful for both theoretical and empirical modeling. In both areas, computational modeling allows researchers to build and estimate more complex models than was previously possible. Instead of attempting to review what has become a vast field, the main goals of this article will be to provide answers to the following questions:

- What types of research problem have led to an increased reliance on computation?
- What hazards exist in relying on computational models?

- Has computation substantially improved the reach of our models, either theoretical or empirical?

In considering these questions, we will focus on three main topics. The first is the use of computational social science to extend the reach of *theoretical models*. Models in this tradition are often referred to as 'agent based models' or 'complex adaptive systems' and several good overviews exist of this area of research (Miller and Page, 2007; de Marchi and Page, 2008, 2014).

The second is the use of computational social science to engage in the *estimation* of a variety of statistical models, especially with the use of machine learning. As noted by Alvarez (2016: 1–2):

Because students are accustomed to having such powerful tools as probit and logit, they often do not understand just how far applied social scientific research has advanced in the past few decades... Few students today realize that back when Rosenstone and Wolfinger were estimating the

models for their paper, when probit models would take 50 minutes to run, they were likely using a mainframe computer. Accessing computer time usually required payment in the form of university research funds. These computational barriers seriously limited social science research.

This anecdote highlights the most common use of computation in applied social science – as raw horsepower to carry out tasks such as the optimization required to estimate statistical models. This has had transformative effects: one notable example is the increased capabilities of neural networks in both forecasting and the estimation of latent variables.

The last area of application that has benefited from increased computational power is *representation* and unsupervised learning, where the main payoff to the discipline has been new sources of data and approaches to measurement. Text as data serves as one central area in which computation has transformed our ability to use large bodies of text to create variables for use in applied statistical models.

In all of the above cases, the rise of computation has been a response to both increased supply of computational power and increased demand for a way to deal with complexity (broadly conceived). Almost always, we are relying on computation because our theories, statistical models or data require its use – without computation, it would be impossible to use agent based models (Kollman et al., 1992; Laver, 2005), event data (Yang et al., 2016; Nguyen and Grishman, 2018), Twitter data (Barberá et al., 2015), text as data (Roberts et al., 2013; Grimmer and Stewart, 2013), network-based models (Fowler, 2006; Dorff et al., 2018; Minhas et al., 2019; Rozenzas et al., 2019), or employ complex models for forecasting (de Marchi et al., 2004). That said, each of these applications has the same underlying cost: computational social science involves choices and the parameter spaces involved are often large.<sup>1</sup> Computational models focused on theory or applied statistics are much more difficult to understand than prior research.

In what follows, we will explore these issues for computational models that focus on theory as well as estimation and measurement. Our hope is to highlight the progress that has been made on all of these fronts, but also to highlight some of the general challenges involved in doing this kind of work. This overview will not be expansive – there is simply too much work that falls under the umbrella of ‘computational social science’. Rather, we will use highly salient examples to illustrate our general points about the use of computation for both theory and empirics.

## THEORY

The field covered in this book [*Theory of Games and Economic Behavior*] is very limited, and we approach it in this sense of modesty ... Its first applications are necessarily to elementary problems where the result has never been in doubt and no theory is actually required. At this early stage the application serves to corroborate the theory. The next stage develops when the theory is applied to somewhat more complicated situations in which it may already lead to a certain extent beyond the obvious and the familiar. Here theory and application corroborate each other mutually. Beyond this lies the field of real success: genuine prediction by theory. It is well known that all mathematized sciences have gone through these successive phases of evolution. (von Neumann and Morgenstern, 1944: 8)

The most common computational models that generate theory are named agent based models (ABM) and consist of a set of actors, called agents, represented in a computer language by behavioral algorithms (Holland and Miller, 1991). Compared to game theoretic models, agents in ABMs have a range of different behavioral rules rather than relying on the more common assumption of backwards induction on an extensive form game. ABMs can, of course, rely on backwards induction and agents may act as perfectly rational actors, but this is not always the best choice for a given context if one wants to understand human behavior. In fact, exploring different behaviors is one of the focal points of

computational models in this tradition; in most models, agents optimize with plausible limitations on memory and cognition (Kim et al., 2010). Given that agents in ABMs are not constrained to use backwards induction as an algorithm, the focus is not typically on finding equilibria. Rather, in computational models the goal is to build more complex models that have enough verisimilitude to real-world contexts to generate precise empirical results. The real distinction between purely deductive and computational approaches is whether one produces a simple model that permits an analytical solution or a more complex model that must be exercised computationally (that is, as a Monte Carlo).

The diversity of behavioral algorithms in computational models has been a source of some controversy due to many game theorists' belief that formal models should rely on rational actor assumptions and not model more complex behavior (Diermeier and Krehbiel, 2003). This ignores, however, the large role that behavior has in all models, even those in the non-cooperative game theory tradition. Equilibrium refinements such as sequential equilibrium (Kreps and Wilson, 1982), trembling hand equilibrium, subgame perfection, and so on, are themselves behavioral assumptions, as are crucial assumptions like stationarity that are not a necessary part of a best response function. Moreover, the proof is in the pudding. The original goal of game theory was to use human games such as poker and chess as a window to more complex behaviors in legislatures and bargaining between nations.<sup>2</sup> The fact, though, is that the proliferation of relatively simple games is due to a general lack of progress by game theorists to expand the complexity of games that may be examined. Poker, one of the longest studied problems by game theorists (see for example Nash and Shapley, 1950; Binmore, 1992; Rapoport et al., 1997), produced virtually no progress (that is, empirical results) until modern computational methods were used to study the problem (Billings et al., 1998; Gilpin and Sandholm, 2007; Sandholm, 2010; Brown and

Sandholm, 2018). Currently, computational models of poker can beat expert humans in two player games of Texas Hold 'em, which is an astonishing level of success compared to earlier, purely deductive approaches.

The reason why computational models have succeeded in some areas where purely deductive models have not is because the price for pure deduction is almost always a focus on simple games. One cannot study poker or other games which humans actually play within the confines of deductive models, so instead one studies radically simpler games that bear only a passing resemblance to real-world contexts. One thus has a choice as a researcher: choose to privilege deductive models, rational actors, and focus on simple games *or* instead focus on more complex games with verisimilitude to human choice contexts and explicitly model behavior with algorithms that stray from backwards induction.<sup>3</sup> Ultimately, insights from deductive approaches can be extended with computation to produce more fruitful models – computation and deduction can and should be complementary.<sup>4</sup>

The other key difference between computational social science models and game theoretic models is that computational models are concerned with behavior in like regions of the model parameter space. In game theoretic models, agents optimize perfectly and most often researchers settle on a set of assumptions designed to produce plausible equilibria. Game theorists do not often deal with dynamics, nor do they expose their models to robustness checks (that is, testing whether or not results survive perturbations to the model parameters).<sup>5</sup> Put another way, one should ask if an equilibrium result meets two criteria: small perturbations to non-essential parameters should not produce discontinuous changes to the results; human actors must be able to find equilibria given reasonable cognitive endowments. Judd and Page (2004) note that:

When we prove a theorem characterizing the set of equilibria in a model or class of models, we like to think that we have explained what will happen in that model – the likely end result. However,

there are important differences between (1) proving the existence of an equilibrium, (2) being able to solve for it, and (3) constructing a set of agents who can learn the equilibrium (p. 197).

Finding an equilibrium for any particular game is valuable, but it is not the end of the road for research. The strength of game theory is that assumptions matter for producing results; the weakness is that there is not a lot of intuition in most cases about what happens to these results if one changes the parameters of the model or if agents do not conform to the behavioral assumptions of the model (de Marchi, 2005).

In a nutshell, computational models aim to accomplish the following:

- i. relax assumptions in existing formal models to generate more general results;
- ii. include algorithms for more plausible behavior (compared to assuming backwards induction), which often means situating actors in time/space/on a network;<sup>6</sup>
- iii. examine relationships between parameters and target variables using applied statistics. This develops an equivalence class for similar models or alerts the researcher that no such class exists;<sup>7</sup>
- iv. develop a correspondence to a real-world outcome and conduct an empirical test of the model. The goal of science is not to write down a deductive model in most cases – there are many such models. The goal is to discover the small subset of deductively true models that have a mapping to real-world outcomes.<sup>8</sup>

To provide additional context for this outline, we will examine two areas where computational models are prevalent: bargaining theory and elections.

### ***Bargaining Theory and Coalition Formation***

Bargaining theory is one of the more important areas of research in economics and political science, as it concerns the allocation of a resource between strategic parties. Typical contexts include bargaining between an employer and a union, nations considering

alliances and parties forming coalitions in proportional representation systems. The most common family of non-cooperative bargaining models starts with Rubinstein (1982) and Baron and Ferejohn (1989). To be clear, these models were crucial in suggesting ways in which the complexity of bargaining might be limited to produce expectations about how human decision-making operates in these contexts. But it is worth noting that many of the key assumptions in these models – especially when generalized to n-players bargaining in coalition politics or legislatures – represent very strong (and largely unmotivated) behavioral assumptions. As Laver et al. (2011) point out, models in the Rubinstein and Baron–Ferejohn tradition have results that are driven by the following assumptions:

- i. *A particular understanding of power (for example, minimum integer weights).* For example, parties in a legislature may have raw seat totals of (10, 10, 1); the minimum integer weights representation<sup>9</sup> of this same legislature (that is, how pivotal each party is) suggests the power of these parties is (1, 1, 1). Focusing on raw seats versus pivotality has a dramatic effect on the expected payoffs (continuation values) of the parties.
- ii. *A strict sequence of play where an exogenous, uniform selection mechanism selects a single party that is able to propose a coalition at each point in time.* Changing this to an endogenous part of the game would have huge implications for the results, as would allowing for an open-outcry approach where multiple actors would be able to make offers simultaneously.<sup>10</sup>
- iii. *A reversion point of 0 for all players if negotiations fail.* This, combined with assumption ii, is largely responsible for the proposer advantage result in many models. Changing the reversion point to a status quo position (for example, a caretaker government) or a new election entirely would remove much of this advantage.
- iv. *Stationarity.* This ensures that history does not matter and that strategies remain constant given ambient conditions at each stage. Without this, the set of equilibria would in most cases be very large.

Worse, attempts to generalize these games to deal with  $n > 2$  players and the non-homogeneous

case<sup>11</sup> have led to a great deal of mayhem (Laver et al., 2010). We are not claiming that formal theorists have not produced deductive models altering or generalizing the above assumptions – multiple innovations exist in the literature (Morelli, 1999; Banks and Duggan, 2006; and Martin and Vanberg, 2019 are excellent examples). But, in the framework of purely deductive models, this process is slow and it is difficult to build intuition about how changes to any of the above parameters affects the general results of bargaining models. Put another way, the combinatorics of the above parameters are bad for modeling – one would need to write hundreds of models to cover the set of parameters listed here.

The bottom line is this: choosing assumptions to limit the complexity of the game is suboptimal if one wishes to study human bargaining. For anyone who has ever been in a faculty meeting, it is reasonably obvious that open-outcry models (to focus on assumption ii above) are a more interesting theoretical approach to understand real-world bargaining. We have equilibria in Baron–Ferejohn style games, though at a price in terms of the complexity of the games we can study. This is not to say that one should do without deductive models – there are important insights to be gained from formalizing a model – but rather that if we are interested in building theories that will map to actual human behavior, we should seek to expand the reach of these models using computation where appropriate.

In the area of coalition bargaining in proportional representation systems, de Marchi and Laver (2020) and de Marchi et al. (n.d.) provide examples of how one may use computation to generalize game theoretic bargaining models. In the first paper, logrolling is added as a way to include policy along with perquisites in the bargaining process; by doing so, Condorcet winners are generated. This purely synthetic variable (that is, the likelihood of observing a Condorcet winner in the computational model) predicts the duration of coalition bargaining in proportional

representation systems. Similarly, the second paper relaxes the assumption of a zero reversion point for all players in the case that bargaining fails and replaces it with the possibility of new elections. A new election, in addition to being the constitutionally mandated reversion point in most systems, can affect bargaining in the present – players that derive a greater advantage from holding new elections may receive higher payoffs in observed bargaining outcomes.

### ***Electoral Models***

Bargaining theory is not the only area in which computational social science has had a positive effect on generating theory. Perhaps the most extensive tradition of computational modeling concerns how preferences are aggregated by institutions into outcomes. Path-breaking work by Kollman et al. (1992, 1998) focused attention on the fact that real-world contexts feature high-dimensional policy spaces and candidates that use optimizing algorithms rather than backwards induction. Constituents are also not fixed; they may move geographically, which has implications for outcomes at the district level (Kollman et al., 1997).

While these early papers were purely theoretical models (albeit computational), empirical support was found for both the American case and the Irish case. Ensley et al. (2009) found that measures of the complexity of electoral landscapes determined the success enjoyed by incumbents. Due to experience and the fact that they had won at least one prior election (Alchian, 1950), complex electorates favor incumbents over challengers, as one would expect from Kollman et al.'s work. Laver (2005) expanded on Kollman et al.'s models and explored the interplay between the algorithms employed by parties to search for platforms and the responses by voters in elections. Relatively simple behavioral rules in Laver's agent based models recover time

series variation in public opinion polls as well as empirically observed party sizes.

Too often, the response of formal theorists to computational models is: ‘that could have been proven deductively’. This claim ignores the hard fact that formal models most often make simplifying assumptions to produce sharp results and this limits the mapping of models to empirical tests of real-world behavior. And too often, computational models ignore prior work by formal theorists. Formal models should be used to help define the actors, strategy sets and payoffs in any given context and serve as necessary starting points for building more complex models.

### *Problem 1: Multiplicity of parameter choices*

Bargaining and elections are not the only examples of computational models that have extended the reach of theory. They do, however, illustrate one of the central problems with the use of computation in theory building: many parameters underlie these models. While there are obvious problems with simply selecting one vector of parameters and generating a simple model solved deductively, building a more complex model comes at the expense of larger parameter spaces which one can only investigate with applied statistics.<sup>12</sup> Often, these parameter choices are poorly detailed in the text of papers, and few researchers in the social sciences have the time or training to read the code that forms the basis of these models.

Ultimately, replicating code at the level of reproducing tables in papers is quite easy. Replicating a computational model at the level of validating that the code does what the author says it does and coming to an understanding of the role of the salient parameters in the model is quite difficult. The main advice we offer is that researchers should produce models that are as complex as they need to be. As with machine learning models (detailed in the next section), code needs to be made available but, more important, modelers need to investigate the relationship

between the parameters in their model and outcome variables and present these relationships clearly. Crucially, instead of treating parameters in an *ad hoc* fashion, theory and the available empirical evidence should guide these choices in models. For example, whenever possible, empirical probability density functions should be used instead of atheoretic distributions being chosen for convenience.

The lesson for us is that whether one is doing non-cooperative game theory or computational models, there are risks that one is overfitting one’s model to produce a desired result.<sup>13</sup> This is, in fact, no different from the curve-fitting that exists in applied statistical models. And, for us, the answer is the same: use theory to pick parameters and expose one’s theoretical model to frequent tests against data.

### *Problem 2: Opacity of the model*

A related issue is that more complex models come at the cost of interpretability. The folk theorem aside, when one examines a simple model like the iterated prisoner’s dilemma (IPD), the highly constrained strategy space and ease of finding equilibria means that we can understand the ‘moving parts’ of the model. With computational models involving much more complex strategy spaces and behavioral algorithms, it is not always easy to grasp what is driving the results.

For example, while computational models of poker have had a great deal of success in two-player games against humans, they have had a much more difficult time generalizing to n-players. Worse, the strategies employed by these poker models are themselves often opaque.<sup>14</sup> There is – as with machine learning models, detailed in the next section – a tradeoff between predictive power and verisimilitude on the one hand and interpretability on the other. Which point on this spectrum a model sits at depends on the application (for example, a simpler model may be chosen due to a desire for policy relevance), but this remains a genuine concern and subject for future research.

To summarize, the central point is this: we understand that many readers are likely suspicious of computational models used to generate theory. These models are not purely deductive, the code is difficult to validate without using statistical models relating parameters to outcomes and computer science is not a skill that is emphasized in our discipline. All of these concerns are valid. But there are tradeoffs that make the use of computation essential, as the examples of poker, bargaining theory and elections demonstrate. Game theoretic models that study simplified games without any mapping to data (and without providing dispositive statistical tests of the implications of the models) mean we are not all that close to the human behavior that motivates our research. While we may find an equilibrium that confirms a stylized fact about politics, this is not evidence any more than a purely correlative statistical model based on observational data is causal. Theorists creating deductive models are in many cases forced, in the concluding remarks of papers, to make verbal analogies for why the IPD has a lot to say about the Vietnam war or why constituents are really just playing signaling games when choosing between candidates. All of the mathematical precision that characterizes these models is thus thrown out of the window when we are forced to bridge the gap between where the games end and the behaviors of interest begin. Computation is the glue that can narrow that gap, and in our view this narrowing is essential if theory is going to play a primary role in guiding research in the social sciences in the future.

## **MACHINE LEARNING AND ESTIMATION**

Just as with theory, increased computational power has given rise to machine learning as an alternative to and extension of traditional statistical models. While machine learning

defies succinct description, it may be easiest to think about as the science of predicting an output given a set of inputs. While models common to social science, such as Ordinary Least Squares (OLS), assume a simple additive, linear structure, in its simplest form the introduction of machine learning offers the ability to move beyond additivity and linearity to model how different features interact in complex and non-linear ways.<sup>15</sup> The introduction of these methods parallels the benefits of using computation for theory by allowing us to move beyond the simplifying assumptions of traditional estimation strategies. This leads to an increased focus on estimation in causal inference and demands a broader discussion of the role of prediction in testing and evaluating theories.

In this section we examine the role of new computationally intensive forms of estimation in causal inference and prediction. While machine learning may allow us to tackle more complicated estimation problems, the core of social science research – how do we learn about social phenomena from data – has largely remained unchanged. We discuss the strengths and weaknesses of machine learning in causal inference and prediction and offer a perspective on what has and has not changed in these fields.

### ***Machine Learning***

Much of machine learning is about estimating a function,  $f()$  which maps from an input,  $X$ , to a prediction,  $\hat{Y}$ , which approximates a target variable,  $Y$ . Many quantitative social scientists are most comfortable with this idea in terms of OLS, which predicts the outcome using a linear function of the inputs:  $\hat{Y} = f(X) = X\beta$ . The fitted function is chosen based on how well  $\hat{Y}$  approximates  $Y$  based on the squared error loss (that is, the one that minimizes the sum of squared errors). While OLS has many virtues, it, as well as other common methods in the social science toolbox, is often not the prediction method with the highest accuracy



due to constraints on functional form, overfitting when using many variables and similar concerns.<sup>16</sup> Machine learning provides alternative methods to learn functions that may be more predictive in settings with many observations. This has allowed for estimation with higher-dimensional feature spaces (more variables) and more complex functional forms (interactions and non-linearities among many variables which do not have to be pre-specified). These gains can be substantial. In Kleinberg et al. (2018), a machine learning approach using gradient boosted decision trees substantially outperformed a basic logistic regression in predicting risk of releasing a defendant on bail. Defendants predicted by the machine learning algorithm to be in the top 1% of the risk distribution committed crimes at a rate of 54%, compared to 40% for those similarly flagged by the logistic regression.<sup>17</sup> Like agent based modeling, the value of machine learning for social scientists is the ability to cope with increased complexity in estimation beyond the standard tools in the social scientist's toolkit.

Flexibility in estimation comes with two related problems: consequential tuning parameters which must be chosen by the user, and opacity of predictions. By contrast, for a given set of input variables, output variables and observations, the OLS estimator only has one solution.<sup>18</sup> The predictions made by OLS might not be very good but they are easy to understand (or at least seem easy to understand!): as we perturb the input variables, the predictions change in a predictable (linear) way. The flexibility of many machine learning methods means that small perturbations in the inputs can map to large changes in the predictions due to local effects, making it difficult to understand why two similar observations yield vastly different predicted outcomes. This flexibility can be a great thing – sometimes the world really is complex – but it also raises different challenges than the ones that are often covered in social science statistics courses.

## Causal Inference

Machine learning tools can also aid social scientists in the goals they are already pursuing, such as causal inference. A new wave of interest in causal inference over the past decade has helped bring attention to the challenges of making inferences about counterfactual situations (Angrist and Pischke, 2008; Morgan and Winship, 2014; Imbens and Rubin, 2015; Pearl and Mackenzie, 2018). Causal inference separates identification of the causal effect from estimation. Identification is the process of linking the estimand of interest (generally involving an unobserved counterfactual quantity) and a statistical estimand that describes the observed data. Estimation then involves estimating the statistical estimand from a finite sample of data. While the two steps are often conflated by practitioners, separating them is essential for deploying machine learning in causal inference. While social scientists must use their understanding of the world to focus on choosing an estimand and identifying it, machine learning helps with the estimation step.

Consider, for example, selection on observables, the identification strategy that powers most of observational social science. Suppose our interest is in the average treatment effect of some treatment  $D$ :  $\tau = E[Y_i(1) - Y_i(0)]$  (written using potential outcomes notation: see Imbens and Rubin, 2015). The treatment effect for an individual is  $Y_i(1) - Y_i(0)$ , the difference for the unit between its outcome if it had taken treatment ( $Y_i(1)$ ) and the unit's outcome if it had not ( $Y_i(0)$ ). The average treatment effect is simply the expectation of these individual effects over the units. Of course, we can never observe both of these quantities and so we appeal to two identification assumptions: (1) positivity ( $0 < p(D_i = 1|\mathbf{X}, Y(1), Y(0)) < 1$ ), which says that every unit has the chance of being treated (or not), and (2) no unmeasured confounding  $p(D_i = 1|\mathbf{X}, Y(1), Y(0)) = p(D_i = 1|\mathbf{X})$ , which says that within levels defined by a set of covariates

$\mathbf{X}$ , the assignment of the treatment  $D$  is not associated with the potential outcomes. These two assumptions allow us to rewrite our statistical estimand  $\tau$  in terms of conditional expectations over observed data  $E_X [E[Y_i|D_i = 1, \mathbf{X}_i] - E[Y_i|D_i = 0, \mathbf{X}_i]]$ . The subsequent estimation problem is fitting a function to approximate these conditional expectations in our observed data. Identification then tells us what that quantity means for us as social scientists.

Separating identification from estimation has allowed machine learning to act as a substitute for more traditional models when tackling difficult estimation problems. In theory, the conditional expectations that we use to reason about identification are arbitrarily non-linear and interactive. Historically, we may have approximated those conditional expectations using linear, additive OLS (and perhaps more recently matching or weighting: see for example Ho et al., 2007 and Hainmueller, 2012). Machine learning provides the possibility to tackle these estimation problems with more accurate tools that can handle more complex non-linearities and interactions and/or higher-dimensional confounders  $\mathbf{X}$ . The past five years has seen an avalanche of work in this space, with much of that effort going into tuning machine learning algorithms toward maximizing performance for the parameter we care about most (in this case, the average treatment effect) and working out the theoretical properties of various machine learning estimators. What the separation of identification and estimation highlights is that machine learning cannot help us with problems that are, at the core, identification issues (such as omitted variable bias).<sup>19</sup>

In practice, there are many ways to use machine learning for estimation in causal inference. For those looking to explore these methods, one approach is the augmented inverse propensity score weighting estimator (Robins et al., 1994), for which Glynn and Quinn (2010) offer an accessible introduction. Here the researcher specifies regression models for the treatment and outcome,

using the predictions for each to construct an estimate of the causal effect. There have been several advances on these methods across a variety of disciplines, but many of the core ideas are similar (Van der Laan and Rose, 2011; Athey and Imbens, 2016; Chernozhukov et al., 2018).<sup>20</sup> The advantage of estimators of these forms is that they allow the researcher to do the work that is closest to their expertise, choosing confounding variables for treatment and outcome models, while delegating the shape of the functional form (about which we generally have no strong theory) to the estimation algorithm.

## Prediction

There is a small but growing interest in prediction in the social sciences.<sup>21</sup> Some focus on the value of forecasting as a value in itself (Ward, 2016), while others have extolled the benefits of prediction for the evaluation of theory (de Marchi, 2005; Watts, 2014; Cranmer and Desmarais, 2017). Prediction also has a role in modern approaches to measurement such as text-as-data methods which replicate human content analysis coding at scale (Grimmer and Stewart, 2013), or techniques of ‘amplified asking’ which combine a relatively small survey with a much larger data source to predict what would have happened if a larger group of people had answered the survey (Blumenstock et al., 2015; Salganik, 2018).

In many of these settings our primary concern is that the predictions are accurate. When we can hold out a randomly selected test set to evaluate performance (for example, using k-fold cross validation or by applying a model to novel data), it is possible to directly estimate predictive accuracy in future data.<sup>22</sup> When would we expect to see large performance improvements from applying machine learning techniques? Machine learning excels in settings where the number of predictors is high relative to the size of the dataset and/or the features are individually relatively

uninformative, but jointly predictive. Thus, we will tend to see performance gains in settings where we have a substantial amount of data (so that it is feasible to estimate the model well), a difficult prediction problem (so that the baseline performance is poor) and noise swamping the signal in our predictors. Decades of careful work developing theory, designing measures and coding variables may mean that a parsimonious set of variables provides most of the available predictive information. However, as we transition into new sources of data – particularly those not explicitly designed by social scientists, such as administrative data or digital trace data – machine learning may become essential for identifying the predictive signal among the noise.

Similar to the discussion above on causal inference, computational methods from machine learning can help us develop the predictive algorithm, but have little to say about what the prediction actually *means*.<sup>23</sup> In causal inference, the meaning of the estimated quantity was conferred by the identification strategy. The significance of the predictive task itself seems rarely to be carefully evaluated in applied work, which can lead to dangerous results. Put simply, predictive methods will map the relationships that you have in your data, not the ones you wish you had. For example, you can use machine learning to predict whether or not a human loan officer will grant someone a loan. However, without a careful set of assumptions, akin to the identification task in causal inference, the machine learning algorithm will learn to predict like the human loan officer, replicating, for example, any racial or gender biases that the officer may have. Predictive algorithms that embed the discriminatory patterns of human decision makers have recently gained widespread attention and criticism (O’Neil, 2017; Eubanks, 2018) and as we move towards more forms of algorithmic governance, they will play an increasingly important role in how governments assess risk, allocate scarce resources and function on a day to day level.

In response to these concerns, computer scientists and economists have begun rigorous study of algorithmic fairness. Other than perfect predictions, it is impossible to simultaneously satisfy many intuitive definitions of fairness (Kleinberg et al., 2016; Corbett-Davies and Goel, 2018). These results imply that there are no easy answers and that considering the ethical implications of these tools will require substantial engagement with the social institutions within which these decision systems are embedded. This is why it is essential that more social scientists engage in these debates and study the deployment of these predictive tools in government settings.

In all of these areas of application, machine learning methods are not a panacea. They can certainly improve our ability to solve estimation problems, but that does not of itself solve the problem of connecting theory to empirics.

### *Problem 1: Multiplicity of parameter choices*

In machine learning algorithms, there are generally tuning parameters chosen by the researcher, such as the regularization penalty in Lasso regression or the number and depth of trees in a random forest. These choices can have a substantial impact on performance. When the objective is clear – for example, prediction accuracy on some target variable – a straightforward way (in simple settings) to set these parameters is through cross-validation (Hastie et al., 2009). Frustratingly, cross-validation itself has a tuning parameter (the number of folds), but the choice tends to matter less as we acquire more data.

The sensitivity of machine learning algorithms to parameters and other details of implementation varies substantially with the problem and the estimator.<sup>24</sup> This sensitivity is a challenge because researcher discretion can be harmful for the accumulation of scientific knowledge in the aggregate even if no individual researcher engages in egregious conduct (Simmons et al., 2011; King and Nielsen, 2019). While existing methods have always had space for researcher ‘discretion’,

the diversity and complexity of machine learning methods means there is considerably more space for these choices.

A second concern is computational reproducibility. Workhorse algorithms in machine learning such as Markov Chain Monte Carlo, stochastic gradient descent, the bootstrap and cross-validation rely on random number generators and can yield different results when started from different random seeds.<sup>25</sup> When the optimization problem being solved is straightforward, the observed variations may be slight. When the methods are solving complex optimization problems, the results can differ substantially (Roberts et al., 2016).<sup>26</sup> Even when using standard solutions such as fixing the random seed, reproducibility may be challenging given differences in the numerical precision of the linear algebra libraries or elements that arise for random assignment of jobs to different cores when using parallel processing. These issues are more pronounced on newer machine learning systems, but older machine learning methods such as support vector machines or Kernel Regularized Least Squares (Hainmueller and Hazlett, 2013) have a few of these issues as well. There is also an emerging exploration within statistics of a broader notion of stability in estimators (see a compelling state-of-the-research agenda in Yu, 2013).

A particularly challenging setting is when we are using machine learning to facilitate scientific discovery. For example, in causal inference we might be trying to identify subsets of the data in a randomized experiment where treatment effects are particularly strong (Imai and Ratkovic, 2013; Athey and Imbens, 2016; Grimmer et al., 2017; Ratkovic and Tingley, 2017). A promising approach in this setting is the use of a split sample (Fafchamps and Labonne, 2017; Egami et al., 2018), where we divide the data into a discovery set and an estimation set. In the discovery set, we look for interesting relationships. Once we have chosen the exact model we want to evaluate, we access the estimation set one time to produce our estimate of interest.

With well-intentioned researchers who are only accessing the estimation set *once*, this can address concerns of searching over endless sets of parameters and ultimately finding only noise. Any relationship that was random noise in the discovery set is highly unlikely to occur in the estimation set by construction. This strategy assumes a relatively straightforward randomized experiment with independent units. An important subject for future work is extending the sample splitting framework to more complex datasets (time-series cross-sectional, network and hierarchical) that we often find in political science.

While there are individual solutions to problems that arise from estimator sensitivity, computational instability and data-driven discovery, these concerns point jointly to a need to invest more heavily in scientific replications within and across research teams. Science is an iterative process and only by returning and reengaging with important problems will we be able to identify findings which are robust.

### *Problem 2: Opacity of the model*

A second concern is the opacity of the model – while regression provides us with a small set of parameters that define clear linear relationships, a modern machine learning algorithm might involve millions of parameters that control interactions and non-linear functions of thousands of inputs, obscuring the role each variable has in generating a prediction. The complexity of the model means that it is relatively difficult to guess how a small perturbation in one input of a model will affect the resulting prediction. This has led to calls in computer science for interpretable or explainable approaches to machine learning. One of the most popular methods to come out of this space is Local Interpretable Model-agnostic Explanations (LIME), which uses a simple (linear) approximation around a prediction of interest to produce an *ex post* explanation of a model (Ribeiro et al., 2016).<sup>27</sup>

Opacity of the machine learning model is only problematic if our use of the method

actually requires interpretability. In the causal inference setting, we often build up to an estimate of a single interpretable parameter (the average treatment effect) by estimating a series of functions which don't themselves need to be interpretable.<sup>28</sup> We can use as complicated a method as we want to estimate the various conditional expectations that make up our estimate of the average treatment effect  $\tau$  and the interpretation of  $\tau$  itself won't change. That is, machine learning is only being used to estimate the nuisance parameters that are necessary to focus on our real estimate of interest (Van der Laan and Rose, 2011; Chernozhukov et al., 2018). This makes explanation of the machine learning components largely unnecessary in the causal inference setting.

In fact, the opacity of some of these components might turn out to have the ancillary positive consequence that variables lacking theoretical interest in the model are not incorrectly interpreted. Interpretation of multiple parameters often requires extremely strong assumptions that practitioners do not realize they are making. Many perfectly plausible heuristics can, however, backfire spectacularly in common settings (see for example the use of heuristics for mediation models discussed in Glynn, 2012). If these parameters are unavailable to the practitioner, we may see fewer misinterpretations in practice.

In other settings, such as algorithms deployed for public policy, there may be a legal or policy requirement to be explainable.<sup>29</sup> The response to this in the machine learning community, as we alluded to above, has been to develop techniques for *ex post* explanations of the predictions. In practice, this typically means using a much simpler model to locally approximate the more complex model. This approach can be doubly misleading. First, we don't know that the simple model captures the complexities of the more complex model (even locally). Second, because the predictive model is not causally identified, explanations can appear to the untrained eye to be causal. For example,

we might observe that an actor is predicted to turn out to vote largely on the basis of the fact that they are a frequent reader of political news; however, this does not imply that if we could somehow convince a non-reader to engage more with political news, they would be more likely to vote.

If transparency of the model is important, a better alternative to *ex post* explanation is using a machine learning model that is designed to produce interpretable results by construction (Rudin, 2018). For example, Rudin's lab has developed a number of high performance machine learning algorithms based on various types of rule lists which can often be summarized in short lists of IF/AND/OR statements (Yang et al., 2017). The results from these techniques suggest that the common understanding that there is a trade-off between accuracy and interpretability may be (at least partly, and in some contexts) an illusion.

There is one subtle opacity problem that we don't have a good answer for: the issue of common support. Common support implies that for a given value of our control covariates, we have both treated and untreated observations. For example, in studying the effect of career choice on political preferences controlling for education, we may not have any doctors who only finished high school in the sample. When we lack common support, our estimates are necessarily reliant on extrapolation and thus become increasingly dependent on arbitrary modeling choices (King and Zeng, 2006). One benefit of approaches like matching is that they drop units where we lack common support, giving us more confidence in our causal estimates. This is easy enough to do with one control variable, such as education in our above example, but is much more complicated when we have more than a handful of variables. By obscuring the prediction process with black-box machine learning algorithms, we may be extrapolating beyond the common support in various regions of the data without realizing it. D'Amour et al. (2017) suggest that this lack

of common support is an inevitable and pervasive problem in the high-dimensional settings where causal inference techniques are being applied. Identifying and addressing regions where we lack common support strikes us as a valuable area of future research.

To summarize, machine learning provides an improved approach to estimation, but estimation only gets us so far. In causal inference, we must separate identification concerns (which machine learning does not currently help with) and estimation concerns (where machine learning may be a useful replacement for existing tools). Although somewhat less common in practice, a similar thought process applies to prediction, where it is necessary to think through the implication of the estimation problem rather than treating it as an end unto itself. Theory retains a central role in setting the target estimand (whether predictive or causal) and shaping the inclusion and measurement of key inputs. Even in ideal settings, machine learning methods struggle with problems that arise due to an expansion of parameter options and opacity, but these problems are limited in scope and we are optimistic about the long-term future prospects for addressing them.

## REPRESENTATION

Computation has fundamentally altered not just estimation, but also the way we think about the measurement of the variables we study. Some of these changes are conceptually straightforward and use computation to automate basic human tasks. For example, in the field of computer-assisted text analysis, machine learning algorithms are used to mimic human coders and classify documents into pre-specified categories, identify passages of copied text (Grimmer and Stewart, 2013; Wilkerson et al., 2015) and classify text based on its complexity (Benoit et al., 2018). We focus this section on approaches to measurement that leverage unsupervised

machine learning and representation learning (Bengio et al., 2013).

Unsupervised learning is an area of machine learning that is less focused on prediction and more focused on dimension reduction and classification. The mathematical goal is to compress a high-dimensional observation into a low-dimensional representation while preserving as much information as possible. For example, we might represent a legislator's voting record by a single number, which stands in for the full set of votes in a given legislative session. The social scientist then makes an important conceptual move: interpreting and naming this dimension 'ideology'. This basic idea has been cross-applied to survey analysis (latent class analysis, factor analysis), text analysis (topic models, scaling, word embedding), voting patterns (ideal points) and networks (latent spaces, community detection). The interpretation of latent structure is not new (factor analysis is more than a hundred years old) but computational power has extended the reach of these tools.

It is important to acknowledge that the label that we place on the latent structure is an *ex post* interpretation and does not necessarily guarantee verisimilitude. For example, ideal point algorithms measure 'ideology' not because they have access to a notion of what ideology is, but because the dimension which captures the most variation in votes may be related to ideology. This emphasizes the point that the interpretation is not justified by the method; rather, it is guided by theory and validated by further empirical tests. To draw an example from the text analysis literature, the Wordfish model (Slapin and Proksch, 2008) introduced a means to analyze left-right positions based on party manifestos. However, nothing in the method entails that the method captures ideology in documents; in general, it will only do so when ideology is the dimension of maximal variation in the text themselves. In demonstrating this point, Grimmer and Stewart (2013) show that applying Wordfish to a corpus of US Senate

press releases does not return a measure of ideology at all (and in fact even fails to separate press releases by party). This is not a failure of the method, just a reminder that the interpretation is a distinct and difficult step.

What then are we to do? Whether the representation involves latent classes in a survey, topics in a document or communities in a network, it is important to recognize that the phenomenon being captured by the latent structure is different than the parsimonious label we put to it. Take for example a topic model with a topic we choose to label ‘economics’. The ‘economics’ topic is actually a distribution of weights over the entire vocabulary (containing often tens of thousands of words). This is at once substantially more complex than the five to ten words that we might display in a figure for the reader and substantially simpler than a human understanding of economics as a field. The name may evoke in the reader the idea that the measurement instrument contains things the topic doesn’t capture and the topic may pick up much more than the reader believes. There are better and worse ways to handle this conceptual gap between the label and the reality, but we echo the call in Grimmer and Stewart (2013) for a substantial focus on validation (not just in text applications but in all uses of unsupervised methods). Unfortunately, as these methods have become more routinized, the extensive validations that were present in early work (for example Grimmer, 2010) have become less rigorous.

It may be helpful to think of the results of (particularly unsupervised) machine learning models for representation as producing a kind of ‘found data’. Much of the data we analyze in the social sciences comes from academics designing a data collection procedure and creating a measure (for example, designing a survey question and then fielding the survey). Increasingly, we are turning to more instances of found data – administrative records, digital trace data and other forms of data that were collected for a purpose different than our own research. These data can

be useful, but it is best to approach their use with a healthy degree of skepticism and carefully investigate and validate that they measure what we hope they measure. In much the same way, representations in machine learning models are often designed to meet a certain criterion which is different than the interpretation that we later put on it.

At some level, it is remarkable that this approach to interpretation works at all. Yet, carefully validated representations have led to numerous interesting discoveries in social science. For example, Garip (2012) discovered theoretically interesting categories of Mexican migrants using cluster analysis and Catalinac (2016) used topic models to develop a theory of how electoral reform in Japan turned national policy from a focus on pork spending to foreign policy. In both cases, the representations serve an important role as measurements of theoretical concepts. What establishes the validity of these measurements, though, is not the fact that they are produced by a machine learning algorithm, but rather the careful theoretical work and subsequent validation that supports them.

### *Problem 1: Multiplicity of parameter choices*

Unsupervised machine learning techniques, as with other computational models, have many parameter choices which impact the estimates we see. We use the example of unsupervised text analysis here, although the problem is general. Denny and Spirling (2018) show that different choices of pre-processing (stemming, stop-word removal and other data cleaning tasks for text analysis) can result in learning very different topics for the same documents. Perhaps even more disturbingly, the same topic model run on the same data can produce qualitatively different answers using different starting values due to the difficulty of the optimization problem (Roberts et al., 2016).

The problem is most acute if you believe there is a single correct representation for a given document that the model ought to be

able to recover; that is, that there is a ‘right answer’. This is not generally how we think about measurement, though. There are many possible readings of a text and many different ways of organizing a text collection. In one of the earliest applications of statistical text analysis, Mosteller and Wallace (1962) analyzed the *Federalist Papers* to determine who authored them (Hamilton or Madison). This question has a correct answer: one of those two almost certainly wrote each essay, and we would expect an authorship model to be able to correctly recover the author. But consider if we were to ask a topic model to characterize what the essays are about. There is no single correct way to characterize what the different *Federalist Papers* are about, even if some interpretations are more or less useful.

A given machine learning algorithm with a given set of parameters provides access to one possible organizational structure. Social science theory takes an analyst from a set of parameters that compress the data effectively, to a substantively meaningful measure of a social or political quantity. From this perspective, a large set of parameter choices and non-unique solutions may be an unavoidable consequence of a complex and interesting world.

### ***Problem 2: Opacity of the model***

Machine learning models for representation learning that have seen use in the social sciences are often very simple. Topic models, for instance, represent a given document as having proportional membership across a fixed set of classes; for example, a document could be 60% about economics and 40% about cooking. While the topic model assumes that the document can be represented by multiple topics, the form of those topics is additive, meaning that ‘cooking’ words look the same regardless of whether they are combined with ‘economics’ or with ‘science’. More complex models that are interactive, allowing for the words associated with a topic to be different based on the topic mixture, have existed for some time

(Larochelle and Lauly, 2012), but they have not found use in the social sciences precisely because they complicate the simple interpretation of the topics. Simplicity in a representation is an important part of its virtue because it makes the measure easily understandable.

We note that our call for simplicity in representations is in contrast to our discussion of theoretical models, where we argued that simplicity should be cast aside to more closely represent the real world. However, a lack of simplicity in representation is distinct because it threatens the conceptual homogeneity of the measurement. That is, if the topic for ‘cooking’ can vary dramatically when combined with other topics, it is no longer clear that multiple documents which share the cooking label are really about the same thing. We can add more dimensions (topics) to our measurement in order to increase the complexity of the phenomena we can represent, but to improve our understanding of what we are measuring, simplicity of the individual topics is extremely helpful.

Even though the representation of the model may be simple, the inference procedures behind these methods are often quite complex. This can make it challenging for applied scholars to know what can and cannot be inferred from the model. To give a trivial example, the ordering of the topics in most topic models is unidentified (that is, you can swap topic 1 and topic 2 without changing the objective function). Thus, we would not want to try to draw conclusions about which topic was first. Other examples abound: the question of what is and is not identified has been carefully studied for ideal point voting models (Rivers, 2003) but this has not been accomplished for newer forms of unsupervised learning. As a result, we have seen numerous examples of dubious inferences drawn from latent variable models such as conclusions about the sparsity of topics in documents, the orientation of a low-dimensional projection or the magnitude in a unidimensional scaling model. In the long



term, the field will develop a more careful study of such models, but in the short term we emphasize again that validating findings with information directly from the raw documents can help avoid these kinds of issues.

The opacity of even simple models may however blind people to a basic point about representations: they will reflect the data they are trying to represent. Word embeddings are a recently popular representation of language in modern natural language processing. The embeddings learn a vector for each word in a low-dimensional space that predicts the kinds of words that are nearby (that is, in a neighborhood – see Mikolov et al., 2013). They are well known for being able to solve analogic reasoning tasks by adding and subtracting the learned vectors. For example, the analogy ‘France is to Paris as Spain is to \_\_\_\_’ can be solved by taking the vector for ‘Paris’ minus ‘France’ plus ‘Spain’. The resulting vector will be close to the correct answer, ‘Madrid’. While these analogy tasks are interesting, the real advantage is that the representations encode similarity between words that allows information to be efficiently shared. The vectors can be quickly trained on very large corpora (for example, all of Wikipedia or the common crawl of the internet) and the information-sharing leads to substantial performance improvements in downstream models.

That said, Caliskan et al. (2017) demonstrate that word embeddings show evidence of human-like biases (such as racial and gender stereotypes). This piece attracted considerable attention and prompted a sequence of papers demonstrating other biases (Garg et al., 2018). While this was surprising to some, from a social scientist’s perspective it was inevitable. Sexism and racism are pervasive in the language people use on the internet and thus representations trained on that language will reflect those associations. That is, the representation learning does not learn an ‘objective’ property of language (if that even exists), but the associations of how that language is deployed in the training corpus. In this case the opacity of the method may

have obscured a more basic truth about how the representation was constructed and what it means.

To summarize, representation learning has become a powerful method for measurement and discovery in the social sciences. While machine learning is complex in these settings and occasionally opaque, the ever-present challenges are more about the interpretations of measures than anything else. This places the challenge firmly in the theory and substantive understanding of the problem at hand.

## CONCLUSION

The future of computation models in the social sciences is bright. Increased computation has led to improvements in theoretical models, estimation and representation, as well as an expansion of the evidence base in social science (Lazer et al., 2009). To the extent that we desire more general models, new sources of data and an increased ability to predict human behavior, there is no substitute for computation. There are, however, genuine limitations and challenges to the increased reliance on computation. Computation will not solve fundamental underlying social science problems such as the need for validation in measurement or identification in causal inference. We identified two common sets of challenges: the sensitivity to parameters and the potential for opacity. In both of these cases, substantive theory is most often the best guide to making choices and producing results that are interpretable. The best use of computation thus requires an iterative process in which theory and empirical work are more closely combined.

We are also concerned that current reproducibility standards, although vastly improved in recent years, are not always adequate for the task of validating computational models. We hope that more attention

will be devoted to these issues, with a particular focus on going beyond reproducing tables in a paper and instead understanding the logic of the procedure and the role key methodological choices have in producing results. The capacity for computation will keep increasing; let's make the best use of it we can.

## Notes

- 1 Most social scientists readily understand the parameters involved in applied statistical modeling; for example, for a Normal distribution we have two parameters, the mean and the variance. The definition of parameter spaces is broader than this, however, and includes all of the choices involved in producing a mathematical model in both applied statistics and formal theory. For example, the choice of equilibrium refinement is a parameter.
- 2 It was clear at the onset of Von Neumann and Morgenstern's work that the focus on 'simple' games was a starting point and not the end point of theory: 'the discussions which follow will be dominated by illustrations from Chess, Matching Pennies, Poker, Bridge, etc., and not from the structure of cartels, markets, oligopolies, etc.' (von Neumann and Morgenstern, 1944: 47).
- 3 Backwards induction scales poorly as the complexity of the choice context increases. As the standard algorithm for 'rational' choice, it fares poorly in most settings of interest (Russell and Norvig, 2016).
- 4 In large part, the National Science Foundation's Empirical Implications of Theoretical Models (EITM) tradition in political science has yet to come to grips with this issue. EITM's original conception was that one would write down a non-cooperative game and then test it with data (see, for example, Signorino, 1999). The difficulty that persists is that the results of games are far afield from the data, complicating empirical testing.
- 5 For more on the importance of robustness in the area of political institutions, see Bednar (2008). The point raised in this work is general, however.
- 6 Backwards induction on a tree is often conflated with 'rational choice' actors. As decision-making algorithms go, backwards induction is not very efficient and does not scale well with the size of the extensive form of the game (that is, the complexity of the game – see Russell and Norvig, 2016, for an overview).
- 7 For more on this topic, see Laver and Sergenti (2011).
- 8 For an overview of the issues involved, see the EITM report generated by the National Science Foundation: <https://www.nsf.gov/sbe/ses/polisci/reports/pdf/eitmreport.pdf>.
- 9 Minimum integer weights, such as the Shapley and Banzhaf value, focus on the pivotality of parties in a bargaining or weighted voting game. Loosely, they are the smallest vector of integers that reproduce the coalitions derived from the raw weights for each actor and represent how pivotal each actor is relative to the others.
- 10 And, even if it remains exogenous, changing this assumption on proposals to one that is a monotonic function of player strength would provide a greater advantage to players with more power.
- 11 Homogeneous games have the property that all minimal winning coalitions are of equal voting weight.
- 12 Best practice is to approach computational models of this sort as Monte Carlo – parameters are varied across many (many!) trials.
- 13 To be concrete, one can focus on the proposer advantage in bargaining models as a salient example. This advantage springs from the closed rule version of Baron and Ferejohn's model and is oft-cited in the literature. But, noting the dominant empirical expectation that parties receive a strictly proportional payoff (that is, Gamson's Law), Morelli's demand bargaining model does not produce a proposer advantage. By manipulating parameters, it is possible to produce any outcome one wants; this is not a productive way to build confidence in the actual causal mechanisms underlying real-world outcomes.
- 14 The same is true of other, related endeavors such as Google's machine learning model AlphaGo, which has been able to beat the best humans on the planet.
- 15 Many areas of social science could benefit from this flexibility. For example, the causes of war are complex and depend on multiple factors, and effects may be conditional based on the historical context (see Jenke and Gelpi, 2017). Machine learning approaches are likely dominant in such applied settings.
- 16 Easy extensions such as regularization can address some of these problems. See [https://web.stanford.edu/~hastie/glmnet/glmnet\\_beta.html](https://web.stanford.edu/~hastie/glmnet/glmnet_beta.html) for a good introduction to regularization.
- 17 The logistic regression used the same set of covariates in a linear additive form. As noted in note 28, adding all two-way interactions induced overfitting and worse test-set performance.

- 18 We are glossing over the edge cases, such as linear dependence in the predictors or a number of predictors that exceed the number of observations.
- 19 There is an area of research on using machine learning to help with identification (Spirtes et al., 2000; Peters et al., 2017) by filling in unknown edges in a directed acyclic graph. These methods still involve strong theoretical assumptions and we think they are unlikely to enter into the mainstream of social science in the near future.
- 20 In R, we recommend the grf package as a starting point (Tibshirani et al., 2018).
- 21 We are not certain why prediction is not more prominent in the social sciences. It may be due to a lack of agreement on dependent variables (Spirling, personal communication) or due to the lack of coordination between theorists and empiricists (<https://www.nsf.gov/sbe/ses/polisci/reports/pdf/eitmreport.pdf>).
- 22 It is worth emphasizing that in true forecasting situations this isn't possible. Political phenomena are dynamic in a way that means models which can predict the past effectively may perform quite poorly in predicting the future (Bowlsby et al., 2019). Similarly, when phenomena are interdependent, simple cross validation is not straightforwardly applicable anyway.
- 23 It is also worth noting that feature selection usually precedes building forecasting models and this step is most often guided by theory (see Bishop, 1995).
- 24 It can be difficult to tell when a choice is particularly important. For example, there are different ways of solving the Lasso regression problem (coordinate descent, least angle regression, etc.) which will in general converge to extremely similar solutions. However, the way in which the regularization parameter is chosen can completely change the quality of the resulting model.
- 25 Non-stochastic algorithms solving non-convex optimization problems depend on the initial starting values. This important issue is generally less about randomness in the algorithm than about the difficulty of the optimization problem to be solved.
- 26 This means that the algorithm may converge to only a local optimum and not the 'true' global solution, which would further minimize the loss function on the observed data. This is not surprising given the very large size of the parameter spaces involved. The value of reproducibility is thus lessened in cases where all one is reproducing is a local optimum.
- 27 While an improvement, local approaches such as LIME still present challenges for applying models in policy settings. It is fair to say that most interven-

- tions/programs are relatively coarse and expect that the main effects of key variables are monotonic.
- 28 Causal estimands are often written as the difference between two conditional expectation functions. The estimand itself needs to be interpretable but the two conditional expectation functions don't need to be.
- 29 For example, the European Union's General Data Protection Regulation (GDPR). More generally, if we want to make policy interventions based on a model, prediction alone is not sufficient if we cannot understand the relationship between variables in the model. It is fair to say that most policy makers expect that when they pull a lever (that is, conduct a program of intervention), it has a monotonic effect – they do not expect an effect that is conditional on other (perhaps unobserved) levers or that the effect changes based on how hard one pulls.

## REFERENCES

- Alchian, Armen A. 1950. Uncertainty, evolution, and economic theory. *Journal of Political Economy* 58(3): 211–221.
- Alvarez, R. Michael, ed. 2016. *Computational Social Science*. Cambridge University Press.
- Angrist, Joshua D. and Jörn-Steffen Pischke. 2008. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.
- Athey, Susan and Guido Imbens. 2016. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences* 113(27): 7353–7360.
- Banks, Jeffrey S. and John Duggan. 2006. A general bargaining model of legislative policy-making. *Quarterly Journal of Political Science* 1(1): 49–85.
- Barberá, P., J. T. Jost, J. Nagler, J. A. Tucker and R. Bonneau. 2015. Tweeting from left to right: is online political communication more than an echo chamber? *Psychological Science* 26(10): 1531–1542.
- Baron, David, and John Ferejohn. 1989. Bargaining in legislatures. *American Political Science Review* 83(4): 1182–1202.
- Bednar, Jenna. 2008. *Robust Federation: Principles of Design (Political Economy of Institutions and Decisions)*. Cambridge University Press.

- Bengio, Yoshua, Aaron, Courville and Pascal Vincent. 2013. Representation learning: a review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35(8): 1798–1828.
- Benoit, Kenneth, Kevin Munger and Arthur Spirling. 2018. *Measuring and Explaining Political Sophistication through Textual Complexity*. Available at SSRN 3062061.
- Billings, Darse, Denis Papp, Jonathan Schaeffer and Duane Szafron. 1998. Opponent modeling in poker. *Proceedings of the Fifteenth National Conference on Artificial Intelligence, Madison, Wisconsin*, pp. 493–499.
- Binmore, Ken. 1992. *Fun and Games: A Text on Games*. DC Heath and Co.
- Bishop, Christopher M. 1995. *Neural Networks for Pattern Recognition*. Oxford University Press.
- Blumenstock, Joshua, Gabriel Cadamuro and Robert On. 2015. Predicting poverty and wealth from mobile phone metadata. *Science* 350(6264): 1073–1076.
- Bowlsby, Drew, Erica Chenoweth, Cullen Hendrix and Jonathan D. Moyer. 2019. The future is a moving target: predicting political instability. *British Journal of Political Science* 1–13.
- Brown, Noam and Tuomas Sandholm. 2018. Superhuman AI for heads-up no-limit poker: Libratus beats top professionals. *Science* 359(6374): 418–424.
- Caliskan, Aylin, Joanna J. Bryson and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356(6334): 183–186.
- Catalinac, Amy 2016. *Electoral Reform and National Security in Japan: From Pork to Foreign Policy*. Cambridge University Press.
- Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey and James Robins. 2018. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal* 21(1): C1–C68.
- Corbett-Davies, Sam and Sharad Goel. 2018. The measure and mismeasure of fairness: a critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*.
- Cranmer, Skyler J. and Bruce A. Desmarais. 2017. What can we learn from predictive modeling? *Political Analysis* 25(2): 145–166.
- D'Amour, Alexander, Peng Ding, Avi Feller, Lihua Lei and Jasjeet Sekhon. 2017. Overlap in observational studies with high-dimensional covariates. *arXiv preprint arXiv:1711.02582*.
- de Marchi, Scott. 2005. *Computational and Mathematical Modeling in the Social Sciences*. Cambridge University Press.
- de Marchi, Scott and Scott E. Page. 2008. Agent-based modeling. In Janet M. Box-Steffensmeier, Henry E. Brady, and David Collier (eds) *The Oxford Handbook of Political Methodology*. Oxford University Press.
- de Marchi, Scott and Scott E. Page. 2014. Agent-based models. *Annual Review of Political Science* 17: 1–20.
- de Marchi, Scott and Michael Laver. 2020. Government formation as logrolling in high-dimensional issue spaces. *The Journal of Politics*. <https://doi.org/10.1086/706462>.
- de Marchi, Scott, Christopher Gelpi and Jeffrey D. Grynawski. 2004. Untangling neural nets. *American Political Science Review* 98(2): 371–378.
- de Marchi, Scott, Michael Laver and Georg Vanberg. Government Formation in the Shadow of an Uncertain Future Election. Working paper.
- Denny, Matthew J. and Arthur Spirling 2018. Text preprocessing for unsupervised learning: why it matters, when it misleads, and what to do about it. *Political Analysis* 26(2): 168–189.
- Diermeier, Daniel and Keith Krehbiel. 2003. Institutionalism as a methodology. *Journal of Theoretical Politics* 15(2): 123–144.
- Egami, Naoki, Christian J. Fong, Justin Grimm, Margaret E. Roberts and Brandon M. Stewart. 2018. How to make causal inferences using texts. *arXiv preprint arXiv:1802.02163*.
- Ensley, Michael J., Michael W. Tofias and Scott de Marchi. 2009. District complexity as an advantage in congressional elections. *American Journal of Political Science* 53(4): 990–1005.
- Eubanks, Virginia. 2018. *Automating Inequality: How High-tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press.

- Fafchamps, Marcel and Julien Labonne. 2017. Using split samples to improve inference on causal effects. *Political Analysis* 25(4): 465–482.
- Fowler, James H. 2006. Connecting the Congress: a study of cosponsorship networks. *Political Analysis* 14(4): 456–487.
- Dorff, Cassy, Max Gallop and Shahryar Minhas. Networks of violence: predicting conflict in Nigeria. *Journal of Politics*, 2018.
- Garg, Nikhil, Londa Schiebinger, Dan Jurafsky and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences* 115(16): E3635–E3644.
- Garip, Filiz. 2012. Discovering diverse mechanisms of migration: the Mexico–US stream 1970–2000. *Population and Development Review* 38(3): 393–433.
- Gilpin, Andrew and Tuomas Sandholm. 2007. Better automated abstraction techniques for imperfect information games, with application to Texas Hold'em poker. *Proceedings of the 6th International Joint Conference on Autonomous Agents and Multiagent Systems*. ACM.
- Glynn, Adam N. 2012. The product and difference fallacies for indirect effects. *American Journal of Political Science* 56(1): 257–269.
- Glynn, Adam N. and Kevin M. Quinn. 2010. An introduction to the augmented inverse propensity weighted estimator. *Political Analysis* 18(1): 36–56.
- Grimmer, Justin. 2010. A Bayesian hierarchical topic model for political texts: measuring expressed agendas in Senate press releases. *Political Analysis* 18(1): 1–35.
- Grimmer, Justin and Brandon M. Stewart 2013. Text as data: the promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis* 21(3): 267–297.
- Grimmer, Justin, Solomon Messing and Sean J. Westwood. 2017. Estimating heterogeneous treatment effects and the effects of heterogeneous treatments with ensemble methods. *Political Analysis* 25(4): 413–434.
- Hainmueller, Jens 2012. Entropy balancing for causal effects: a multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis* 20(1): 25–46.
- Hainmueller, Jens and Hazlett, Chad. (2013). Kernel regularized least squares: reducing misspecification bias with a flexible and interpretable machine learning approach. *Political Analysis*. DOI: <http://dx.doi.org/10.2139/ssrn.2046206>
- Hastie, Trevor, Robert Tibshirani and Jerome Friedman. 2009. *The Elements of Statistical Learning*. Springer.
- Ho, Daniel E., Kosuke Imai, Gary King and Elisabeth A. Stuart. 2007. Matching as non-parametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis* 15(3): 199–236.
- Holland, John H. and John H. Miller. 1991. Artificial adaptive agents in economic theory. *The American Economic Review* 81(2): 365–370.
- Imai, Kosuke and Marc Ratkovic. 2013. Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics* 7(1): 443–470.
- Imbens, Guido W. and Donald B. Rubin. 2015. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press.
- Jenke, Libby and Christopher Gelpi. 2017. Theme and variations: historical contingencies in the causal model of interstate conflict. *Journal of Conflict Resolution* 61(10): 2262–2284.
- Judd, Kenneth and Scott E. Page. 2004. Computational public economics. *Journal of Public Economic Theory* 6(2): 195–202.
- Kim, Sung-young, Charles S. Taber and Milton Lodge. 2010. A computational model of the citizen as motivated reasoner: modeling the dynamics of the 2000 presidential election. *Political Behavior* 32(1): 1–28.
- King, Gary and Richard, Nielsen. 2019. Why propensity scores should not be used for matching. *Political Analysis* 27(4): 435–454.
- King, Gary and Langche Zeng. 2006. The dangers of extreme counterfactuals. *Political Analysis* 14(2): 131–159.
- Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig and Sendhil Mullainathan. 2018. Human decisions and machine predictions. *The Quarterly Journal of Economics* 133(1): 237–293.
- Kleinberg, Jon, Sendhil Mullainathan and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807* (2016).

- Kollman, Ken, John H. Miller and Scott E. Page. 1992. Adaptive parties in spatial elections. *American Political Science Review* 86(4): 929–937.
- Kollman, Ken, John H. Miller and Scott E. Page. 1997. Political institutions and sorting in a Tiebout model. *The American Economic Review* 87(5): 977–992.
- Kollman, Ken, John H. Miller and Scott E. Page. 1998. Political parties and electoral landscapes. *British Journal of Political Science* 28(1): 139–158.
- Kreps, David M. and Robert Wilson. 1982. Reputation and imperfect information. *Journal of Economic Theory* 27(2): 253–279.
- Larochelle, Hugo and Stanislas Lauly. 2012. A neural autoregressive topic model. In P. Bartlett, F. Pereira, C. J. C. Burges, L. Bottou and K. Q. Weinberger (eds) *Advances in Neural Information Processing Systems 25*, Curran Associates, Inc., pp. 2708–2716.
- Laver, Michael. 2005. Policy and the dynamics of political competition. *American Political Science Review* 99(2): 263–281.
- Laver, Michael and Ernest Sergenti. 2011. *Party Competition: An Agent-based Model*. Vol. 18. Princeton University Press.
- Lazer, David, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis et al. Computational social science. *Science* 323(5915): 721–723. 2009.
- Martin, Lanny and Georg Vanberg. 2019. What You See Is Not Always What You Get: Bargaining before an audience under Multiparty Government. Working paper.
- Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pp. 3111–3119.
- Miller, John and Scott Page. 2007. *Complex Adaptive Systems: Computational Models of Social Life*. Princeton University Press.
- Minhas, Shahryar, Peter D. Hoff and Michael D. Ward. 2019. Inferential approaches for network analysis: AMEN for latent factor models. *Political Analysis* 27(2): 208–222.
- Morelli, Massimo. 1999. Demand competition and policy compromise in legislative bargaining. *American Political Science Review* 93(4): 809–820.
- Morgan, Stephen L. and Christopher Winship. 2014. *Counterfactuals and Causal Inference (2nd ed.)*. Cambridge University Press.
- Mosteller, Frederick and David L. Wallace. 1962. Notes on an authorship problem. In *Proceedings of a Harvard Symposium on Digital Computers and Their Applications, 3–6, April, 1961*. Harvard University Press, pp. 163–197.
- Nash, John F. Jr. and Lloyd S. Shapley. 1950. A simple three-person poker game. *Essays on Game Theory* (1993). Edward Elgar Press.
- Nguyen, Thien Huu and Ralph Grishman. 2018. Graph convolutional networks with argument-aware pooling for event detection. *Thirty-Second AAAI Conference on Artificial Intelligence*.
- O’neil, Cathy. 2017. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Broadway Books.
- Pearl, Judea and Dana Mackenzie. 2018. *The Book of Why: The New Science of Cause and Effect*. Basic Books.
- Peters, Jonas and Dominik Janzing, Bernhard Schölkopf. 2017. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press.
- Rapoport, Amnon, Ido Erev, Elizabeth V. Abraham and David E. Olson. 1997. Randomization and adaptive learning in a simplified poker game. *Organizational Behavior and Human Decision Processes* 69(1): 31–49.
- Ratkovic, Marc and Dustin Tingley. 2017. Sparse estimation and uncertainty with application to subgroup analysis. *Political Analysis* 25(1): 1–40.
- Ribeiro, M. T., S. Singh and C. Guestrin 2016. Why should I trust you? Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144. ACM.
- Rivers, Douglas. 2003. Identification of multidimensional spatial voting models. Typescript. Stanford University.
- Roberts, Margaret E. Roberts, Brandon M. Stewart and Dustin Tingley. 2016. Navigating the local modes of big data: the case of topic models. In R. Michael Alvarez, ed. *Computation Social Science: Discovery and Prediction*. Cambridge University Press, pp. 51–90.

- Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley and Edoardo M. Airoldi (2013, January). The structural topic model and applied social science. In *Advances in Neural Information Processing Systems Workshop on Topic Models: Computation, Application, and Evaluation* (pp. 1–20).
- Robins, James M., Andrea Rotnitzky and Lue Ping Zhao. 1994. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* 89(427): 846–866.
- Rozenas, Arturas, Shahryar Minhas, and John Ahlquist. 2019. Modeling asymmetric relationships from symmetric networks. *Political Analysis* 27(2): 231–236.
- Rubenstein, Ariel. 1982. Perfect equilibrium in a bargaining model. *Econometrica* 50(1): 97–109.
- Rudin, C. 2018. Please stop explaining black box models for high stakes decisions. *arXiv preprint arXiv:1811.10154*.
- Russell, Stuart J. and Peter Norvig. 2016. *Artificial Intelligence: A Modern Approach*. (3rd ed.) Pearson Education.
- Salganik, Matthew J. 2018. *Bit by Bit: Social Research in the Digital Age*. Princeton University Press.
- Sandholm, Tuomas. 2010. The state of solving large incomplete-information games, and application to poker. *AI Magazine* 31(4): 13–32.
- Signorino, Curtis S. 1999. Strategic interaction and the statistical analysis of international conflict. *American Political Science Review* 93(2): 279–297.
- Simmons, Joseph D., Leif D. Nelson, Uri Simonsohn. 2011. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science* 22(11): 1359–1366.
- Slapin, Jonathan B., and Sven-Oliver Proksch. 2008. A scaling model for estimating time-series party positions from texts. *American Journal of Political Science* 52(3): 705–722.
- Spirites, Peter, Clark N., Richard Scheines, David Heckerman, Christopher Meek, Gregory F. Cooper and Thomas Richardson. 2000. *Causation, Prediction, and Search*. (2nd ed.) MIT Press.
- Tibshirani, Julie, Susan Athey, Stefan Wager, Rina Friedberg, Luke Miner and Marvin Wright. 2018. *grf: Generalized Random Forests (Beta)*. R package version 0.10.2.
- van der Laan, Mark J. and Sherri Rose. 2011. *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer Science & Business Media.
- von Neumann, John and Oskar Morgenstern. 1944. *Theory of Games and Economic Behavior*. Princeton University Press.
- Ward, Michael D. 2016. Can we predict politics? Toward what end? *Journal of Global Security Studies* 1(1): 80–91.
- Watts, Duncan J. 2014. Common sense and sociological explanations. *American Journal of Sociology* 120(2): 313–351.
- Wilkerson, John, David Smith, and Nicholas Stramp. 2015. Tracing the flow of policy ideas in legislatures: a text reuse approach. *American Journal of Political Science* 59(4): 943–956.
- Yang, Bishan and Tom Mitchell. 2016. Joint extraction of events and entities within a document context. *arXiv preprint arXiv:1609.03632*.
- Yang, Hongyu, Cynthia Rudin and Margo Seltzer. 2017. Scalable Bayesian rule lists. In *Proceedings of the 34th International Conference on Machine Learning, Volume, 70*: 3921–3930.
- Yu, Bin. 2013. Stability. *Bernoulli* 19(4): 1484–1500.



# Learning and Diffusion Models

Scott J. LaCombe and Frederick J. Boehmke

## **POLICY DIFFUSION**

In this chapter, we provide an overview of different methodological approaches used to study diffusion and learning. In political science, studies of diffusion seek to capture interdependence across units in the spread of an idea or action, such as the adoption of a policy by a state or country, the signing of compacts or treaties, the adoption of a norm of behavior, or the occurrence of political conflict. A variety of methods has been developed to understand how such items spread, each with its own advantages and disadvantages. We will highlight the theoretical foundations of diffusion research as well as the corresponding methodologies to understand how phenomena spread.

In American politics, diffusion research often focuses on the spread of policies across states and cities. US states make an ideal unit of analysis for diffusion studies because they all operate at the same level in the same federal system. Policy adoptions also represent

easily observable, discrete events that are conducive to large-N quantitative analysis. Research in US state policy diffusion typically asks a variety of questions: why do some policies spread widely across states, countries or cities, and others do not? What makes a government innovative? Are some governments innovative in certain policy areas but not others? How do states make decisions about what states to use as a model for policy ideas? Since Walker's (1969) foundational work on policy diffusion, political scientists have become increasingly focused on understanding how policies spread from one governing body to another. Diffusion scholars want to know what makes some policies spread widely and not others (Boushey, 2010; Nicholson-Crotty, 2009). In American politics, some states, such as California and Massachusetts, are historically innovative (Boehmke and Skinner, 2012), while others, such as Wyoming and West Virginia, usually wait for others to go first before adopting a policy. Studies of American politics have



helped scholars understand what makes policies more likely to spread and states or cities more likely to innovate.

International relations and comparative politics research also consider diffusion (Zhukov and Stewart, 2013). As with researchers in American politics, scholars of cross-national diffusion recognize the role of international interdependence in the spread of political phenomena (Gilardi, 2012): the actions of one country often influence the decisions of other actors in the international system. Countries often use a combination of carrots and sticks to explicitly influence other countries to adopt their preferred policies or trade practices, or engage in a variety of other interstate behavior (Simmons et al., 2006). Research ranges from the diffusion of macro-economic policies cross-nationally (Simmons and Elkins, 2004) to the diffusion of democratic governments (Starr, 1991), government spending levels (Lee and Strang, 2006) and conflict (Gleditsch, 2009).

The prior examples illustrate that many phenomena are influenced by diffusion. But why does diffusion occur? Researchers seeking to understand the causes of diffusion rely on four widely agreed upon mechanisms: coercion, imitation, competition and learning (Simmons et al., 2006; Shipan and Volden, 2008; Jensen and Lindstädt, 2012). In this chapter, we will focus on learning. Diffusion via learning occurs when a government or other actor uses information resulting from the policy or other choices made by other actors. If those choices prove beneficial, then other governments will tend to make similar choices (Butler et al., 2017). Learning requires that policymakers are aware of both what policies other actors are adopting and the outcomes produced by those adoptions.

The primary empirical method for studying diffusion has long been event history analysis (EHA). In recent years, however, a wide variety of alternatives have emerged that can incorporate the diffusion of multiple policies and more explicitly model and account for

interdependence among observations. We will provide an overview of the following methodological approaches that have been used to study diffusion in political science: event history analysis, dyadic event history analysis, pooled event history analysis, network analysis, latent networks and spatial regression.

## THEORETICAL BACKGROUND

This section focuses on the four mechanisms of diffusion described above and briefly traces the evolution of the literature on policy diffusion, with examples from related areas (e.g., conflict, monetary policy). We recap many of the original ideas on the diffusion of innovations, starting with Rogers in 1939, and briefly trace their development in the political science literature (Weyland, 2005; Graham et al., 2013; Karch, 2007; Meseguer and Gilardi, 2009; Berry and Berry, 2007).

Despite increased attention to understandings of how innovations spread from state to state, researchers have struggled to develop tests to isolate the four theorized mechanisms of diffusion. When they do test mechanisms, researchers rarely agree on appropriate indicators to distinguish between them. A recent survey of the literature found that the ‘same mechanisms are operationalized using different indicators, and different mechanisms are operationalized using the same indicators’ (Maggetti and Gilardi, 2016: 3). Given this, researchers have effectively focused on predicting state innovativeness rather than on specific mechanisms that may cause states to adopt policies sooner (Gilardi, 2016). Four primary theoretical mechanisms have been identified: imitation, learning, coercion and competition (Simmons et al., 2006; Shipan and Volden, 2008).

The most clearly identifiable mechanism is coercion. Coercion occurs when an actor compels another to adopt a policy (Simmons et al., 2006; Berry and Berry, 2007). In international

relations, more powerful countries could pressure less powerful ones into implementing certain reforms or taking a desired action (Henisz et al., 2005). In the American context this occurs when the federal government pushes states to adopt policies, either by using a combination of carrots and sticks (often in the form of federal funds) or via Supreme Court rulings. Diffusion through coercion, unlike other mechanisms, is a vertical relationship in the US context,<sup>1</sup> which makes it easier to observe. The federal government has a clear position above the states when coercion is used. States are horizontal actors operating at the same level within the government, so none has the authority to legally force another to take an action. A similar dynamic has been studied in which US states coerce cities to adopt preferred policies, or refrain from adopting a policy the state opposes (Shipan and Volden, 2006).

The other three mechanisms mostly focus on the horizontal spread of policies across political actors, meaning that adoptions spread from actor to actor at the same level of government. Within a country this could mean spreading from state to state or city to city; in international relations it means country to country diffusion. Learning occurs when a state observes the policy outcome(s) in another state, and uses that information to make a decision on policy adoption. If that policy produced an outcome that a state desires (e.g., reduced crime), then it learns from the policies and choices of other states that have achieved that outcome and adopts the relevant components. This mechanism often occurs when a state is faced with a new problem, and is looking for potential policy solutions (Walker, 1969). Learning can also incorporate more than just policy outcomes by providing information about a policy's political effects (Graham et al., 2013; Gilardi, 2010). Butler et al. (2017) use an experimental approach to demonstrate that even when given information on a policy's success, municipal policymakers are less likely to support policies that they

ideologically oppose. Policymakers have additional considerations when deciding to adopt a policy. Is the policy politically viable? Does it produce spillover effects or negative externalities?

Learning requires not only that policymakers are aware of what other governments are doing, but that they understand the effect of their actions. The Japanese 'Economic Miracle' provided a model for other developing countries to grow their economies. Other countries, such as South Korea and Chile, successfully learned from the Japanese case by adopting similar economic liberalization policies as Japan (Simmons and Elkins, 2004). Countries that successfully followed Japan's example had to identify which policies led to the desired outcome (increased growth and development).

The next mechanism, imitation, occurs when a state looks to another state for policy ideas, but without looking at the policy outcome (Shipan and Volden, 2008). In this case, actors are adopting policies because of the other state's attributes. Mississippi may adopt a gun law first passed in Alabama not because the policy was a success, but because Mississippi looks at shared conservatism (or another aspect) between the states and follows Alabama's lead even if the policy does not have the intended effect. Similarly, urban states may look to urban states, and wealthy states to wealthy states, for policy ideas. This is done irrespective of policy outcomes. Learning focuses on the policy outcome, whereas imitation uses actor characteristics to inform policy adoption decisions.

Finally, competition occurs when states look to gain an advantage, often economic, over other actors. Incentives to innovate frequently arise from negative externalities fueled by other states' previous adoptions (Baybeck et al., 2011). Tax rates may be lowered, business incentives increased or welfare spending altered to prevent one state from gaining a competitive advantage over the others (Peterson and Rom, 1990). While learning influences decisions to adopt

policies, competition has been found to motivate post-adoption behavior. Boehmke and Witmer (2004) argue that while competition and learning often both matter for initial adoptions, future adjustments to a policy result solely from competition. States have already adopted the policy, so they need not look to other actors for information about outcomes.

Recent work attempts to focus on developing theories and implications about learning-based diffusion. The goal is typically to offer guidance about how learning might produce specific patterns of diffusion and methodological tools that could be used to detect it. One branch of this research typically relies on conditional effects – identifying circumstances in which learning is more likely and then specifying an empirical model to test for the predicted pattern. For example, diffusion by learning might be more likely to influence those that have no experience with a given policy or behavior (Boehmke and Witmer, 2004), or it might be more likely to occur among those with resources to support more careful consideration and exploration of different options (Shipan and Volden, 2008).

Theoretical work has also attempted to clarify the conditions for learning and the patterns that might result from it. These typically create opportunities for learning by making the benefits or outcomes of policies uncertain and partially revealed only by experimenting with policy choices. Volden et al. (2008) develop a formal model of policy adoption that demonstrates an incentive for potential adopters to free-ride on others by waiting to adopt in order to see how the policy plays out in other jurisdictions. This can lead to a delay in adoptions and specific patterns of adoption based on states' underlying predilection toward the policy. A related model from Callander and Harstad (2015) finds that the incentives for free riding can be reduced within a federal structure if states expect the national government to harmonize policy after a period of time. Within the context of a single decision maker, Callander

(2011) examines optimal policy search in an environment with policy uncertainty characterized by Brownian motion. This leads to certain patterns of policy search that feature stickiness and typically end once the results seem 'good enough'. This model has clear applications for policy choices by multiple units. Finally, there have been some attempts to develop empirical estimators corresponding to game-theoretic models of strategic interaction and learning. Building on the development of Quantal Response Equilibrium estimators (Signorino, 2003), Esarey et al. (2008) develop an estimator for incomplete information games that explicitly incorporates learning. Whang et al. (2013) apply a similar estimator to determine that sanctions threats do not lead to the targets updating their beliefs about the sanctioning country's resolve but rather shape the cost-benefit calculation directly.

## EVENT HISTORY ANALYSIS

Walker's (1969) research on policy innovation was quickly met by theoretical and methodological critique from other scholars in the field (Gray, 1973). Further limiting progress in diffusion research was an inability to model both internal characteristics of states (such as legislative professionalism and demographics) and external predictors (contiguous state adopters) of policy diffusion simultaneously. Berry and Berry (1990) introduced event history analysis to the discipline and provided a framework that would allow for modeling both internal and external determinants of policy adoption in the same model.

Event history analysis typically uses a logit (or probit) estimator to model the likelihood of a state adopting a policy in a given year.<sup>2</sup> The logic of the analysis was taken from medical researchers that modeled medical survival data. The 'event' would often be death, and the model would estimate the risk

of an individual dying at different in time. In policy diffusion research, the event is the adoption of a policy, and the risk is the probability of a government adopting a policy in a given year. The risk set is the sample of units that are at risk of the event occurring to them. Once a state adopts a policy, it is no longer at risk of adopting it, and is removed from the analysis.<sup>3</sup> The definition of the event can vary from state policy adoption to signing international treaties to engaging in military conflict, and is signified by a binary variable that is 1 in the time period the policy is adopted, and 0 when it is at risk of adoption but has not done so.

Scholars must consider which units qualify to be part of the risk set, how to measure the event and when units begin to be at risk. For diffusion scholars, the risk set typically begins once any actor in the risk set experiences the event. For example, if a scholar studied the diffusion of the ballot initiative process in US states (Smith and Fridkin, 2008), the risk set would initially include all 50 states. The risk would begin once the first state, South Dakota, adopts the initiative process in 1898. In 1899 the other 49 states are still at risk of adopting the policy, whereas South Dakota is not, because it already experienced the event. Each time a state adopts the initiative, it drops out of the risk set in the subsequent year. Observations that have not adopted the policy by the time data collection ends are considered right censored – they could still adopt that policy but have not done so yet.<sup>4</sup>

The event does not have to be a policy adoption, and event history analysis has been used in a variety of studies. Lektzian and Souva (2001) estimate an event history model for when a country returns to pre-sanction levels of international trade after being sanctioned by the international community. A country is at risk once sanctions are put in place and the event occurs when trade reaches its pre-sanction level. After the event occurs a country drops out of the risk set. Countries that had not yet reached

pre-sanction levels of trade when data collection finished would be considered right censored. Note that while the concept being measured is not binary (amount of trade), the event itself is operationalized as a discrete event that either occurs or does not occur in a discrete time period.

Event history analysis has long been the workhorse method for studies of diffusion in political science and has contributed greatly to how scholars understand policy diffusion (Berry and Berry, 1990; Mintrom, 1997; Balla, 2001). As Gilardi (2016) summarizes, this literature has established that diffusion occurs in a wide variety of phenomena, both within and across countries. The challenge, as he sees it (and we agree), is to move beyond cataloguing the presence of diffusion to understanding the mechanisms behind it. The policy diffusion literature, especially in the American states, has typically accounted for diffusion by including a count of prior adoptions by contiguous states (e.g. Berry and Berry, 1990). Yet a variety of mechanisms could drive findings that find support for this effect. Given data limitations on what the relevant actors know and when, scholars have often turned to identifying conditional diffusion effects to help sort out the mechanisms. For example, Shipan and Volden (2006) find that states with greater legislative professionalism – and therefore greater capacity to learn – respond to the adoption of anti-smoking laws by cities within their borders with state-level adoptions, whereas those with lower levels do not.

In addition to challenges in sorting out mechanisms, EHA presents several other drawbacks. While early studies provided important theoretical contributions to the literature, single-policy studies are increasingly providing diminishing returns (Boehmke 2009a). Scholars hoping to produce generalizable findings about broad diffusion dynamics struggle to aggregate findings from multiple event history analyses. Finally, an EHA model only somewhat acknowledges interdependence among units by focusing

on exogenous and time-lagged outcomes in other states, while diffusion scholars are explicitly arguing that actions in one unit affect actions in another. As the study of diffusion has developed and grown, a number of new methods have been used to overcome some of the shortcomings of single-policy event history analyses.

### POOLED EVENT HISTORY ANALYSIS

Pooled event history analysis (PEHA) was proposed as a way to apply the logic of single-policy EHA to multiple policies (or multiple components of the same policy) in a unified model (Shipan and Volden, 2006; Boehmke, 2009a). PEHA seeks to address the problem of interpreting a possibly wide range of findings from single-policy EHAs by estimating a single model. This allows it to identify systematic commonalities in the determinants of diffusion and innovation across policies and reduce the idiosyncrasies of each individual policy. Analyses with a large number of policies incorporate dramatically more information, which can offer a more accurate estimate of the average effect of a variable. The unit of analysis for PEHA shifts from unit-year, as in EHA, to policy-unit-year. Each policy has a separate risk set that begins when the first unit adopts a policy, just as in single-policy EHA. The difference is that a pooled EHA stacks each policy's risk set into a single model (Makse and Volden, 2011; Shipan and Volden, 2006). The dependent variable,  $Y_{ikt}$ , is still a binary indicator, coded 1 if unit  $i$  adopts policy  $k$  in time period  $t$  and 0 when it is at risk of adopting policy  $k$  but does not adopt.

This approach can identify common diffusion trends across policies, but makes the assumption of fairly homogeneous diffusion pathways across policies. Single-policy analyses have demonstrated significant heterogeneity in diffusion pathways by policies. For example, morality policies may have distinct

patterns of diffusion from non-morality policies (Mooney and Lee, 1995) while crime and law policies may diffuse differently from education policies. Republican control of a state legislature likely predicts diffusion positively for some policies, but negatively for others. As policy databases become increasingly large (Boushey, 2010; Kreitzer, 2015; Makse and Volden, 2011; Boehmke et al., 2018), the likelihood of heterogeneity among diffusion pathways increases. If scholars fail to account for this heterogeneity, then conclusions will simply be a function of the sample of policies chosen for analysis, which is similar to the problem posed by single-policy case studies. If the coefficients differ across policies, then both the coefficient estimates and standard errors will be incorrect.

To overcome this obstacle, Kreitzer and Boehmke (2016) propose running PEHAs as a multilevel model using random intercepts and random coefficients across policies or units to model heterogeneity in the baseline rates of adoption and heterogeneous effects of variables by policy. This form of PEHA seeks to find a middle ground between running separate single-policy EHA models and a single pooled model with all policies treated identically. Scholars must decide which variables merit random (or fixed) effects. Decisions should be guided by the nature of the data being used and the covariates included in the model. These models can be computationally intense, particularly when using a large number of random coefficients. When properly constructed, a multilevel PEHA acknowledges heterogeneity in diffusion pathways while still identifying common, generalizable trends across units and policies.

PEHA's ability to capture both common effects and heterogeneity can be of value for identifying whether mechanisms of diffusion, including learning, have consistent effects in a collection of policies or just within a subset of policies in the collection. That subset may be identified by the researcher and captured in the model via interactions with unit-level covariates (Boushey, 2016; Makse and

Volden, 2011; Shipan and Volden, 2006), or it may be left unspecified and detected via the estimated distribution of a random coefficient across policies, some values of which may indicate the presence of learning while others show no such effect.

Finally, although a PEHA can identify commonalities among diffusion pathways, it still can only partially recognize interdependence among observations. Monadic analysis assumes independence among observations, yet diffusion implies interdependence. Many researchers include variables to capture adoptions in other units, most often a lagged count of the number of contiguous units that have adopted the policy. Both the single-policy and pooled EHA can include external and internal determinants of policy adoption, but the analysis still seeks to explain policy adoption or event occurrence more generally. The methods we outline next represent attempts to directly model diffusion pathways at the level at which they occur.

## DYADIC EVENT HISTORY ANALYSIS

A critical shortcoming of the previous approaches is that they fail to capture the specific sources of learning. Theories of learning and diffusion examine how one unit learns from other units, but it does not necessarily learn from all other units equally. For example, a country can learn from the experiences of countries that have adopted a policy and from those of countries that have not adopted it. It may learn more from the adoption decisions of countries that have similar demographic and political environments or that face a similar scope of the problem the policy seeks to address. Learning therefore typically occurs from one unit to another, and often from multiple units in different ways. Single-policy and pooled EHA models struggle to capture the different sources of learning since they lump the influence of other units into a single variable or series of variables

(though they may be weighted according to their likely influence). They can therefore capture whether the choices of other countries influence the outcome in a single country in the aggregate, but provide little leverage on differential impacts needed to identify sources of learning and to help distinguish them from other mechanisms.

To help address this problem, diffusion scholars have often turned from monadic to dyadic structures. These have been applied extensively in the study of international conflict (Maoz and Russett, 1993b; Leeds, 2003; Danilovic and Clare, 2007) but have more recently been adapted to the diffusion framework to explicitly model the spread of policies or behaviors through the development of dyadic event history analysis. Rather than a single country (or state, city, etc.) being the unit of analysis, a dyadic approach evaluates dyads, or pairs of units. This approach allows for modeling sender and receiver states in order to evaluate the direction of the diffusion event flow (Volden, 2006). In international relations applications, the dependent variable often measures the probability of two states going to war (Maoz and Russett, 1993a; Leeds, 2003) or signing bilateral trade agreements (Elkins et al., 2006). In the case of policy diffusion, the dependent variable shifts from 'policy adoption' to 'dyadic policy similarity' where a receiver unit is either adopting a policy already present in the other unit in the dyad (Gilardi and Füglistler, 2008) or moving its policy closer, in a possibly multidimensional space, to that in the other unit (Volden, 2006).

Importantly, dyadic EHA is directional. Each dyad consists of a sender and a receiver unit. The dyad enters into the risk set once the sender has adopted the policy (see Gilardi and Füglistler (2008) for information on how to format the data). This method allows direct modeling of diffusion pathways (Boehmke, 2009b), meaning scholars using this approach are modeling diffusion itself rather than diffusion as a part of a broader policy adoption or innovation process. Hinkle (2015) uses a dyadic logit to identify signals of policy

success or failure (in this case, being ruled unconstitutional in the courts) which states consider when learning about policy solutions. If scholars can identify cases of successful policy outcomes, a dyadic approach can allow for modeling the extent to which actors are learning from previous adoptions.

While dyadic analysis offers several advantages, scholars have found theoretical and methodological issues with this approach (e.g., Erikson et al., 2014). As noted above, the dependent variable shifts from being the occurrence of the event of interest to the convergence of one unit's outcome with another unit's existing choices. Thus information on why some units tend to experience the event earlier than others is lost in favor of explaining the choices of later adopters vis-à-vis the choices of previous adopters. This may be helpful when the focus is explicitly on learning or other mechanisms but researchers need to be cognizant of the change in interpretation. For example, Boehmke (2009b) demonstrates the need to remove observations that already have the outcome in question since they are not at risk of learning in many applications of dyadic EHA. Including them risks conflating factors that influence adoption of the underlying policy with those that influence convergence between two units.

Dyadic EHA has not yet addressed the more challenging limitation that while it can capture policy convergence between pairs of states, it cannot differentiate from among a host of potential sources. That is, the dependent variable captures whether a state moves toward a set of states but its movement toward all units with the same current policy is coded the same way. Conceptually, researchers must rely on richer measures of convergence or the inclusion of lots of data to try to tease out which of the units to which a state converges matter and which may be coincidences, e.g., since they may all be converging to a single leader state at different points in time. Statistically, this creates problems since it induces correlation between the error terms as a state must be treated as converging

to all states with similar outcomes, even if it is only converging to a subset of them. The dyadic approach therefore explicitly models some interdependence among observations, but misses and can even create other forms. Compared to a monadic analysis, dyadic models may misidentify independent observations as interdependent, and thus underestimate the effect of internal characteristics on policy adoption and overstate the role of external characteristics. Finally, researchers should also address why dyads are the proper unit of analysis, and not triads, or, as we discuss in the next section, even a full network.

The dyadic approach's ability to include characteristics of both units in a pair means that it can account for features of states seeking to learn, those they might learn from, and relative features of the two (such as similarity). While this is helpful for studying all mechanisms of diffusion, it proves particularly valuable for studying learning since it offers an opportunity to account for the performance of a policy in units that already have it. The dyadic approach therefore offers great leverage for establishing that learning occurs when such measures are available. For example, Volden (2006) analyzes convergence in US states' children's health insurance programs over six dimensions and finds that states were considerably more likely to revise their policy to more closely mirror those in states that had successfully reduced their uninsured rate among poor children. In contrast to this evidence of policy learning, Gilardi (2010) finds evidence of political learning in his study of unemployment benefits in OECD countries: right wing governments are more likely to move their policies toward those of countries in which reforms have produced electoral benefits.

## **DIFFUSION NETWORKS**

Another way for scholars to incorporate interdependence among observations is to use network analysis to model event

pathways. This approach allows for a more comprehensive incorporation of interdependence than a dyadic approach by incorporating higher ordered network processes such as transitivity (Valente, 1995). Public policy research has used network analysis for decades to understand how policies spread (Coleman and Perl, 1999; Klijn, 1996; Koppenjan and Klijn, 2004; Thatcher, 1998), and international relations research has used networks to explain topics ranging from the structure of alliance networks (Cranmer et al., 2012) to international trade flows (Hafner-Burton et al., 2009). The basic assumption behind a network approach is that the behavior of an actor, e.g., the adoption of a policy, affects the behavior of other actors in the group. Network scholars argue that events flow through a network of actors. Countries make decisions to ratify a treaty or go to war based on their own characteristics and in response to other countries. Network analysis better approximates the theoretical flow of policies (Lubell et al., 2012). When the United States makes a decision to sign an international treaty, it is likely influenced by whether allies (or competitors) have signed the treaty, and its decision to sign (or not sign) likely influences other nations.

The shift to networks now means that the focus of analysis is how an event spreads, not whether individual actors or dyads behave in a particular way. In the context of policy diffusion, a tie forms between two nodes when they adopt the same policy. So in a network of US states, after the first state adopts a policy the network would have no ties because the policy has not spread. As the policy spreads to another state, a directed tie forms from the first adopter to the second, and the network continues to build as policies go from receiver to sender states. However, scholars may struggle to identify the source of policy ties. If Florida adopts a policy previously adopted by Indiana and Wyoming, should Florida receive a tie from Indiana, Wyoming, or both states? Garrett and Jansa (2015) use bill text to identify the source of a policy and

where it spreads. Their analysis reveals that interest groups can act as a policy resource. Researchers must decide what a tie means in the network, and how to determine the source of diffusing policies. A network of directed ties better mirrors the diffusion theories that policies spread from one unit to another, but there may be cases where an undirected network is the appropriate choice. Mutual defense treaties or free trade agreements imply a reciprocal relationship. Those using networks must decide how to define a tie in the network and whether it should be directed or undirected.

Network analysis presents other advantages beyond better mirroring the structure of diffusion theories. Networks have been used to identify where policies originate and how they spread through the network (Garrett and Jansa, 2015). A variety of network statistics can be used to evaluate which units are the most or least central to a unit. For example, when two countries share a similar network position (measured by structural equivalence), they are likely to compete with each other (Cao, 2010). Unless these two units were directly connected, a non-network approach would be unable to observe this higher-order relationship between actors. For example, during the Cold War the USSR and United States were the two poles of the international alliance network. They responded to each other throughout the Cold War, but did not have direct alliance ties to each other. With a network approach, unlike monadic or dyadic analyses, scholars can model how the United States deciding to join an alliance affects the alliance network for the USSR (Chyzh and Kaiser, n.d.). Network processes and statistics can provide insights into diffusion pathways that would otherwise go unobserved in a monadic analysis (Robins et al., 2012).

Beyond network statistics, estimators such as QAP or ERGMs can be used to determine the correlates of ties between the states (Cranmer and Desmarais, 2010). These models can be very powerful for understanding



policy diffusion because they use a combination of network structure and edge-level attributes (such as population differences or trade between units) to model diffusion (Robins et al., 2012). This approach directly mirrors theories of diffusion that argue that both external and internal characteristics contribute to diffusion and policy adoption. Notably, ERGMs include the dyadic logit as a special case in which there are no network-level effects. This analysis takes a similar approach of a dyadic logit but adds network dynamics to the model of nodal, edge and system level variables. These network effects can potentially be used to distinguish between learning, emulation and competition (Maggetti and Gilardi, 2016).

ERGMs allow researchers to include features of the network structure as part of the explanatory model. For example, Thurner and Binder (2009) use an ERGM to understand how the structure of the European Union affects the network connections between high ranking bureaucrats in member states. They test if the institutionalization of the European Union replaced existing networks of communication between nation states. The ERGM they estimate includes both network statistics (reciprocity) and edge-level covariates (economic interdependence) to predict communication ties between policymakers. They find that the structural components of the network and the edge-level covariates have a significant relationship with the existence of communication ties between policymakers. ERGMs allow researchers to leverage both network structure variables and edge covariates to understand how actor characteristics and the structure in which they are nested affect diffusion pathways.

Diffusion can also be viewed as occurring within an existing network structure. Examples of this abound for exogenous networks such as contiguity, ideological similarity or trade. In monadic models, researchers often include the network-weighted sum of policies or behaviors in other states to explain occurrence in the current state. In

dyadic models such pairwise features may be included as dyadic covariates. But in some cases one might worry about coevolution of the behavior of interest and the transmission network. For example, Chyzh (2016) argues that states' human rights policies depend on their position in the international trade network, but also contends that position in the network depends on human rights policies since states often forego trade with countries whose protections they deem insufficient. To address this, Chyzh (2016) employs a coevolutionary actor-oriented longitudinal-network model (see, e.g., Steglich et al., 2010). This estimator, referred to as RSiena, models the network connections simultaneously with a behavioral outcome, such as human rights policies, at the nodal level within that network. Both equations can include features of the network. The results indicate that states that rely more on indirect trade links (trading through mutual partners rather than directly) tend to score lower on human rights and that states that score lower on human rights have few direct trade connections to other states. Thus the network shapes policy, but policy also shapes the network.

While network models have been used to study diffusion generally, to our knowledge no published research has used an ERGM to identify learning in a policy diffusion network. However, researchers could take a few approaches to identify learning in the network. Many of these will mirror the options for a dyadic logit since they share the same underlying structure of modeling links between units, but an ERGM allows researchers to study additional features related to network structure. Researchers could evaluate how signals of success or failure alter the diffusion network, e.g., do states with successful policy outcomes take a more central role in a diffusion network? Do we see a greater level of transitivity when the policy is deemed a success? Are isolates (states that do not have diffusion ties to any state) less common when a policy's success is clear? Additionally, different diffusion mechanisms imply different

network connections. Competition suggests reciprocity between nodes as they act and react to the other's behavior. Learning, on the other hand, should mostly be a uni-directional relationship because actors are responding to the policy success of a previous adoption. Researchers can include network statistics to identify these types of connections between actors. We believe a network approach could be a fruitful avenue for identifying learning in diffusion.

One of the drawbacks of the network approach in general centers on how to pool results. Almost every policy will have a unique adoption network when considering both how and when policies spread. Different policy areas may also have different policy leaders. This presents a similar dilemma to the single-policy event history analyses discussed in an earlier section. Different networks will likely result in different conclusions. Without a systematic way to aggregate findings, scholars must assume that the chosen diffusion network is representative of general trends in the diffusion network if they wish to make generalizable claims about diffusion networks.

## LATENT NETWORKS

More recently, network scholars have begun to use latent network analysis to examine diffusion pathways. Rather than studying observed diffusion networks, this approach utilizes data from a large number of such networks to estimate a single, underlying latent network that contains the ties that best explain the observed diffusion patterns. A latent network approach operates in a similar way to latent factor analysis by using observed policy cascades to infer an optimized network of ties between units (Gomez-Rodriguez et al., 2010). The network is constructed using an algorithm that infers diffusion ties based on the number of cascades in which state  $i$  adopts before state  $j$ ,

the length of time between these adoptions and how well state  $i$  adoption predicts state  $j$ 's adoption as opposed to that of other states that tend to adopt before  $j$ . Latent networks are not directly observed, but represent the most likely network of diffusion ties given the policy adoption networks used to infer the network. Rather than predicting the adoption of a single policy, researchers estimate latent diffusion ties, which can then be analyzed using appropriate models, such as dyadic logit, QAP logit or an ERGM.

Desmarais et al. (2015) apply this algorithm to a sample of more than 100 policy adoption cascades in the American states to recover a latent diffusion network. The ties represent the most likely diffusion connections between states. Their results reveal a network that evolves over time based on a 40-year rolling window of adoption data. With 100 years' worth of data, this produces 60 estimated networks. This means that scholars can estimate and then model how diffusion networks evolve over time. To study this evolution, Desmarais et al. (2015) employ a QAP logit model to find that directed ties in the latent state policy diffusion network depend on theoretically relevant dyadic features such as geographic distance, ideological similarity and political similarity. They also find that more populous states send out more ties and receive more ties.

Because the networks cannot be directly observed, scholars must carefully consider how they infer a network, including what observed diffusion events to use, how quickly the influence of an adoption should decay over time, how dense the network should be and what the appropriate universe of cases is when deciding nodes in the network. Each of these decisions will determine what the network looks like. The network will be less dense if sparse policy adoption networks are used to construct the latent network. Additionally, the longer the window of time for previous policy adoptions to influence the current diffusion network, the denser the

network. When constructed appropriately, latent networks are representative of the general diffusion network in a given era. Even though there is still only a single network produced for a given time period, scholars can feel more comfortable that the network is generalizable. Additionally, latent networks represent more than just observed adoption, but also the flow of information and other factors that predict diffusion between units.

Latent networks allow for a variety of diffusion studies. Network descriptives can be used to identify the leader and follower states, as well as the most central actors in a diffusion network. The network can be the dependent variable, where scholars analyze the determinants of ties between units in a network. This approach allows for examining the competing roles of state characteristics and network forces such as triadic closure or in-degree in the same model. Alternatively, these networks can be utilized in event history analyses as a measure of the external diffusion influences on policy adoption. In much the same way that adoptions by contiguous neighbors or ideologically similar states increase the chance of adoption in a given state – whether in a monadic, dyadic or network analysis – past adoptions by latent sources likewise predict adoption (Boehmke et al., 2017). Once the network has been produced, scholars can proceed using the same types of network analysis that they would use for other types of diffusion networks, including ERGMs, QAP and other model specifications where the latent network is used as either an independent or dependent variable to understand the structure of ties between states. Notably, in contrast to the dyadic EHA models discussed earlier, the latent network approach facilitates analysis of specific links since it leverages the spread of many policies to determine the presence of a diffusion tie between all pairs of states.

Political scientists have just begun to explore the ability to estimate latent networks for studying policy diffusion. An important next step will be using the estimated

networks to help study the mechanisms of diffusion, including learning. While they require a considerable amount of data and optimization, they also offer some advantages. Most importantly, they offer a direct estimate of a tendency for diffusion to occur between all pairs of units, possibly varying over time. Rather than relying on the information contained in the adoption of a single or small number of policies, researchers can obtain direct estimates of the item of interest: policy ties between units across a large number. These ties can then be modeled with variables intended to more directly capture the mechanisms of diffusion.

## SPATIAL ECONOMETRIC MODELS

Spatial econometrics offers a potential middle ground between dyadic analysis of ties in the diffusion network and monadic analysis of the choices by individual units. It allows the researcher to capture a variety of dependencies between units, both endogenous and exogenous. It does so by requiring the researcher to specify a spatial dependency matrix indicating how each unit connects to all other units. This matrix is essentially a representation of a continuous-valued network. In fact, most existing diffusion studies already use a version of spatial econometrics via the inclusion of a lagged count of adoptions in contiguous units or related measures. The lagged count comes from the multiplication of the contiguity matrix for all units by a vector capturing the presence of the policy in every unit. Since the presence of the policy variable is usually lagged, this is just a case of spatial regression with an exogenous lag.

Spatial econometric methods offer much more than this, however. They can accommodate any matrix of connections between units, whether geographically based or not. This feature proves critical for studying diffusion mechanisms since many of them are

not based on notions of geography: fashion may diffuse through friend networks or via social media ties; policies may diffuse via ideological or problem similarity; or conflict may spread through terrorist networks or ethnic groups that straddle international borders. One can include a sum or weighted average of any feature of other units as an exogenous influence by specifying the spatial weights matrix (Neumayer and Plümper, 2016); more than one such feature can be included via multiple weights matrices.

Even more powerfully – and much less commonly utilized in the study of diffusion – spatial econometric models permit capturing endogenous dependencies via these spatial weights matrices. Rather than include the weighted value of an exogenous covariate, one can include the weighted value of the error terms or, even better, of the outcome variable. Conceptually, the latter means that spatial autoregressive (SAR) and spatio-temporal autoregressive (STAR) models can capture the simultaneous way in which the outcome in one unit explicitly depends on the outcomes in other units and within the same unit over time (Franzese Jr and Hays, 2007). This makes them valuable for studying certain forms of diffusion, including those based on contemporaneous learning or strategic interactions.

An early application of STAR models to policy diffusion concerns the question of welfare benefits in the American states. Longstanding concerns about a race to the bottom in which states work to keep their benefits levels below those of nearby states to thereby avoid attracting too many potential recipients make this a strong candidate for spatial analysis. Rom et al. (1998) conduct just such an analysis with contiguity as their spatial weights matrix and find evidence of positive spatial correlation: an increase of \$100 in benefits per person in a state leads to a contemporaneous increase of \$27 in its neighbor (which is then compounded over time and across space). A useful comparison can be made with Volden

(2002), which conducts an EHA model for a binary measure of benefit increases and accounts for changes in neighboring states via a time-lagged exogenous variable, reaching similar conclusions. In contrast to these findings, Franzese Jr and Hays (2006) find evidence of free riding in European countries' support for labor market policy, with a negative spatial correlation that produces a drop in domestic spending when neighboring states increase their spending.

Spatial autocorrelation arises from a number of possible mechanisms (Franzese Jr and Hays, 2007): interdependence, unmodeled heterogeneity (in the form of spatially correlated random shocks or omitted variables) and selection (e.g., via homophily in the connectivity matrix). Interdependence includes the common forms of diffusion such as learning, emulation, competition and coercion. As with the other models we have discussed, it is often difficult to determine which mechanism undergirds a finding of spatial correlation (though see Mitchell (2018) for a recent proposal for how to do so). As with the EHA model and its variants, scholars often turn to conditional interactive effects to identify learning. For example, Arel-Bundock and Parinandi (2018) study tax competition in the American states and find that corporate tax policy in states with better resourced (and therefore more able to learn) legislatures more closely tracks policy in connected states.

## MOVING FORWARD

Each of the methodological approaches outlined in this chapter has its trade-offs for studying policy diffusion and innovation. There has been tremendous growth in the diffusion literature over the past 30 years, particularly after Berry and Berry (1990) introduced EHA to political scientists and noted its strengths in incorporating internal and external characteristics to determine

when events occur. This offered a way to test for the influence of diffusion mechanisms while accounting for unit-level differences in the probability of an event occurring. As the field has moved forward greater emphasis has been placed not just on understanding whether diffusion occurs, but on testing and identifying the role of specific mechanisms.

These demands have pushed researchers to develop and apply new empirical methods for studying diffusion. EHA was introduced nearly 30 years ago, but none of the other methods discussed here was used much, if at all, just over a decade ago. This new menu of estimators offers diffusion scholars a range of options for identifying and testing for these mechanisms. As we see it, they fall broadly into two groups. The first, including PEHA, dyadic EHA and latent network estimation, provides opportunities to leverage information from large data sets to identify common and possibly small diffusion effects while allowing for heterogeneity across events. The second group, including ERGMs, RSiena and spatial econometrics, provides estimators designed to explicitly capture endogenous interdependencies in diffusion networks.

On top of these differences in their orientation towards data and diffusion processes, each of these methods has its own strengths and weaknesses. The choice of which to use may therefore depend on the questions asked or the types of analyses needed. And as many of them have only relatively recently been applied to studying diffusion, work remains to be done to more fully determine their strengths and weaknesses and adapt them to diffusion applications to maximize the former while minimizing the latter. Ultimately the ability to identify the presence of a distinct mechanism of diffusion, such as learning as opposed to emulation, requires careful thinking about how theoretical concepts map into measures and which methods provide the most appropriate features for estimating them.

## Notes

- 1 While horizontal coercion is common in international politics (countries using their economic or military clout to force others to comply), the constitutional context of the United States makes horizontal coercion rare.
- 2 Continuous-time duration models like the Weibull or Cox may also be employed (Jones and Branton, 2005).
- 3 For some policies, re-adoption of a policy is possible, but generally researchers remove an actor once the event, however defined, occurs.
- 4 See Box-Steffensmeier and Jones (2004) for a guide to the application and estimation of EHA.

## REFERENCES

- Arel-Bundock, Vincent and Srinivas Parinandi. 2018. 'Conditional tax competition in American states.' *Journal of Public Policy* 38(2):191–220.
- Balla, Steven J. 2001. 'Interstate professional associations and the diffusion of policy innovations.' *American Politics Research* 29(3):221–245.
- Baybeck, Brady, William D. Berry and David A. Siegel. 2011. 'A strategic theory of policy diffusion via intergovernmental competition.' *The Journal of Politics* 73(1):232–247.
- Berry, Frances Stokes and William D. Berry. 1990. 'State lottery adoptions as policy innovations: an event history analysis.' *The American Political Science Review* 84(2):395–415.
- Berry, Frances Stokes and William D. Berry. 2007. 'Innovation and diffusion models in policy research.' In *Theories of the Policy Process*, ed. Paul A. Sabatier. Second ed. Westview pp. 223–260.
- Boehmke, Frederick J. 2009a. 'Approaches to modeling the adoption and modification of policies with multiple components.' *State Politics and Policy Quarterly* 9(2):229–252.
- Boehmke, Frederick J. 2009b. 'Policy emulation or policy convergence? Potential ambiguities in the dyadic event history approach to state policy emulation.' *The Journal of Politics* 71(3):1125–1140.
- Boehmke, Frederick J., Abigail Matthews Rury, Bruce A. Desmarais and Jeffrey J. Harden. 2017. 'The seeds of policy change: leveraging diffusion to disseminate policy

- innovations.' *Journal of Health Politics, Policy and Law* 42(2):285–307.
- Boehmke, Frederick J., Mark Brockway, Bruce Desmarais, Jeffrey J. Harden, Scott LaCombe, Fridolin Linder and Hanna Wallach. 2018. 'State Policy Innovation and Diffusion (SPID) Database v1.0.'
- Boehmke, Frederick J. and Paul Skinner. 2012. 'State policy innovativeness revisited.' *State Politics and Policy Quarterly* 12(3):303–329.
- Boehmke, Frederick J. and Richard Witmer. 2004. 'Disentangling diffusion: the effects of social learning and economic competition on state policy innovation and expansion.' *Political Research Quarterly* 57(1):39–51.
- Boushey, Graeme. 2010. *Policy Diffusion Dynamics in America*. New York City: Cambridge University Press.
- Boushey, Graeme. 2016. 'Targeted for diffusion? How the use and acceptance of stereotypes shape the diffusion of criminal justice policy innovations in the American states.' *American Political Science Review* 110(1):198–214.
- Box-Steffensmeier, Janet M. and Bradford S. Jones. 2004. *Event History Modeling: A Guide for Social Scientists*. Cambridge University Press.
- Butler, Daniel M., Craig Volden, Adam M. Dynes and Boris Shor. 2017. 'Ideology, learning, and policy diffusion: experimental evidence.' *American Journal of Political Science* 61(1):37–49.
- Callander, Steven. 2011. 'Searching for good policies.' *American Political Science Review* 105(4):643–662.
- Callander, Steven and Bård Harstad. 2015. 'Experimentation in federal systems.' *The Quarterly Journal of Economics* 130(2):951–1002.
- Cao, Xun. 2010. 'Networks as channels of policy diffusion: explaining worldwide changes in capital taxation, 1998–2006.' *International Studies Quarterly* 54(3):823–854.
- Chyzh, Olga. 2016. 'Dangerous liaisons: an endogenous model of international trade and human rights.' *Journal of Peace Research* 53(3):409–423.
- Chyzh, Olga V. and Mark S. Kaiser. (n.d.). 'A local structure graph model: modeling formation of network edges as a function of other edges.' *Political Analysis* 27(4), 397–414.
- Coleman, William D. and Anthony Perl. 1999. 'Internationalized policy environments and policy network analysis.' *Political Studies* 47(4):691–709.
- Cranmer, Skyler J. and Bruce A. Desmarais. 2010. 'Inferential network analysis with exponential random graph models.' *Political Analysis* 19(1):66–86.
- Cranmer, Skyler J., Bruce A. Desmarais and Elizabeth J. Menninga. 2012. 'Complex dependencies in the alliance network.' *Conflict Management and Peace Science* 29(3):279–313.
- Danilovic, Vesna and Joe Clare. 2007. 'The Kantian liberal peace (revisited).' *American Journal of Political Science* 51(2):397–414.
- Desmarais, Bruce A., Jeffrey J. Harden and Frederick J. Boehmke. 2015. 'Persistent policy pathways: inferring diffusion networks in the American states.' *American Political Science Review* 109(2):392–406.
- Elkins, Zachary, Andrew T. Guzman and Beth A. Simmons. 2006. 'Competing for capital: the diffusion of bilateral investment treaties, 1960–2000.' *International Organization* 60(4):811–846.
- Erikson, Robert S., Pablo M. Pinto and Kelly T. Rader. 2014. 'Dyadic analysis in international relations: a cautionary tale.' *Political Analysis* 22(4):457–463.
- Esarey, Justin, Bumba Mukherjee and Will H. Moore. 2008. 'Strategic interaction and interstate crises: a Bayesian quantal response estimator for incomplete information games.' *Political Analysis* 16(3):250–273.
- Franzese Jr, Robert J. and Jude C. Hays. 2006. 'Strategic interaction among EU governments in active labor market policy-making: subsidiarity and policy coordination under the European employment strategy.' *European Union Politics* 7(2):167–189.
- Franzese Jr, Robert J. and Jude C. Hays. 2007. 'Spatial econometric models of cross-sectional interdependence in political science panel and time-series-cross-section data.' *Political Analysis* 15(2):140–164.
- Garrett, Kristin N. and Joshua M. Jansa. 2015. 'Interest group influence in policy diffusion networks.' *State Politics and Policy Quarterly* 15(3):387–417.

- Gilardi, Fabrizio. 2010. 'Who learns from what in policy diffusion processes?' *American Journal of Political Science* 54(3):650–666.
- Gilardi, Fabrizio. 2012. 'Transnational diffusion: norms, ideas, and policies.' In *Handbook of International Relations 2*. Sage pp. 453–477.
- Gilardi, Fabrizio. 2016. 'Four ways we can improve policy diffusion research.' *State Politics and Policy Quarterly* 16(1):8–21.
- Gilardi, Fabrizio and Katharina Füglistner. 2008. 'Empirical modeling of policy diffusion in federal states: the dyadic approach.' *Swiss Political Science Review* 14(3):413–450.
- Gleditsch, Kristian Skrede. 2009. *All International Politics Is Local: The Diffusion of Conflict, Integration, and Democratization*. University of Michigan Press.
- Gomez-Rodriguez, Manuel, Jure Leskovec and Andreas Krause. 2010. Inferring Networks of Diffusion and Influence. In *The 16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*.
- Graham, Erin R., Charles R. Shipan and Craig Volden. 2013. 'The diffusion of policy diffusion research in political science.' *British Journal of Political Science* 43(3): 673–701.
- Gray, Virginia. 1973. 'Innovation in the states: a diffusion study.' *The American Political Science Review* 67(4):1174–1185.
- Hafner-Burton, Emilie M., Miles Kahler and Alexander H. Montgomery. 2009. 'Network analysis for international relations.' *International Organization* 63(3):559–592.
- Henisz, Witold J., Bennet A. Zelner and Mauro F. Guillén. 2005. 'The worldwide diffusion of market-oriented infrastructure reform, 1977–1999.' *American Sociological Review* 70(6):871–897.
- Hinkle, Rachael K. 2015. 'Into the words: using statutory text to explore the impact of federal courts on state policy diffusion.' *American Journal of Political Science* 59(4):1002–1021.
- Jensen, Nathan M. and René Lindstädt. 2012. 'Leaning right and learning from the left: Diffusion of corporate tax policy across borders.' *Comparative Political Studies* 45(3):283–311.
- Jones, Bradford S. and Regina P. Branton. 2005. 'Beyond logit and probit: Cox duration models of single, repeating, and competing events for state policy adoption.' *State Politics and Policy Quarterly* 5(4):420–443.
- Karch, Andrew. 2007. 'Emerging issues and future directions in state policy diffusion research.' *State Politics and Policy Quarterly* 7(1):54–80.
- Klijin, Erik Hans. 1996. 'Analyzing and managing policy processes in complex networks: a theoretical examination of the concept policy network and its problems.' *Administration and Society* 28(1):90–119.
- Koppenjan, Joop and Erik-Hans Klijin. 2004. *Managing Uncertainties in Networks: Public Private Controversies*. Routledge.
- Kreitzer, Rebecca J. 2015. 'Politics and morality in state abortion policy.' *State Politics and Policy Quarterly* 15(1):41–66.
- Kreitzer, Rebecca J. and Frederick J. Boehmke. 2016. 'Modeling heterogeneity in pooled event history analysis.' *State Politics and Policy Quarterly* 16(1):121–141.
- Lee, Chang Kil and David Strang. 2006. 'The international diffusion of public-sector downsizing: network emulation and theory-driven learning.' *International Organization* 60(4):883–909.
- Leeds, Brett Ashley. 2003. 'Do alliances deter aggression? The influence of military alliances on the initiation of militarized interstate disputes.' *American Journal of Political Science* 47(3):427–439.
- Lektzian, David and Mark Souva. 2001. 'Institutions and international cooperation: An event history analysis of the effects of economic sanctions.' *Journal of Conflict Resolution* 45(1):61–79.
- Lubell, Mark, John Scholz, Ramiro Berardo and Garry Robins. 2012. 'Testing policy theory with statistical models of networks.' *Policy Studies Journal* 40(3):351–374.
- Maggetti, Martino and Fabrizio Gilardi. 2016. 'Problems (and solutions) in the measurement of policy diffusion mechanisms.' *Journal of Public Policy* 36(1):87–107.
- Makse, Todd and Craig Volden. 2011. 'The role of policy attributes in the diffusion of innovations.' *The Journal of Politics* 73(1):108–124.
- Maoz, Zeev and Bruce Russett. 1993a. 'Normative and structural causes of democratic peace, 1946–1986.' *American Political Science Review* 87(3):624–638.

- Maoz, Zeev and Bruce Russett. 1993b. 'Normative and structural causes of democratic peace, 1946–1986.' *The American Political Science Review* 87(3):624–638.
- Meseguer, Covadonga and Fabrizio Gilardi. 2009. 'What is new in the study of policy diffusion?' *Review of International Political Economy* 16(3):527–543.
- Mintrom, Michael. 1997. 'Policy entrepreneurs and the diffusion of innovation.' *American Journal of Political Science* 41(3):738–770.
- Mitchell, Joshua L. 2018. 'Does policy diffusion need space? Spatializing the dynamics of policy diffusion.' *Policy Studies Journal* 46(2):424–451.
- Mooney, Christopher Z. and Mei-Hsien Lee. 1995. 'Legislative morality in the American states: the case of pre-Roe abortion regulation reform.' *American Journal of Political Science* 39(3):599–627.
- Neumayer, Eric and Thomas Plümpner. 2016. 'W.' *Political Science Research and Methods* 4(1):175–193.
- Nicholson-Crotty, Sean. 2009. 'The politics of diffusion: public policy in the American states.' *The Journal of Politics* 71(1):192–205.
- Peterson, Paul E. and Mark C. Rom. 1990. *Welfare Magnets: A New Case for a National Standard*. Brookings Institute Press.
- Robins, Garry, Jenny M. Lewis and Peng Wang. 2012. 'Statistical network analysis for analyzing policy networks.' *Policy Studies Journal* 40(3):375–401.
- Rom, Mark Carl, Paul E. Peterson and Kenneth F. Scheve Jr. 1998. 'Interstate competition and welfare policy.' *Publius: The Journal of Federalism* 28(3):17–37.
- Shipan, Charles R. and Craig Volden. 2006. 'Bottom-up federalism: the diffusion of anti-smoking policies from U.S. cities to states.' *American Journal of Political Science* 50(4):825–843.
- Shipan, Charles R. and Craig Volden. 2008. 'The mechanisms of policy diffusion.' *American Journal of Political Science* 52(4):840–857.
- Signorino, Curtis S. 2003. 'Structure and uncertainty in discrete choice models.' *Political Analysis* 11(4):316–344.
- Simmons, Beth A. and Zachary Elkins. 2004. 'The globalization of liberalization: policy diffusion in the international political economy.' *American Political Science Review* 98(1):171–189.
- Simmons, Beth A., Frank Dobbin and Geoffrey Garrett. 2006. 'Introduction: the international diffusion of liberalism.' *International Organization* 60(4):781–810.
- Smith, Daniel A. and Dustin Fridkin. 2008. 'Delegating direct democracy: interparty legislative competition and the adoption of the initiative in the American states.' *American Political Science Review* 102(3):333–350.
- Starr, Harvey. 1991. 'Democratic dominoes: diffusion approaches to the spread of democracy in the international system.' *Journal of Conflict Resolution* 35(2):356–381.
- Steglich, Christian, Tom A. B. Snijders and Michael Pearson. 2010. 'Dynamic networks and behavior: separating selection from influence.' *Sociological Methodology* 40(1):329–393.
- Thatcher, Mark. 1998. 'The development of policy network analyses: from modest origins to overarching frameworks.' *Journal of Theoretical Politics* 10(4):389–416.
- Turner, Paul W. and Martin Binder. 2009. 'European Union transgovernmental networks: The emergence of a new political space beyond the nation-state?' *European Journal of Political Research* 48(1):80–106.
- Valente, Thomas W. 1995. 'Network models of the diffusion of innovations.' *Computational and Mathematical Organization Theory* 2(2):163–164.
- Volden, Craig. 2002. 'The politics of competitive federalism: a race to the bottom in welfare benefits?' *American Journal of Political Science* 46(2):352–363.
- Volden, Craig. 2006. 'States as policy laboratories: emulating success in the Children's Health Insurance Program.' *American Journal of Political Science* 50(2):294–312.
- Volden, Craig, Michael M. Ting and Daniel P. Carpenter. 2008. 'A formal model of learning and policy diffusion.' *American Political Science Review* 102(3):319–332.
- Walker, Jack L. 1969. 'The diffusion of innovations among the American states.' *American Political Science Review* 63(3):880–899.
- Weyland, Kurt. 2005. 'Theories of policy diffusion: lessons from Latin American pension reform.' *World Politics* 57(2):262–295.



Whang, Taehee, Elena V. McLean and Douglas W. Kuberski. 2013. 'Coercion, information, and the success of sanction threats.' *American Journal of Political Science* 57(1):65–81.

Zhukov, Yuri M. and Brandon M. Stewart. 2013. 'Choosing your neighbors: networks of diffusion in international relations.' *International Studies Quarterly* 57(2):271–287.

PART III

# Conceptualization and Measurement



*This page intentionally left blank*



# Conceptualization and Measurement: Basic Distinctions and Guidelines

Gerardo L. Munck, Jørgen Møller  
and Svend-Erik Skaaning

One of the key aims of the social sciences is to describe the social world. Descriptions are one of the most powerful products of the social sciences. Based on descriptions, countries are ranked as being more or less democratic or respectful of human rights or corrupt; the level of violence over time within and between particular groups is gauged; political parties are compared on a left–right spectrum; citizens are held to have more or less liberal or religious values, and so on. Much of what we know about the social world is due to research that seeks to provide descriptions. In addition, research oriented to offering descriptions provides important input for research that aims at explaining the social world.

In this chapter, we offer an overview of the issues involved in producing the data that are used in descriptions. The overview is divided into three main sections. We begin by focusing on the task of conceptualization. *Concepts* play a fundamental but frequently unappreciated role in the production of data.

We clarify the components of concepts, discuss how concepts can be organized and distinguish among different kinds of conceptual systems. We next turn to measurement, distinguishing between the production of *data on indicators* and *data on indices*. The notions of indicators and indices are sometimes used interchangeably. However, the tasks and choices involved in producing data on indicators and indices, respectively, are distinct and better addressed one at a time. Thus, in the second section we focus on data on indicators, and address the task of selecting indicators, designing measurement scales and collecting data. Subsequently, in the third section, we turn to data on indices, where we develop a key distinction between two kinds of indices – those that combine data on multiple units and those that combine data on multiple indicators measuring different properties in one unit – and discuss key options concerning these two kinds of indices.

Data can be good or poor, and we are also concerned with ensuring that data are of high

quality. Thus, we discuss not only what is involved in *producing* data but also what is involved in *evaluating* descriptions. Ideally, as we suggest, evaluations would feed back into the production of data, but frequently evaluations are carried out as a post-production task. To this end, we discuss various criteria that are relevant to an evaluation of data. However, because this chapter focuses on concepts and the link between concepts and measures, and does not provide a full discussion of measurement, we emphasize the criterion of validity and conceptualize it more broadly than is customary.

We provide many examples to illustrate our points about methodology. However, one of our recurring examples is democracy. This is a concept that has been the center of attention in much of the methodological literature.<sup>1</sup> Moreover, it is a concept that is central to a broad body of substantive research in political science and other disciplines.

## CONCEPTUALIZATION

Concepts are the building blocks of the social sciences, as they are of all sciences. There is no theory without concepts, there is no description without concepts, and there is no explanation without concepts. Thus, concept formation – conceptualization – has logical priority in research because it is a key input in all subsequent steps, including those concerned with the production of data. Moreover, though quantity and quality are mutually complementary, every quantitative concept presupposes a qualitative concept. Indeed, as Sartori (1970: 1038) put it, because we cannot measure something if we have not specified its meaning, ‘concept formation stands prior to quantification’. Or, more broadly, as Bunge (1995: 3; 2012: 122) argues, ‘concept formation precedes empirical test’ and ‘in concept formation quality precedes quantity’ (see also Lazarsfeld and Barton, 1951: 155–6). Thus, researchers

need to focus on the formation of concepts and to recognize the qualitative foundations of all research.

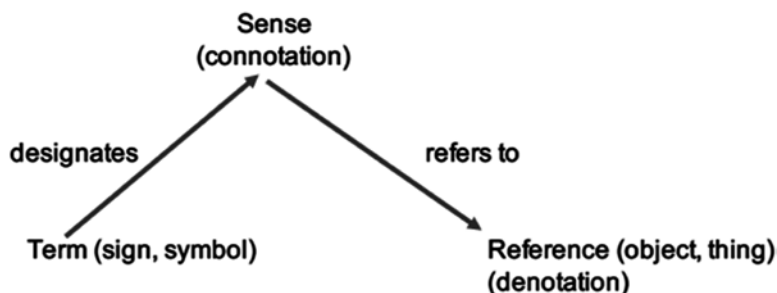
There are no rules on how to form a concept, just as there are no rules that can be followed to create a theory. Concepts are formed through a combination of induction and deduction. As suggested by Adcock and Collier (2001: 531–3), the decisions about a concept that is to be used in the social sciences are frequently made in light of a dialogue with ‘the broader constellation of meanings and understandings associated with a given concept’, or what they label the ‘background concept’. Moreover, the link between conceptualizing and theorizing is very close: as Kaplan (1964: 53) notes, ‘proper concepts are needed to formulate a good theory, but we need a good theory to arrive at the proper concepts’. Concept formation, like theory building, is largely an art.

Nonetheless, the product as opposed to the process of concept formation can surely be assessed. Concepts can be clear or vague. Concepts can be well formed or poorly formed. Concepts can be more or less elaborate and systematized. Indeed, there are various features of concepts that are used to distinguish good from bad concepts. At the very minimum, it is important to be clear about the various parts of a concept and the ways in which the sense or meaning of a concept is organized, which are two matters we address next.

### *Term, Sense and Reference*

A concept consists of three interrelated elements. The first is a *term*. This is a sign that designates the *sense* or connotation of a concept – the part of the concept that is frequently understood as its meaning – and the latter in turn refers to the objects that are included in the *reference* or denotation of a concept (see Figure 19.1).

Most of the discussion about concepts rightly focuses on concepts’ *sense*, which is



**Figure 19.1 The parts of a concept: term, sense and reference**

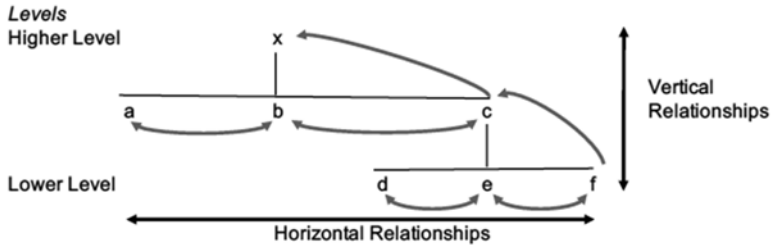
*Note:* This depiction is an adaptation of what is commonly known as the semantic triangle or the Ogden/Richards triangle (Ogden and Richards, 1923: 11).

given by the conceptual attributes that represent properties of objects and the relationship among conceptual attributes. Indeed, the meaning of a concept can largely be taken to be conveyed by a concept's sense, and debates about concepts focus mainly on this part of a concept. For example, debates about the concept of democracy since the work of Schumpeter (1942) hinge on matters such as what the conceptual attributes of democracy are and what the relationship among conceptual attributes is (Collier and Levitsky, 1997). We discuss this aspect of concepts more fully below. However, first, a few brief comments regarding a concept's term and reference are in order.

First, the role played by the term of a concept might seem rather simple. But Dahl's (1971) effort to introduce the term 'polyarchy', so as to avoid the possible confusion created by the multiple uses given to the term 'democracy', shows that terminological issues are not trivial. Indeed, there are many terms that are given different meaning. Furthermore, understanding how a term is used requires some knowledge of the broader semantic field in which it is embedded. For example, though the term 'regime' has a different meaning in the fields of comparative politics and international relations, the difference is clarified once the term is placed within the semantic field of these two fields of research. Thus, while terminological

matters are not the most important ones, they certainly deserve some attention (Sartori, 2009 [1975]: 61–9; 2009 [1984]: 111–15, 123–5).

Second, the idea of the *reference* of a concept needs to be clarified at the outset. A concept's reference (aka the domain of a concept) is all objects to which a concept refers and is thus related to the unit of analysis of a study. In contrast, a concept's *extension* is those objects which actually have certain properties. It is important to grasp the distinction between reference and extension, and the relationship between them. Though statements about reference rely on theoretical concepts and do not presuppose that of truth, statements about extension rely on empirical concepts and do presuppose that of truth. For example, it is one thing to say that democracy is a series of properties of political communities and another to say country *x* is a democracy. Indeed, the latter is an empirical claim, which could be factually true or false and can only be addressed once data has been collected, and hence is not strictly a conceptual matter (Bunge, 1974a: ch. 2; 1974b: 133–53; 1998a [1967]: 73–80).<sup>2</sup> Thus, we start our discussion here by considering theoretical concepts and purely conceptual operations, before turning to empirical concepts and empirical operations, such as the construction of indicators and data collection.



**Figure 19.2** The structure of a concept: levels and relationships

### ***The Attributes and Structure of a Concept***

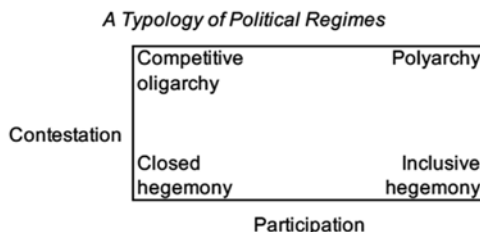
Turning to a more detailed discussion of a concept's sense, it is critical to recognize that a concept's sense is conveyed by (i) the conceptual attributes that represent properties of objects and (ii) the relationship among conceptual attributes or, for short, the structure of a concept. For this reason, the meaning of concepts is not fully conveyed by a simple listing of conceptual attributes, a common feature of definitions.

Listing the defining attributes of a concept is useful. It puts the focus on what conceptual attributes should be *included* in a concept. Moreover, inasmuch as a definition also clarifies what conceptual attributes should be *excluded* from a concept (even though they are included by some scholars), such an exercise is critical. For example, one of the ongoing concerns in the discussion about the concept of democracy is how to strike the right balance between expanding the concept of democracy beyond the sphere of elections. This can, for example, be done by adding attributes considered to be part of democracy (e.g. horizontal accountability), and expanding the concept of democracy in such a way that what might be considered extraneous attributes are included in the concept of democracy (e.g. the economic equality of citizens) (Munck, 2016).

However, it is important to note that any such list offers an incomplete sketch of a

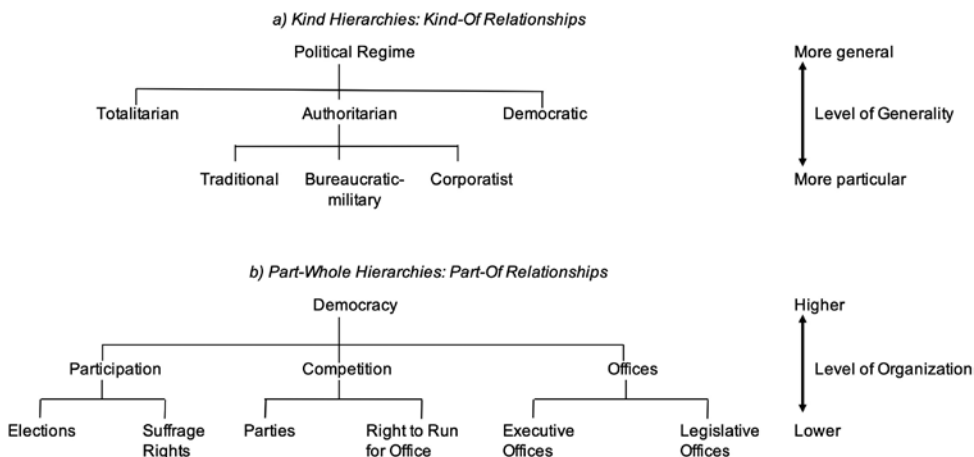
concept. Indeed, inasmuch as more than one conceptual attribute is posited, it is necessary to inquire about the structure of a concept, which is given by the relationships among conceptual attributes at the same level (horizontal relationships) and at different levels (vertical relationships) (see Figure 19.2). That is to say, the meaning of a concept might not be conveyed by each attribute taken individually, in an additive manner, and the structure of a concept could be key to its meaning. Thus, to fully and correctly grasp the meaning of a concept, it is crucial to appreciate that concepts can be – indeed, usually are – *conceptual systems*, which in turn are part of larger conceptual systems or semantic fields.<sup>3</sup>

It is also important to distinguish among different kinds of conceptual systems that connect and systematize multiple concepts that share, at least partially, their sense or their reference.<sup>4</sup> The simplest form is the *typology*, which unifies a series of concepts of different connotation but at the same level and of the same scope by proposing the underlying dimensions of multiple concepts. An example of such conceptual systems is Aristotle's (1995 [c. 330 BC]: Book III, chs 6 and 7) classical typology of political regimes, which relies on the underlying dimensions of the number of persons who exercise power and the ends they seek. Another is Dahl's (1971: 7) modern typology of political regimes, which relies on the underlying



**Figure 19.3 Conceptual systems I: typologies**

Source: Dahl (1971: 7).



**Figure 19.4 Conceptual systems II: two hierarchical structures**

Note: The example of a kind hierarchy draws on Linz (1975); the example of a part-whole hierarchy draws on Schumpeter (1942), Dahl (1971), and Przeworski et al. (2000).

dimensions of contestation and participation (see Figure 19.3).<sup>5</sup>

A different, and more complex, conceptual system is the *taxonomy*, which connects concepts at different levels in a hierarchical structure, in one of two ways. One hierarchical structure, sometimes called a *kind hierarchy*, organizes concepts that partition collections of objects into groups and subgroups, and yields a classic taxonomy. An example of this kind of conceptual system is Juan Linz’s (1975) encompassing and nuanced classification of 20th-century

political regimes (see Figure 19.4, panel a), which are defined in terms of the underlying dimensions of pluralism, ideology, mobilization and leadership. But there is another hierarchical structure, sometimes called a *part-whole hierarchy*, that organizes concepts that decompose wholes into parts and also connects parts to the whole.<sup>6</sup> A classic example of such a hierarchy is the conceptualization of democracy that decomposes a whole – democracy – into parts at various levels of organization (see Figure 19.4, panel b).



## Evaluation

Concepts are not true or false. Nonetheless, not all concepts rest on an equally sound foundation; some have been carefully elaborated and justified, while others are merely stipulated without much in the way of reflection. Without making any claim to exhaustiveness, we propose a set of criteria to assess whether concepts are good or bad.

Most basically, concepts have to be *intelligible*. This means that we should be able to answer the following questions: what is the concept designed by the term or symbol used for? What are the conceptual attributes? What is the structure of a concept, that is, what are the relationships among conceptual attributes? What is the reference of a concept?

Several criteria regarding a concept's term can be highlighted (Sartori, 2009 [1975]: 61–9; 2009 [1984]: 111–15, 123–25, 132–3; Gerring, 1999). Most critical is the criteria of *terminological univocality*, that is, the avoidance of terminological ambiguity introduced through the use of homonyms, terms with multiple meanings, and synonyms, multiple terms with the same meanings. Other criteria concern the *fit* of the term with the terminology used in prior research, and the *familiarity* and *resonance* of the term.<sup>7</sup>

Another criterion is *logical formation*. Inasmuch as any concept consists of more than one conceptual attribute, it is important to ask whether the proposed conceptual system fulfills, for a given domain, two logical requirements. First, they should be mutually exclusive, meaning that no concept or conceptual attribute at the same level overlaps with the meaning of another concept or conceptual attribute; and they should be collectively exhaustive, meaning that no concept or conceptual attribute that is part of a conceptual space is excluded. In addition, inasmuch as any concept consists of more than one conceptual attribute at different levels, whether or not the conceptual attributes

are logically organized by level of generality or organization is a key consideration (Lazarsfeld and Barton, 1951: 156–9).

Yet another key criterion, which deserves some elaboration, is *conceptual validity*, understood here with reference to the sense of the concept and both the conceptual attributes and the structure of a concept.<sup>8</sup> The inclusion and exclusion of conceptual attributes is a key decision in the formation of a concept. The same goes for any decision regarding the relationship among conceptual attributes. And each decision can and should be assessed in terms of the extent to which the decision is theoretically justified.<sup>9</sup>

It bears noting that the assumption underpinning this criterion – that concepts can and should be assessed in light of their theoretical justification – is not universally accepted. On the one hand, many scholars posit that a number of concepts, and especially those that have an obvious normative connotation, are ‘essentially contested’ and that they will always remain ‘open’ in the sense that a research community will never agree on a definitive definition (Gallie, 1956; Gray, 1978). From this relativist perspective, any claim that a certain concept is more theoretically justified than another can be portrayed as arbitrary or subjective. This perspective could even lead to the view that since disputes over the meaning of concepts cannot be resolved, any effort at measurement is futile, in that claims about what is measured cannot be settled.

However, it is not obvious that, for example, democracy, seen as the ‘essentially contested’ concept *par excellence*, merits such a characterization (Bobbio, 1989: ch. 4, 2003: 42; Beetham, 1994: 27; see also Arblaster, 2002: 6–10). Though disagreements about the concept of democracy persist, it is clear that the research by Schumpeter (1942) and Dahl (1971) has led to widespread consensus about the core meaning of democracy in research on democratization (Munck, 2009: 16–23, 2016). The same can be said about other concepts with strong normative

resonance. For example, Waldron (2002) observes that while the institutional or political arrangements required by the rule of law – another concept frequently characterized as essentially contested – are subject to disagreement, there is actually considerable consensus about its basic formal–legal requirements, such as that laws are prospective, open and clear and that there is congruence between official action and declared rule (see also Collier et al., 2006: 228–30; Møller and Skaaning, 2014: ch. 1).

On the other hand, a common epistemology, empiricism, holds that knowledge only concerns observable properties and that empirical concepts but not theoretical ones are acceptable (Bridgman, 1927; Carnap, 1936, 1937). From this perspective, the suggestion that concepts could be assessed in light of theory would be deemed unjustified and all work on theoretical concepts would be no more than a distraction from, and even a hindrance to, the real work of measurement (King et al., 1994: 25, ch. 2, 109–10). However, the distinction between, and mutual irreducibility of, theoretical and empirical concepts is well established (Kaplan, 1964: 54–60; Sartori, 2009 [1975]: 83–4; Laudan, 1977: chs 1 and 2). And the shortcomings of the empiricists' endeavor to reduce the theoretical to the empirical are evident (Bunge, 2012: ch. 13). Indeed, the main concepts in the social sciences are theoretical as opposed to empirical. Key examples are society, economy, class, ideology, politics, state, power, rights, constitutionalism, democracy, rule of law, welfare and peace. Few scholars are willing to remain silent about these concepts.

In short, these two extremes can and should be avoided. *Contra* Gallie, many key concepts have been theoretically developed enough to have some shared meanings, and measurement does not have to wait until all conceptual disputes are resolved. *Contra* empiricists, the banishment of theoretical concepts is simply a self-defeating position that is hard to consistently maintain. Thus,

the validation of concepts by reference to theory is both viable and central.

## MEASUREMENT I: DATA ON INDICATORS

Turning from conceptualization to measurement opens up a whole new series of challenges. Theoretical concepts refer to at least some imperceptible facts. Thus, inasmuch as social scientists seek to describe and explain the world, they must address some complicated empirical operations involved in measurement (see Figure 19.5). First, to bridge theoretical concepts and facts, they must develop *indicators*, which relate observable properties to unobservable ones, and propose how to draw distinctions based on indicators. Second, to produce data, they must engage in *data collection*, which assigns (qualitative or quantitative) values to indicators in light of observables about objects. In other words, they must design and use measuring instruments. Thus, though any attempt to produce data must begin with a clear idea of *what* is to be measured, the distinct issues involved in *how* to measure some theoretical concept –

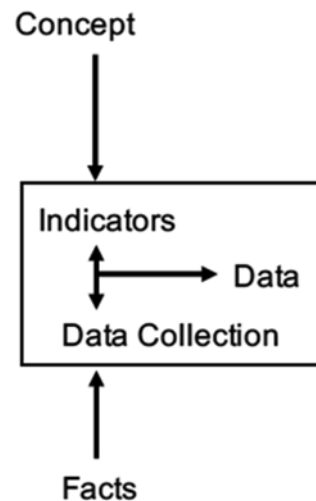


Figure 19.5 The concept–fact interface

the development of indicators and, relatedly, of measurement scales, and data collection and coding – deserve scrutiny.

### **Indicators**

The general challenge in developing indicators, sometimes called operationalization, is to build a bridge between unobservables and observables, that is, to link theoretical concepts that refer to facts about properties of objects with empirical concepts, an observable property of the same object (Bunge, 1998b [1967]: 192–6).<sup>10</sup>

Due to the nature of indicators, probably the hardest challenge in the design of a measuring instrument relates to what is usually called *content validity* – the extent to which one or more indicators capture the correct and full sense or content of the concept being measured (Adcock and Collier, 2001: 536–40). The goal of measurement is to generate data that can be used to evaluate the truth of claims about facts (e.g. the US is a democracy in 2019). But data will be useful for this purpose only inasmuch as any data collected on some indicators can be linked back to the concept (e.g. the concept of democracy in this example) used in a factual claim. Building such bridges is anything but an easy task, especially when the concept of interest is multidimensional, that is, has many conceptual attributes.

This task is made harder because researchers also have to be concerned about measurement *equivalence*, the extent to which an indicator captures the same sense or content of the concept being measured in different contexts (Przeworski and Teune, 1970: chs 5 and 6; Adcock and Collier, 2001: 534–6; Davidov et al., 2014). Often it is not obvious that the same indicator will have similar meanings in different countries, for different persons and in different time periods. This means, first, that it is often necessary to use several indicators to measure a concept in order to capture different nuances

of the concept and increase the reliability of any measures, and second, that different contexts sometimes call for different indicators to capture the same facts. This can be understood by invoking the distinction between common and system-specific indicators (Przeworski and Teune, 1970: ch. 6). Common indicators work in the same way across different contexts, while system-specific indicators vary across contexts but are functional equivalents, meaning that they are in principle substitutable. For example, actions considered corrupt in one place are considered appropriate behavior elsewhere, so asking similar questions about corruption will not provide equivalent measures. In survey research, questions should preferably have the same meaning for all respondents, but linguistic, cultural and other differences make it difficult to establish measurement equivalence (see also Locke and Thelen, 1995; van Deth, 1998).

At the same time, the search for system-specific indicators can lead to an excessive, even paralyzing, emphasis on the unique and can open the door to relativism. For example, most current global datasets on democracy rely on common indicators and hence could be criticized for not taking into account how different conceptual attributes of democracy should be adapted to different contexts. In addition, some of these datasets have been criticized, and rightly so, for having a Western bias, in that specific Western institutions are treated as universal standards for assessing other countries. However, it is clear that an attempt to factor in ideas from the literature on ‘non-Western democracies’, especially the argument that democracy takes a different form in non-Western societies, amounts to a rejection of any standard to compare countries around the world. What might at first glance seem a rather simple empirical operation – the design of indicators – actually hides many potential pitfalls. Indeed, for these reasons, the development of indicators that offer a basis for testing factual claims has been recognized as an important accomplishment

(Harré, 1981), and the development of broad cross-national measures, such as those used to measure economic activity around the world, are celebrated (Vanoli, 2005).

### **Measurement Scales**

The design of indicators is inextricably linked with another task, namely, the design of the measurement scales used to discriminate among cases. The standard options are well known: there are, most basically, nominal, ordinal, interval and ratio scales. Moreover, the standard way of comparing these scales barely needs mention: the move from nominal toward ratio scales involves a gain in precision. Thus, all else being equal, scales designed to collect data that is more precise and informative are preferable. Or, as is sometimes argued, inasmuch as nominal and ordinal scales are treated as qualitative scales, quantification is a sign of progress.

However, we add a caveat to this conventional wisdom that is suggested by the debate about whether democracy is best seen as a question of either-or or more-or-less (Sartori, 1987; Bollen, 1990; Collier and Adcock, 1999). In this debate, many authors have suggested persuasively that nominal and ordinal scales are sometimes preferable, in that they actually capture better the concept of interest. For example, the common idea of a democratic transition suggests that some changes are actually qualitative in nature and hence that nominal scales are appropriate (Przeworski et al., 2000: 18). Likewise, a common argument in the literature on democratization is that the extension of the right of suffrage evolved one social group at a time, a change well captured by an ordinal scale. Thus, it is important to note that decisions regarding measurement scales are made in the context of specific concerns and concepts, and hence that, as Collier and Adcock (1999: 537) suggest, these decisions should be justified through ‘specific arguments linked to the goals of research’

rather than by reference to the superior information of certain scales when considered in the abstract.<sup>11</sup>

### **Data Collection**

Once a researcher has designed an indicator or a series of indicators, each with their own measurement scale, the distinct task of data collection – the gathering and categorization of relevant information about a phenomenon of interest for a researcher – can begin in earnest. In this regard, we caution against a narrow view of the possible kinds of data and sources of data, and hence a narrow view of the challenges involved in data collection and the problems that might emerge in the course of data collection. Indeed, an overreliance on data from data-rich countries or time periods (e.g. the US in current times) would likely introduce bias into our knowledge of the social world. Moreover, in thinking about data collection, we draw attention to three questions: (1) When and where was it created? (2) Who created it? (3) For what purposes was it created? Answers to these questions provide the background information required to carry out systematic source criticism (*Quellenkritik*), which is the process of evaluating whether information (of all kinds) is more or less valid, reliable or relevant for a particular purpose.

### **Sources of qualitative data**

A key distinction is frequently made between primary sources and secondary sources. Primary sources provide direct or firsthand evidence about events, objects or persons. They include historical and legal documents, eyewitness accounts, interviews, surveys, audio and video recordings, photographs, speeches, diaries, letters, art objects and various kinds of online communications (e.g. emails, tweets, posts, blog entries). Secondary sources provide some kind of interpretation and analysis of events, conditions or experiences. Hence, newspaper articles and reports can be either

primary or secondary sources, depending on whether they provide information about facts or analysis and interpretation.

The ideas associated with systematic source criticism and the distinction between primary and secondary sources have their origin in the academic discipline of history. Historical data presents a number of attractions for social scientists, including more variation on key variables, the ability to investigate how similar causal mechanisms play out in different contexts and the ability to analyze path dependency. However, the more social scientists delve back in time, the more they come to depend on the prior work of trained historians, who have produced the narrative accounts that social scientists use either to code historical datasets or to produce in-depth historical narratives.

This raises an important but often ignored challenge: that social scientists will be prone to solely enlist or overly emphasize ‘works by historians using implicit theories about how events unfold and how people behave very similar to the theory under consideration’ (Lustick, 1996: 607). To mitigate this risk, social scientists first need to recognize that historical work cannot be seen as theoretically neutral. The implicit or explicit theoretical and historiographical perspectives of historians (e.g. the Marxist or Annales schools) color the ways they interpret their findings. Social scientists must therefore build a representative body of historical data from which to draw inferences. This means that they need a deep knowledge about the development of historiography and the debates of historical work in a particular field (Lustick, 1996; Lange, 2013: 141–8).

Different guidelines have been developed to ensure this. Lustick (1996) proposes four strategies:

- *Explain variance in historiography*: Assume a normal distribution among historical works and then identify the consensus.
- *Be true to your school*: Identify a particular historical tradition or school as superior for the pur-

pose at hand and then accept this interpretation, knowing how it differs from other interpretations.

- *Quasi-triangulation*: Limit the readings of history to those interpretations that have a broader support across historical schools.
- *Explicit triage*: Argue why some historical studies are better than others given the task at hand.

Møller and Skaaning (2019: 6) endorse Lustick’s argument that social scientists need to systematically consider differences between historical interpretations, but they criticize the notion that the average or consensus interpretation is less biased. Instead, to avoid selection bias in the sources of data, they suggest that social scientists should factor in the ‘shape of the distribution within historiography’ in three ways:

- *Aim for conceptual consistency*: Prioritize historical interpretations that are based on similar concepts as those being considered by the social scientist.
- *Clarify the vantage point of historical accounts*: Prioritize historical interpretations that are relatively atheoretical or where the thesis conflicts with the thesis that the social scientist is interrogating.
- *Prioritize updated evidence*: Prioritize historical interpretations that are based on newer evidence.

These three criteria are anchored in a simple Bayesian logic and they enable social scientists to heed what has been termed ‘the Ulysses Principle’, that is, to figuratively tie oneself to the mast in order to take precautions against influencing the evidence that is used to examine descriptive or causal propositions (Møller and Skaaning, 2019).

This principle, it bears noting, is not only relevant when dealing with historical sources. Recent methodological debates have emphasized the possibility of going deep more generally by shifting the focus from the macro level of analysis to the micro level, so as to probe mechanisms (Beach and Pedersen, 2016). While there are different ways of doing this, they all force social scientists to deal with qualitative data sources, such as interviews, archives, newspapers, organization records

and reports and participants' observations (see Tilly, 2008; McAdam et al., 2008). This requires not only a close familiarity with the data, but also careful consideration about how to avoid bias in the identification and reading of qualitative sources. If one takes out the historical part of the criteria mentioned above, they are applicable for processing many different kinds of qualitative data.

### *Sources of quantitative data*

Shifting focus to quantitative data, these normally take one of five forms:

- 1 *Hand-coded data*, such as the CIRI Human Rights Database, the Manifesto Project Database, the Uppsala Conflict Data Program, the Freedom House data on Political Rights and Civil Liberties and the Polity IV Project, where researchers or their assistants code events or conditions based on some predefined criteria.
- 2 *Machine-coded data*, such as the Integrated Crisis Early Warning System, the Global Database of Events, Language, and Tone (GDELT), and the Fragile States Index, where researchers develop automated algorithms that can categorize behavior, conditions or opinions.
- 3 *Ordinary survey data*, such as the World Values Survey, the Afrobarometer and various national election studies, where a sample (often representative) of people belonging to a particular group (citizens of a nation, employees in a firm, parliamentarians, members of an organization, etc.) is enlisted to respond to a number of questions about opinions and behavior.
- 4 *Expert survey data*, such as parts of the Varieties of Democracy dataset, the Chapel Hill Expert Survey, the Perceptions of Electoral Integrity dataset and the Quality of Government Survey, where experts are enlisted to answer questions about a certain topic about which they have special competence.
- 5 *Administrative data*, such as election turnout and vote share, roll call votes, number of state employees and government financial and economic statistics, which have been collected by national public agencies and international organizations (e.g. the UN, the World Bank, the IMF and the OECD).

In situations where we are interested in measuring not opinions but the actual condition

of, say, different aspects of democracy or the prevalence of corruption, another distinction has received much attention: namely, the difference between fact-based and judgment-based indicators. Those favoring fact-based (directly observable and verifiable) indicators emphasize that such data are more transparent and replicable and therefore broadly recognizable. They criticize judgement-based and perception-based data for being based on fuzzy and unsubstantiated inferences and personal biases.

Users and producers of judgement-based indicators have responded to this criticism by pointing out that fact-based indicators are often unable to capture all relevant nuances of particular phenomena. The preference for fact-based data rests, according to Schedler (2012: 28), on two conditions, which are often not fulfilled: '(1) transparent empirical phenomena whose observation do not depend on our judgmental faculties and (2) complete public records on those phenomena'. For example, some aspects of democracy, such as freedom of expression, are not easily observable. More generally, '[s]ome empirical phenomena we cannot observe in principle, others we cannot observe in practice' (Schedler, 2012: 28). In a nutshell, the problem is that directly observable empirical information is often incomplete, inconsistent or insufficient.

Different types of evidence can of course be used simultaneously to answer particular research questions. Just as researchers can make use of methods triangulation in order to appraise theoretical expectations, they can also carry out data triangulation and take advantage of the strengths and shortcomings of different kinds of sources of data (Skaaning, 2018). In general, the combination of information from different kinds of data increases our ability to capture related, but distinct, aspects of the variable in question. In addition, relying on multiple indicators can reduce the impact of idiosyncratic measurement errors associated with single indicators and facilitates systematic assessment of how reliable the data are.

There has recently been much talk of new data collection methods, based on increased computer power and a plethora of new information that is accessible online – what has been referred to as ‘big data’. Web scraping of information from, for example, newspapers or Wikipedia or social media (Twitter, Facebook) allows scholars to build large data-sets. One partial novelty here is to treat text – including alterations of text on Wikipedia and the like – as data. These newer sources of data collection are addressed in other chapters of the *Handbook*. Hence, all we note here is that the issues of conceptualization and measurement discussed in this chapter are also relevant for these new data collection enterprises.

### *On coding*

One of the more versatile means of producing systematic data – whether quantitative or qualitative, whether on variables or causal mechanisms, whether for a large-scale or a small-scale project, whether for the current period or times long past – is hand-coding by a single scholar or a team of scholars. Even though this is only one among various means of assigning values to indicators, given its important role in the social sciences we offer some comments about this procedure.

The production of hand-coded data normally proceeds in particular stages. Relevant information is gathered, after which a coder evaluates the evidence on one or more issues and translates it into a score based on more or less explicit and precise standards or coding rules. Despite careful attention to the selection of sources, training of coders and documentation of coding procedures, specific biases can still influence the scores (Bollen and Paxton, 1998, 2000).

The accessibility and selection of sources is a major issue. Evidence has been through a filtering process in which some information passes through and some is filtered out. This process is likely to introduce problems because the filters are selective in non-random ways, meaning that the

information is generally neither complete nor representative.

If the patterns of incomplete data are not random, descriptive and explanatory analyses using the data will be biased. For instance, Casper and Tufis (2003) have demonstrated that some of the most prominent democracy measures are not genuinely interchangeable, even though they are all anchored in Dahl’s (1971) definition of polyarchy and even though they are highly correlated (between .85 and .92). One reason for this could be systematic missingness. For example, relevant information is frequently not available for poor countries and autocracies. Missingness can be evaluated by simple tests of non-random missingness (see e.g. Ríos-Figueroa and Staton, 2012), where one examines whether there are significant differences between the scores for units covered by the data and those units that are not covered on other variables expected to be related to the outcome that is being researched.

Another issue is how the coders or respondents process the evidence. They can introduce random and systematic measurement errors by interpreting the sources differently, either because they base their evaluation on different pieces of (relevant or irrelevant) information, because they weight the same evidence differently or because they have different understandings of the concepts and scales that are used. More generally, various actors in the ‘data supply chain’ respond to different incentives and have variable capabilities that influence – and sometimes consciously manipulate – the production of data (Herrera and Kapur, 2007).

In addition, the practical procedures in the specific coding processes can introduce method effects. For example, scores can be influenced by how many units and questions the coders process, whether and when revisions can be made or whether they code across cases or over time. All of these factors tend to influence the implicit reference points in the minds of coders and thus the scores that are generated through exercises in coding.

On a more general level of abstraction, the reproducibility of measurement procedures is an important aspect of social science. This requires a systematic approach to data collection, precise descriptions of the procedures and transparency of these issues. Ideally, researchers should be able to reproduce or replicate the scores, and then assess the results of independent coding exercises. For example, where multiple, overlapping indicators exist, if the same variable is coded by several coders for the same units, one can assess the extent to which they generate consistent and converging data. In such cases, inter-coder reliability tests are valuable tools to assess whether the assumptions about consensus among coders are met (Gwet, 2014).

One way to do this is to employ Item-Response Theory (IRT) modeling techniques. These use patterns of agreement between the scores from different coders/indicators (and sometimes also other kinds of information, such as coder characteristics) to identify variations in reliability and systematic bias, and use this information to reduce measurement error in connection to latent concepts and to generate systematic estimates of uncertainty.

## **Evaluation**

An evaluation of measuring instruments and the data on indicators produced by using these instruments, much as with concepts, hinges first of all on *intelligibility*. If an independent scholar is not able to comprehend how the data was produced, what decisions were made to produce the data, and what the reasons were for at least the key decisions, the data cannot be properly scrutinized. In other words, without transparency, there is no possibility of replication and no way of assessing reliability and validity.

The demand for transparency has traditionally been directed mostly at quantitative data, but it has recently been pushed by the DA-RT (Data Access and Research Transparency)

initiative within the American Political Science Association with respect to qualitative research as well. One of the tools that has been proposed is data repositories that allow researchers to store qualitative data in a systematic way. This enables scholars to document their evidentiary record and makes it possible for other scholars to acquaint themselves with what is written in the sources that are referred to for evidence. For instance, the use of active citation gives readers a quick way to assess if a particular observation or interpretation does indeed seem to be supported by the work that is referenced (Lupia and Elman, 2014).

There are many other criteria that could be used to assess measuring instruments and data on indicators. As noted, measuring instruments can be more or less *versatile*, that is, they can be better or worse suited to generate data on various concepts in different domains (that is, temporal and spatial units). Data can be more or less *reliable*, that is, yield the same results when repeated measures are carried out independently. Data can have more or less *measurement error*, and identifying the sources of such error and providing estimates of uncertainty is part of best practice.

Importantly, in contrast to the evaluation of concepts, the evaluation of data on indicators can rely on empirical tests, using the data that has been produced and other available data (Cronbach and Meehl, 1955; Campbell and Fiske, 1959; Adcock and Collier, 2001; Seawright and Collier, 2014; McMann et al., 2016). For example, in a test of *convergent-discriminant validity* a researcher examines to what degree a new measure converges with established measures of the same concept and diverges from established measures of different concepts. In turn, in a test of *nomological validity* a researcher examines to what degree a new measure is able to reproduce well-established relationships among variables. Thus, it is important that researchers take advantage of the various empirical tests that can yield information that is relevant to an assessment of data.



However, the value of such tests depends very much on the current state of empirical knowledge. That is, a test of convergent-discriminant validation requires that a researcher can take for granted that the other measures, the standards with which the measure of interest is compared, are valid. In turn, a test of nomological validation requires that a researcher can take for granted that the established relationship is valid. Yet frequently this is not the case, and hence these tests may simply not be relevant. Moreover, the proponents of new measures frequently challenge existing conceptualizations or explanations, making agreement with prior knowledge an improper standard.

Thus, it is critical to stress the centrality of the question of *content validity*, that is, the extent to which one or more indicators capture the correct and full sense or content of the concept being measured (Adcock and Collier, 2001: 536–40). Assessing the validity of data is complex, because it concerns the link between observables and unobservables. Moreover, unlike estimates of convergent-discriminant and nomological validity, it cannot be quantified through an analysis of the data. However, it is important to recognize some key points about content validity. First, the question of content validity is distinctive. Second, it has priority in an evaluation of measurement validity, in the sense that it should be addressed first, during the process of indicator construction, and that it affects the data that are used in tests of convergent-discriminant and nomological validity. Third, it is an important consideration regardless of the kind of data (quantitative or qualitative) that is produced.

## MEASUREMENT II: DATA ON INDICES

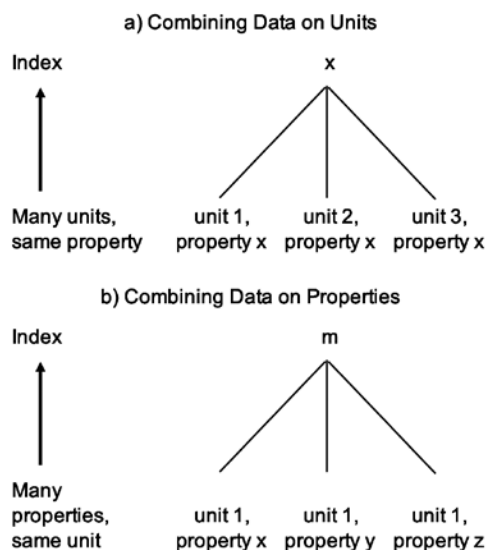
Data analysis for the purpose of description and explanation frequently relies on data on indicators. However, the production of data on indicators frequently raises a new

question: how might these data on indicators be combined? Indeed, there are many reasons why a scholar may want to develop what can generically be called *indices*, which combine data on indicators. The production of indices involves complex considerations, several of which are of a technical nature, and there is a large literature on index formation (e.g. Lazarsfeld, 1958; Lazarsfeld and Menzel, 1961; Blalock, 1982: ch. 7; Bollen and Lennox, 1991; Nardo et al., 2005; Greco et al., 2019). Thus, our discussion is necessarily cursory. Nonetheless, we draw attention to some key distinctions and options that have not always been addressed with clarity in the recent literature, and introduce some considerations that are ignored by the literature on measurement that pays little or no attention to the connection between theoretical concepts and measurement.

At the broadest level, drawing on the distinction between two of the core parts of a concept, its sense and reference (see above), it is possible to distinguish between two kinds of indices: (i) indices that combine data on the same indicator (measuring the same property) in multiple units (e.g. percentage of people in the world earning less than 2 dollars a day), and (ii) indices that combine data on multiple indicators (measuring different properties) in one unit (e.g. how democratic is the US) (see Figure 19.6).<sup>12</sup> In addition, building on these two kinds of indices, megaindices can be, and frequently are, built (e.g. proportion of countries in the world that are democracies, proportions of country dyads in the world that are democratic dyads, etc.). However, the core issues and options concern these two basic situations.

### *Combining Data on Units*

In the social sciences, the lowest level of analysis is the individual, and hence the most fine-grained data that are collected are data on properties of individuals. From this basic starting point, it is possible to combine data



**Figure 19.6 The production of data on indices: two basic situations**

on units all the way up to the highest possible level of analysis, the world system. However, there are two different ways, corresponding to two different social properties, in which data on units can be combined, and the index that is produced is different depending on which option is chosen (Lazarsfeld, 1958: 111–12; Lazarsfeld and Menzel, 1961: 426–8).

When the data on different units (e.g. individuals, firms or states) concerns a property possessed by each unit (e.g. income or life), an index that represents an *aggregate or resultant property* is generated. Examples are GDP, GDP per capita, percent of GDP accounted for by trade, global GDP, number of deaths in war, homicides per 100,000, proportion of the population that supports democracy and percentage of votes won by candidates in an election. In turn, when the data on different units concerns a property a unit has by virtue of a relationship among units (e.g. relative income, capital–labor relations or trading relationship between states), an index that represents a *relational or structural property* is generated. Examples are

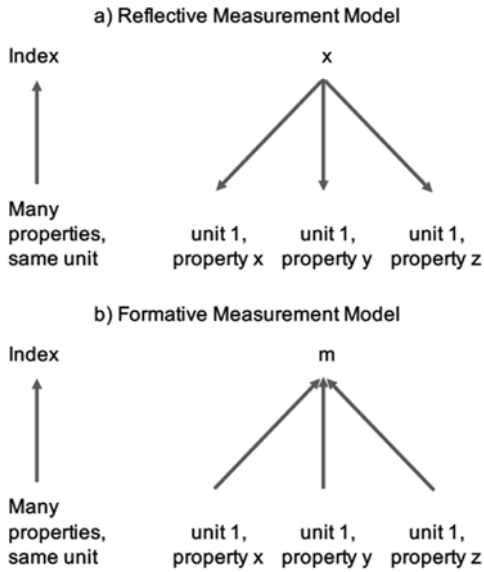
income inequality, polarization of the class structure, conflict levels of industrial relations, judicial independence, state legitimacy, trade dependence between countries and eigenvector centrality.

These are not the only social properties. Indeed, as Lazarsfeld (1958: 112–13; Lazarsfeld and Menzel, 1961: 428–9) pointed out long ago, there is a third kind of social property: *global or emergent properties*. These properties are not based on information about lower level units because they are not possessed by each lower level unit either independently of other units or due to a relationship with other units. Examples of global or emergent properties are crowd behavior, national culture, social cohesion, political stability and the dominant mode of production. The measurement of such social properties does not proceed by combining data on the same property in multiple units.

### **Combining Data on Properties**

A second kind of index is produced by combining data on multiple indicators (measuring different properties) in one unit. To be sure, the production of such indices does not need to be limited to one unit. For example, though some scholars have developed an index of democracy for one country, it is common for scholars to produce indices covering many countries or even the entire world. The increase in the number of units opens some important possibilities, such as tests of dimensionality. But the point is that the focus of such index production is on the question of how data on multiple indicators, each linked with different conceptual attributes, should be combined.

This challenge has been the subject of considerable debate, and different scholars have different views about how such a challenge should be addressed. Nonetheless, in broad strokes, the key choice a researcher faces is whether aggregation, that is, the combination of data on multiple indicators, should be



**Figure 19.7 Combining data on properties: reflective and formative measurement models**

based on what has been called a reflective or a formative aggregation model.<sup>13</sup>

These aggregation models differ both conceptually and substantively. In a *reflective model*, the concept is understood as the common ‘cause’ of the indicators used to measure it. Hence, ‘causation’ runs from the latent concept to the indicators (see Figure 19.7, panel a). Changes in the latent trait (not directly observed) are therefore expected to ‘cause’ a change in the indicator scores, but not vice versa, and that change in the latent variable should simultaneously bring about variation on all indicators. It follows that indicators should have a high positive correlation. This indicates that the multiple indicators and hence conceptual dimensions tap into a single underlying dimension. If so, indicators can be seen as partially interchangeable and dropping one of these indicators would not alter the meaning of the index that is produced. A good example is Teorell’s (2010: 164–5) socioeconomic modernization index, which he constructs, through the use of principal components analysis, by combining information on nine indicators: net

output of the non-agricultural sector as percentage of GDP, gross secondary school enrolment ratio, urban population as percentage of total population, life expectancy at birth, infant mortality rate, the log of GDP per capita, radios per capita, televisions per capita and newspaper circulation per capita. The indicators all load highly on a common latent dimension, which lends support to the index construction.

The assumptions behind a *formative model* are different. A latent concept is construed as the summary of the relevant variation in a set of indicators that are understood as constitutive of a particular concept. In other words, a latent concept is composed of conceptual attributes that are individually important for the meaning of the concept. In this case, ‘causation’ flows from the indicators to the latent concept (see Figure 19.7, panel b). In contrast to reflective models, in formative models the correlation among indicators is considered irrelevant and, since the indicators are understood as defining attributes, excluding one or more of them will fundamentally alter the meaning of the concept that is to be captured. To illustrate, contestation (or competitive elections) and inclusive suffrage are often conceived as the two essential features of representative government (Dahl, 1971; Coppedge et al., 2008). These two conceptual attributes are not necessarily highly correlated with each other. Today, many countries have universal adult suffrage but not much contestation, and historically many countries had a high degree of contestation but highly restrictive voting rights. However, only including indicators that capture either suffrage or contestation would critically alter the core concept that is being measured. Measuring one property cannot substitute for the measurement of another property, and dropping the data on one of the multiple properties would radically alter the meaning of the index that is produced.

### **Evaluation**

The constructors of indices must tackle some distinct choices, beyond those that go into the

production of data of indicators, when they consider whether and how to aggregate data from indicators. In combining data on the same indicator in multiple units, analysts need to be aware of what social property is being measured, and hence whether the appropriate procedure is being used. In turn, in combining data on multiple indicators (measuring different properties) in one (or more) units, they have to be aware at least of the choice between reflective or formative aggregation models. However, as Lazarsfeld (1958: 113) noted, it is by no means self-evident how an analyst should proceed. Indeed, at times it is not even clear whether an analyst faces the challenge of combining data on units or on properties.

Given this uncertainty, the temptation to rely on default options might be strong. But this temptation should be resisted. As with the evaluation of data on indicators, empirical tests, using the data that have been produced and other available data, can be conducted and used to inform the construction of indices. Indeed, various empirical checks can be of help (Bollen and Bauldry, 2011). However, one cannot simply make the data speak for itself. Thus, no matter which of the options is seen as more suitable for a given aggregation task, whatever procedure is used to form an index through the combination of data on indicators needs to be justified theoretically. What this means is that, to ensure what has been called concept-measure consistency (Goertz, 2006: ch. 4), which might be thought of as a counterpart or aspect of the criterion of *content validity* discussed above, what is needed is a theory about how multiple indicators should be combined (Goertz, 2006: 53–65, ch. 5; Munck, 2009: 30–2, 49–51). Indeed, much as with data on indicators, data on indices are valid if they fulfill two criteria: (i) a theoretical concept has been formed in a conscious and careful manner, that is, a theory has been articulated to justify what conceptual attributes are included and excluded, how the included conceptual attributes relate to each other and what the referents of conceptual attributes are;

and (ii) the way in which data on indicators is combined matches the concept that is being measured.

## CONCLUSION

The social sciences, in contrast to disciplines such as logic and mathematics, are factual sciences, given that they refer to facts about the concrete world. Thus, empirics and, more narrowly, measurement, understood as the production of data, are essential parts of social science research. However, empirics should be distinguished from empiricism. Empiricism is a one-sided epistemology that holds that experience is the only source of knowledge and that, in the context of measurement, asserts that theoretical concepts are not different from empirical concepts or that theoretical concepts can be reduced to empirical ones. The history of science reveals the limitations of empiricism. Indeed, a widely recognized indicator of progress is the replacement of classification schemes based on concepts that represent secondary, observable properties with ones based on primary, non-observable properties of things. For example, the conceptualization of chemical elements based on atomic number and electron configuration rather than observable properties such as color or smell, and the classifications in biology based on molecular differences rather than observable morphological traits.

Thus, counter to an empiricist approach to measurement, this chapter places the focus squarely on theoretical concepts and insists on the link between theoretical concepts and measures. Indeed, we have sought to draw attention to various ways in which a clear idea of *what* theoretical concept is to be measured is needed to make decisions regarding *how* to measure that theoretical concept. And to that end, we started by addressing what concepts and conceptual systems are, and then highlighted how both the production of data

on indicators and indices should consider the link between concepts and measures.

We do not seek to convey the message that the link between concepts and data should be the only concern in any measurement project. Other matters are also important. Moreover, not every project on measurement has to be conjoined with a project on conceptualization. There can surely be a division of labor between researchers who seek to form concepts and researchers who produce data. However, for data to be used to ascertain the truth of the kind of factual claims that are routinely made in the social sciences, decisions regarding the production of data *must be guided* by ideas regarding the sense and reference of concepts as well as their structure. Measures that ignore these matters are of limited value and, inasmuch as they are interpreted as measures of theoretical concepts, potentially erroneous.

## Notes

- 1 Collier and Levitsky (1997); Bollen and Paxton (1998, 2000); Collier and Adcock (1999); Munck and Verkuilen (2002); Goertz (2006); Gerring et al. (2019).
- 2 This distinction between reference and extension is frequently overlooked. Indeed, it is not addressed in the influential work on the social sciences by Sartori (1970, 2009 [1984]: 102–6). However, it actually is consistent with Sartori's (2009 [1975]: 84) clarification that 'the rules of transformation along a ladder of abstraction ... apply to observational, not to theoretical, concepts'. That is to say, though the intension and extension of a concept varies inversely (Sartori, 1970: 1040–4; Collier and Mahon, 1993: 846), this statement applies only to empirical concepts and not to theoretical concepts. For more on the distinction between theoretical and empirical concepts, see Kaplan (1964: 54–60).
- 3 Bunge (1998a [1967]: 82–9); Sartori (2009 [1984]: 118–25); Thagard (1992); Collier and Levitsky (2009).
- 4 For useful discussions about concepts and conceptual systems, see Bunge (1998a [1967]: chs 2 and 3), Bailey (1973, 1994), Marradi (1990) and Collier et al. (2012).
- 5 2x2 typologies have been hugely influential in social science, both for descriptive and explanatory purposes. But typological property spaces are often much more complicated, as they can contain more than two dimensions and as each of these dimensions can be divided into more than two classes. For example, one could add the rule of law, the effective power to govern and/or the guiding ideology as separate dimensions to the regime typology in Figure 19.3 if one has good theoretical reason to do so. Or one could subdivide contestation and/or participation into different levels, say low, medium and high. The problem with such operations is that the property space can quickly become too complex to be useful for theorizing and empirical analysis. Thus, the essence of forming a typology is to first identify the dimensions and the classes on each dimension, and then to reduce the property space in order to focus on the most important types (Lazarsfeld and Barton, 1951: 169–80; Elman, 2005; Collier et al., 2012).
- 6 On the distinction between kind and part-whole hierarchical structures, see Thagard (1990, 1992: 7–8, 27–33) and Collier and Levitsky (2009).
- 7 For an exemplary critical analysis of the term 'authoritarianism', as used in the study of political regimes, see Przeworski (2017).
- 8 Though the idea of measurement validity is ubiquitous in the literature on measurement, the distinction between conceptual validity and measurement validity is rarely made; for exceptions, see Jackson and Maraun (1996) and Billiet (2016: 196–200). Yet, inasmuch as the idea that there are theoretical concepts apart from their measures is accepted, as is the case here, this distinction is crucial.
- 9 For exemplary justifications of the concept of democracy, see Dahl (1989) and Saward (1998).
- 10 Inasmuch as some observable property of another object is lawfully related to the observable property of an object under consideration, the observable property of another object could be used as an indicator.
- 11 There is an associated issue that crops up frequently in the measurement of democracy. Scholars have good reasons to want qualitative *and* quantitative distinctions. However, one common practice – the derivation of qualitative distinctions from quantitative distinctions – deserves scrutiny. Indeed, such exercises tend to rely on a rather arbitrary assertion, usually made with little reference to the concept of democracy, that some point on a scale can be treated as the dividing line between democracy and non-democracy. It is preferable to start with qualitative distinctions and then refine these measures by adding quantitative distinctions.
- 12 The problem of combining data also occurs if multiple scores are generated for a single indica-

tor in the same unit (e.g. when multiple coders are used in data based on expert rating) or if data are generated for multiple indicators of the same conceptual property in the same unit (e.g. when a battery of indicators are used to measure some psychological trait). Here we take as our starting point data which can already be treated as data on conceptual properties.

- 13 On reflective and formative aggregation models, see Blalock (1982: ch. 7); Bollen and Lennox (1991); Edwards and Bagozzi (2000); Coltman et al. (2008); Bollen and Bauldry (2011); Edwards (2011).

## REFERENCES

- Adcock, Robert N., and David Collier. 2001. 'Measurement Validity: A Shared Standard for Qualitative and Quantitative Research.' *American Political Science Review* 95(3): 529–46.
- Arblaster, Anthony. 2002. *Democracy*. 3rd ed. Buckingham and Philadelphia: Open University Press.
- Aristotle. 1995 [c. 330 BC]. *Politics*. Oxford: Oxford University Press.
- Bailey, Kenneth D. 1973. 'Monothetic and Polythetic Typologies and Their Relation to Conceptualization, Measurement, and Scaling.' *American Sociological Review* 38(1): 18–32.
- Bailey, Kenneth D. 1994. *Typologies and Taxonomies. An Introduction to Classification Techniques*. Thousand Oaks, CA: Sage.
- Beach, Derek, and Rasmus Brun Pedersen. 2016. *Causal Case Study Methods: Foundations and Guidelines for Comparing, Matching, and Tracing*. Ann Arbor, MI: University of Michigan Press.
- Beetham, David. 1994. 'Key Principles and Indices for a Democratic Audit', pp. 25–43, in David Beetham (ed.), *Defining and Measuring Democracy*. London: Sage.
- Billiet, Jaak. 2016. 'What Does Measurement Mean in a Survey Context?', pp. 193–209, in Christof Wolf, Dominique Joye, Tom W. Smith and Yang-chih Fu (eds), *The SAGE Handbook of Survey Methodology*. Thousand Oaks, CA: Sage.
- Blalock, Hubert M. 1982. *Conceptualization and Measurement in the Social Sciences*. Beverly Hills, CA: Sage.
- Bobbio, Norberto. 1989. *Democracy and Dictatorship: The Nature and Limits of State Power*. Minneapolis, MI: University of Minnesota Press.
- Bobbio, Norberto. 2003. *Teoría general de la política*. Madrid: Editorial Trotta.
- Bollen, Kenneth. 1990. 'Political Democracy: Conceptual and Measurement Traps.' *Studies in Comparative International Development* 25(1): 7–24.
- Bollen, Kenneth A., and Shawn Bauldry. 2011. 'Three Cs in Measurement Models: Causal Indicators, Composite Indicators, and Covariates.' *Psychological Methods* 16(3): 265–84.
- Bollen, Kenneth A., and Richard Lennox. 1991. 'Conventional Wisdom on Measurement: A Structural Equation Perspective.' *Psychological Bulletin* 110(2): 305–14.
- Bollen, Kenneth A., and Pamela Paxton. 1998. 'Detection and Determinants of Bias in Subjective Measures.' *American Sociological Review* 63(3): 465–78.
- Bollen, Kenneth A., and Pamela Paxton. 2000. 'Subjective Measures of Liberal Democracy.' *Comparative Political Studies* 33(1): 58–86.
- Bridgman, P. W. 1927. *The Logic of Modern Physics*. New York, NY: The Macmillan Company.
- Bunge, Mario. 1974a. *Treatise on Basic Philosophy* Vol. 1. *Semantics I: Sense and Reference*. Dordrecht, Holland: D. Reidel Publishing Company.
- Bunge, Mario. 1974b. *Treatise on Basic Philosophy* Vol. 2. *Semantics II: Interpretation and Truth*. Dordrecht, Holland: D. Reidel Publishing Company.
- Bunge, Mario. 1995. 'Quality, Quantity, Pseudo-quantity, and Measurement in Social Science.' *Journal of Quantitative Linguistics* 2(1): 1–10.
- Bunge, Mario. 1998a [1967]. *Philosophy of Science*. Vol. 1. *From Problem to Theory*. New Brunswick, NJ: Transaction Publishers.
- Bunge, Mario. 1998b [1967]. *Philosophy of Science* Vol. 2. *From Explanation to Justification*. New Brunswick, NJ: Transaction Publishers.
- Bunge, Mario. 2012. *Evaluating Philosophies*. New York: Springer.
- Campbell, Donald T., and Donald W. Fiske. 1959. 'Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix.' *Psychological Bulletin* 56(2): 81–105.
- Carnap, Rudolf. 1936. 'Testability and Meaning.' *Philosophy of Science* 3(4): 419–71.

- Carnap, Rudolf. 1937. 'Testability and Meaning.' *Philosophy of Science* 4(1): 1–40.
- Casper, Gretchen, and Claudiu Tufis. 2003. 'Correlation versus Interchangeability: The Limited Robustness of Empirical Finding on Democracy Using Highly Correlated Data Sets.' *Political Analysis* 11(2): 196–203.
- Collier, David, and Robert Adcock. 1999. 'Democracy and Dichotomies: A Pragmatic Approach to Choices about Concepts.' *Annual Review of Political Science* 2: 537–65.
- Collier, David, Fernando Daniel Hidalgo and Andra Olivia Maciuceanu. 2006. 'Essentially Contested Concepts: Debates and Applications.' *Journal of Political Ideologies* 11(3): 211–46.
- Collier, David, Jody LaPorte and Jason Seawright. 2012. 'Putting Typologies to Work: Concept Formation, Measurement, and Analytic Rigor.' *Political Research Quarterly* 65(1): 217–32.
- Collier, David, and Steven Levitsky. 1997. 'Democracy with Adjectives: Conceptual Innovation in Comparative Research.' *World Politics* 49(3): 430–51.
- Collier, David, and Steven Levitsky. 2009. 'Democracy: Conceptual Hierarchies in Comparative Research', pp. 269–88, in David Collier and John Gerring (eds), *Concepts and Method in Social Science: The Tradition of Giovanni Sartori*. New York: Routledge.
- Collier, David, and James E. Mahon. 1993. 'Conceptual "Stretching" Revisited: Adapting Categories in Comparative Analysis.' *American Political Science Review* 87(4): 845–55.
- Coltman, Tim, Timothy M. Devinney, David F. Midgley and Sunil Veniak. 2008. 'Formative versus Reflective Measurement Models: Two Applications of Formative Measurement.' *Journal of Business Research* 61(12): 1250–62.
- Coppedge, Michael, Angel Alvarez and Claudia Maldonado. 2008. 'Two Persistent Dimensions of Democracy: Contestation and Inclusiveness.' *Journal of Politics* 70(3): 632–47.
- Cronbach, Lee Joseph, and Paul E. Meehl. 1955. 'Construct Validity in Psychological Tests.' *Psychological Bulletin* 52(4): 281–302.
- Dahl, Robert A. 1971. *Polyarchy: Participation and Opposition*. New Haven, CT: Yale University Press.
- Dahl, Robert A. 1989. *Democracy and Its Critics*. New Haven, CT: Yale University Press.
- Davidov, Eldad, Bart Meuleman, Jan Cieciuch, Peter Schmidt and Jaak Billiet. 2014. 'Measurement Equivalence in Cross-National Research.' *Annual Review of Sociology* 40: 55–75.
- Edwards, Jeffrey R. 2011. 'The Fallacy of Formative Measurement.' *Organizational Research Methods* 14(2): 370–88.
- Edwards, Jeffrey R., and Richard P. Bagozzi. 2000. 'On the Nature and Direction of Relationships between Constructs and Measures.' *Psychological Methods* 5(2): 155–74.
- Elman, Colin. 2005. 'Explanatory Typologies in Qualitative Studies of International Politics.' *International Organization* 59(2): 293–326.
- Gallie, Walter B. 1956. 'Essentially Contested Concepts.' *Proceedings of the Aristotelian Society* 56: 167–98.
- Gerring, John. 1999. 'What Makes a Concept Good? A Criterial Framework for Understanding Concept Formation in the Social Sciences.' *Polity* 31(3): 357–93.
- Gerring, John, Daniel Pemstein and Svend-Erik Skaaning. 2019. 'An Ordinal, Concept-driven Approach to Measurement: The Lexical Scale.' *Sociological Methods and Research*. DOI: <https://doi.org/10.1177/0049124118782531>
- Goertz, Gary. 2006. *Social Science Concepts: A User's Guide*. Princeton, NJ: Princeton University Press.
- Gray, John. 1978. 'On Liberty, Liberalism and Essential Contestability.' *British Journal of Political Science* 8(4): 385–402.
- Greco, Salvatore, Alessio Ishizaka, Menelaos Tasiou and Gianpiero Torrisi. 2019. 'On the Methodological Framework of Composite Indices: A Review of the Issues of Weighting, Aggregation, and Robustness.' *Social Indicators Research* 141(1): 61–94.
- Gwet, Kilem L. 2014. *Handbook of Inter-Rater Reliability*. 4th ed. Gaithersburg, MD: Advanced Analytics.
- Harré, Rom. 1981. *Great Scientific Experiments: Twenty Experiments that Changed Our View of the World*. Oxford: Phaidon Press.
- Herrera, Yoshiko, M. and Devesh Kapur. 2007. 'Improving Data Quality: Actors, Incentives, and Capabilities.' *Political Analysis* 15(4): 365–85.
- Jackson, Jeremy S. H., and Michael Maraun. 1996. 'The Conceptual Validity of Empirical

- Scale Construction: The Case of the Sensation Seeking Scale.' *Personality and Individual Differences* 21(1): 103–10.
- Kaplan, Abraham. 1964. *The Conduct of Inquiry: Methodology for Behavioral Science*. Scranton, PA: Chandler Publishing Co.
- King, Gary, Robert O. Keohane and Sidney Verba. 1994. *Designing Social Inquiry: Scientific Inference in Qualitative Research*. Princeton, NJ: Princeton University Press.
- Lange, Matthew. 2013. *Comparative-Historical Methods*. Los Angeles, CA: Sage.
- Laudan, Larry. 1977. *Progress and Its Problems: Toward a Theory of Scientific Growth*. Berkeley: University of California Press.
- Lazarsfeld, Paul F. 1958. 'Evidence and Inference in Social Research.' *Daedalus* 87(4): 99–130.
- Lazarsfeld, Paul F., and Allen H. Barton. 1951. 'Qualitative Measurement in the Social Sciences: Classification, Typologies, and Indices', pp. 155–92, in Daniel Lerner and Harold D. Lasswell (eds), *The Policy Sciences: Recent Developments in Scope and Method*. Stanford, CA: Stanford University Press.
- Lazarsfeld, Paul F., and Herbert Menzel. 1961. 'On the Relation between Individual and Collective', pp. 422–40, in Amitai Etzioni (ed.), *Complex Organizations: A Sociological Reader*. New York: Holt, Rinehart and Winston.
- Linz, Juan J. 1975. 'Totalitarianism and Authoritarian Regimes', pp. 175–411, in Fred Greenstein and Nelson Polsby (eds), *Handbook of Political Science* Vol. 3, *Macropolitical Theory*. Reading, MA: Addison-Wesley Press.
- Locke, Richard M., and Kathleen Thelen. 1995. 'Apples and Oranges Revisited: Contextualized Comparisons and the Study of Comparative Labor Politics.' *Politics and Society* 23(3): 337–67.
- Lupia, Arthur and Colin Elman. 2014. 'Openness in Political Science: Data Access and Research Transparency.' *PS: Political Science and Politics* 47(1): 19–42.
- Lustick, Ian S. 1996. 'History, Historiography, and Political Science: Multiple Historical Records and the Problem of Selection Bias.' *American Political Science Review* 90(3): 605–18.
- Marradi, Alberto. 1990. 'Classification, Typology, Taxonomy.' *Quality & Quantity* 24(2): 129–57.
- McAdam, Doug, Sidney Tarrow and Charles Tilly. 2008. 'Methods for Measuring Mechanisms of Contention.' *Qualitative Sociology* 31(4): 307–31.
- McMann, Kelly M., Daniel Pemstein, Brigitte Seim, Jan Teorell and Staffan I. Lindberg. 2016. *Strategies of Validation: Assessing the Varieties of Democracy Corruption Data*. *V-Dem Working Paper* 2016: 23.
- Munck, Gerardo L. 2009. *Measuring Democracy: A Bridge between Scholarship and Politics*. Baltimore, MD: The Johns Hopkins University Press.
- Munck, Gerardo L. 2016. 'What Is Democracy? A Reconceptualization of the Quality of Democracy.' *Democratization* 23(1): 1–26.
- Munck, Gerardo L., and Jay Verkuilen. 2002. 'Conceptualizing and Measuring Democracy: Evaluating Alternative Indices.' *Comparative Political Studies* 35(1): 5–34.
- Møller, Jørgen, and Svend-Erik Skaaning. 2014. *The Rule of Law: Definitions, Measures, Patterns and Causes*. New York: Palgrave Macmillan.
- Møller, Jørgen, and Svend-Erik Skaaning. 2019. 'The Ulysses Principle: A Criterial Framework for Reducing Bias When Enlisting the Work of Historians.' *Sociological Methods and Research*. DOI: <https://doi.org/10.1177/0049124118769107>
- Nardo, Michela, Michaela Saisana, Andrea Saltelli, Stefano Tarantola, Anders Hoffman and Enrico Giovannini. 2005. *Handbook on Constructing Composite Indicators*. Paris: OECD Publishing.
- Ogden, C. K., and I. A. Richards. 1923. *The Meaning of Meaning: A Study of the Influence of Language upon Thought and the Science of Symbolism*. London: Routledge.
- Przeworski, Adam. 2017. 'A Conceptual History of Political Regimes: Democracy, Dictatorship, and Authoritarianism.' *Studia Socjologiczno-Polityczne. Seria Nowa* 7(2): 9–30.
- Przeworski, Adam, Michael E. Alvarez, José Antonio Cheibub and Fernando Limongi. 2000. *Democracy and Development: Political Institutions and Well-Being in the World, 1950–1990*. New York: Cambridge University Press.
- Przeworski, Adam, and Henry Teune. 1970. *The Logic of Comparative Social Inquiry*. New York: Wiley.



- Ríos-Figueroa, Julio, and Jeffrey Staton. 2012. 'An Evaluation of Cross-National Measures of Judicial Independence.' *Journal of Law, Economics, and Organization* 30(1): 104–37.
- Sartori, Giovanni. 1970. 'Concept Misformation in Comparative Politics.' *American Political Science Review* 64(4): 1033–53.
- Sartori, Giovanni. 1987. *The Theory of Democracy Revisited Part 1: The Contemporary Debate*. Chatham, NJ: Chatham House Publishers.
- Sartori, Giovanni. 2009 [1975]. 'The Tower of Babel', pp. 61–96, in David Collier and John Gerring (eds), *Concepts and Method in Social Science: The Tradition of Giovanni Sartori*. New York: Routledge.
- Sartori, Giovanni. 2009 [1984]. 'Guidelines for Concept Analysis', pp. 97–150, in David Collier and John Gerring (eds), *Concepts and Method in Social Science: The Tradition of Giovanni Sartori*. New York: Routledge.
- Saward, Michael. 1998. *The Terms of Democracy*. Cambridge: Polity Press.
- Schedler, Andreas. 2012. 'Judgment and Measurement in Political Science.' *Perspectives on Politics* 10(1): 21–36.
- Schumpeter, Joseph A. 1942. *Capitalism, Socialism, and Democracy*. New York: Harper and Brothers.
- Seawright, Jason, and David Collier. 2014. 'Rival Strategies of Validation: Tools for Evaluating Measures of Democracy.' *Comparative Political Studies* 47(1): 111–38.
- Skaaning, Svend-Erik. 2018. 'Different Types of Data and the Validity of Democracy Measures.' *Politics and Governance* 6(1): 105–16.
- Teorell, Jan. 2010. *Determinants of Democratization*. New York: Cambridge University Press.
- Thagard, Paul. 1990. 'Concepts and Conceptual Change.' *Synthese* 82(2): 255–74.
- Thagard, Paul. 1992. *Conceptual Revolutions*. Princeton, NJ: Princeton University Press.
- Tilly, Charles. 2008. 'Describing, Measuring, and Explaining Struggle.' *Qualitative Sociology* 31(1): 1–13.
- van Deth, Jan W. (ed.). 1998. *Comparative Politics: The Problem of Equivalence*. London: Routledge.
- Vanoli, André. 2005. *A History of National Accounting*. Amsterdam: IOS Press.
- Waldron, Jeremy. 2002. 'Is the Rule of Law an Essentially Contested Concept (in Florida)?' *Law and Philosophy* 21(2): 137–64.

# Measurement Models

Christopher J. Fariss, Michael R. Kenwick  
and Kevin Reuning

## INTRODUCTION

Measurement models in general, and latent variable models in particular, are now common in political science research. This is because political scientists are increasingly focused on improving the measurement of unobservable concepts and understanding the relationships and potential biases between different pieces of observable information and the measurement procedures that link this information to theoretical concepts. Recent methodological and computational advances have led to a flourishing of new latent variable modeling applications. These new tools provide researchers with a means of measuring difficult to observe concepts based on events, ratings or other pieces of observable information that are assumed to be a result of the underlying unobservable latent trait.<sup>1</sup>

Latent variable models are built on the idea that observable variables are manifestations of an underlying conceptual process that is not perfectly observable or knowable

and includes increasingly computationally sophisticated probability models (e.g., Imai et al., 2016; Jackman, 2000, 2001; Martin and Quinn, 2002; Plummer, 2017; Carpenter et al., 2017) and computationally simply additive scales (e.g., Guttman, 1949; van Schuur, 2003). In this chapter, we review the scientific measurement process and the assumptions needed to construct models of unobservable theoretical concepts.

The scientific process of measurement occurs in three iterative stages: *conceptualization* of the sociological or physical system being studied, *operationalization* of the data generating process that approximates the system and *empirical analysis* of the data generated by that system. The relationship between each of these steps is assessed using construct validity tools.<sup>2</sup> Because the measurement process is iterative, it is incumbent on the researcher to (1) acknowledge the starting point of the measurement process and (2) provide an assessment of the quality of the links between these steps. We provide

more details about these recommendations throughout this chapter, although our focus here is on how latent variable models can be used to assess these steps.

Latent variable models allow for the empirical assessment of how the different observed pieces of data relate to one another through their association with the estimated latent trait. Even computationally simple additive scales are models that represent an underlying latent concept. Additive scales require the same process of assessment as more computationally difficult latent variable approaches (van Schuur, 2003). We discuss these additive scaling models as a starting point for thinking about estimating latent variable models more generally, because these models share the same set of assumptions. New computationally sophisticated latent variable models allow the researcher to relax these assumptions in conceptually meaningful ways.

The particular examples of latent variable models that we review in this chapter have been applied across a variety of subfields, encompassing the study of political ideology (Barbera, 2015; Bond and Messing, 2015; Martin and Quinn, 2002; Martin et al. 2005; Caughey and Warshaw, 2015; König et al., 2013; Pan and Xu, 2018; Treier and Hillygus, 2009; Windett et al., 2015), political attitudes, knowledge and preferences (Blaydes and Linzer, 2008; Pérez, 2011; Jessee, 2017; Stegmueller, 2011, 2013), regime institutions (Treier and Jackman, 2008; Pemstein et al., 2010; Kenwick, 2018, Gandhi and Sumner, 2019), UN voting positions (Voeten, 2000), human rights abuse (Schnakenberg and Fariss, 2014; Fariss, 2014, 2019; Fariss et al., 2020), human rights treaty embeddedness (Fariss, 2018b,a), judicial independence (Linzer and Staton, 2016), demographic variables (Anders et al., forthcoming), and institutional transparency (Hollyer et al., 2014). We discuss several latent variable models that are capable of accommodating different forms of conceptual dependencies between units, in particular temporal interdependence in time-series cross-sectional data.

We provide examples that build on insights from a recently published article on temporal dependence and sudden temporal changes in time-series cross-sectional data (Reuning et al., 2019).<sup>3</sup>

After discussing the measurement process and construct validity in more detail and laying out different dynamics of latent variables, we highlight places that we believe are ripe for future research. In particular, we discuss new ways to theoretically include time in latent variable models, ways to scale expert surveys, the use of Multiple-Indicator-Multiple-Causes models and issues with different model fit statistics. Finally, we end with a list of recommendations for the applied researcher using latent variable models.

## THE MEASUREMENT PROCESS

The process of measurement can be broadly characterized as having three steps.<sup>4</sup> The process of measurement allows the researcher to think explicitly about each of these three steps and the relationships between them because it links theories, the *concept*, with operational procedures, the *construct*, which generate observable information, the *data*. We discuss each of these steps here.

In the first step, a researcher generates a systematized definition of a concept in which they are interested. The systematized definition should be specific enough to have intellectual traction, but sufficiently broad so that it can be meaningfully applied to a set of objects across time, space or both (Shadish, 2010). What does this mean in practice? That there is necessarily a trade-off between specificity and generalizability and, when applied, the researcher must clarify the boundary conditions that define the set of objects for which the measurement procedure operates and the set for which it does not. At the extreme, the conceptual process should cover more than one object, but less than all objects. Specifying these boundary conditions is part

of the conceptual step in the measurement process. However, because the measurement process is iterative, the researcher can and should return to this first step in order to make refinements to the systematized definition based on information obtained in the second or third step of the process.

Often in political science, even a well-defined concept cannot be directly observed in the real world. In the second step, the researcher must therefore begin to identify how the latent trait relates to observable information, thereby creating a data generating process from the latent trait to the observed indicators. A researcher interested in democracy might, e.g., identify whether a country holds competitive elections, whether there is a representative legislature with the ability to effectively pass legislation and whether there has been alternation in power among competing political groups. Thus, this second step involves the critical task of designing the data generating procedures used to collect information that relates to the underlying concept of interest for the objects under study.

Once the data generating procedures are defined, the researcher proceeds to the third step, which involves collecting observational information about a set of objects and the categorization or scoring of those objects. This process maps the observed information collected about the objects in the second step back to the concept of interest defined in the first step through a defined categorization or scoring procedure. The definitional rules of the operational procedure should be consistent with the conceptual definition defined in the first step. The creation and use of any operational protocol requires that researchers make decisions about how to weight each piece of information and how they individually or jointly inform the researcher's beliefs about an object's score for the underlying trait.

In sum, the three steps are: (1) define theoretical concept and scope; (2) identify how observational data connects to the theoretical concept by defining the data generating

process; (3) use the operational procedure to categorize or score cases which are the subjects or units of study. Most of our discussion from here focuses on the second and third steps. This procedure highlights the fact that all measurement inherently involves the creation of a measurement model, which is the second step of the measurement process, but with links to both the first and third steps. Like all other models in social science, those used in measurement require careful validation about the relationships between steps.

At the broadest level, measurement validation centers upon what is known as construct validity, which is an assessment of both the theoretical content of the operationalization protocol and the empirical content that is believed to be captured by this construct (e.g., Adcock and Collier, 2001; Jackman, 2008; Shadish, 2010; Shadish et al., 2001). Construct validity encompasses a variety of different ways to evaluate a measure and operationalization.

Two important parts of construct validity are translation validity and measurement validity. Translation validity is an evaluation of the match between the theoretical construct and the proposed data generating procedure, which generates the observed pieces of information. Measurement validity is an evaluation of the fit between the proposed data generating procedure and the actual data obtained from it.

Translation errors occur when the operational protocol does not match the theory of the concept. Measurement errors occur when the fit between representation of the data generating procedure (the measurement model) and the data is poor. As researchers validate their measures along these two related criteria, they may choose to (1) update the types of information to collect, (2) modify the method for linking this information into scores on the latent trait or (3) modify the theoretical concept that the data generating procedure is derived from. The measurement process is an inherently iterative process between each of the three steps outlined above.

Thus, to generate good estimates of a theoretical concept of interest, the research must understand the relationship between each part of the measurement process.

## MEASUREMENT MODELING ASSUMPTIONS

All measurement models, regardless of their complexity, require assumptions about the underlying trait. In this section we provide an overview of these assumptions for some of the measurement models that are most commonly used in the social sciences (additive scales and IRT models). We begin by discussing the assumptions of additive scales, proceed to identify assumptions of latent variable models and finally provide an overview of latent variable model assumptions about dynamics and their relationship to local independence.

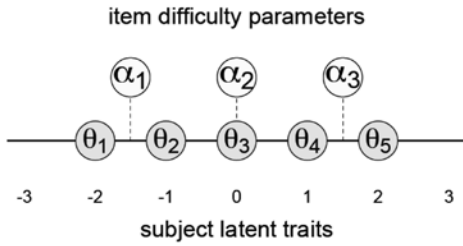
Before proceeding, it is useful to provide a brief overview of the notation we will use in the following section. We denote the latent trait as  $\theta$ , which is observed across units indexed with  $i$ , which takes on values of 1, 2, ...,  $N$ , where  $N$  is the total number of units in the sample. We observe  $\theta$  indirectly through observable pieces of information often referred to as 'items' or 'manifest indicators', each of which is indexed using  $k$  with values 1, 2, ...,  $K$ , where  $K$  is the total number of manifest indicators. The realized values of these indicators are  $y$ , with  $Y$  acting as the manifest indicator yet to be observed. This notation lets us refer empirically to both the potential observed realization of data  $Y$  and the actual realization of data  $y$ . Formally, we let  $Y_{ik}$  denote the score of subject  $i$  on item  $k$ , a random variable with realization  $y_{ik} = \{0, 1\}$ . For simplicity, we assume that each indicator is binary. In the next section, we will continue to build towards an additive scale as a latent variable representation of a concept. We also discuss the assumptions underlying this model and the standard unidimensional item

response theory models which we review later in the chapter.

### *Assumptions of Additive Scale Measurement Models*

To make the notation and formalizations presented in this section more clear, we introduce a small deterministic example that illustrates the relationships between the different model parameters and data. As we mentioned above, we let  $k$  take on integer values from 1, 2, 3, which represents three distinct questions of varying ability that we will ask of five hypothetical subjects. These are the items which generate responses (i.e., the item responses) from each subject. We first introduce a new parameter  $\alpha_k$  which represents a feature of the items. In a testing setting,  $\alpha_k$  parameters represent the difficulty of a particular question as it relates to the ability of the test-takers or subjects, which is represented by  $\theta$ . In additive scales it is assumed that if the latent trait for unit  $i$  is greater than  $\alpha_k$  then we will observe  $y_i = 1$ . More generally,  $\alpha_k$  accounts for the variation in how high (or low) a unit has to be on the latent trait to achieve a positive outcome for indicator  $y_k$ . For this example, we are supposing that we know the true values of this parameter in our measurement model. Later on, we will estimate these parameters.

In our example we consider the following latent traits for five units ( $\theta_1 = -2$ ,  $\theta_2 = -1$ ,  $\theta_3 = 0$ ,  $\theta_4 = 1$ ,  $\theta_5 = 2$ ) and three items ( $\alpha_1 = -1.5$ ,  $\alpha_2 = 0$ ,  $\alpha_3 = 1.5$ ), which are all arrayed along the same unidimensional line. The relationship between the five units and the three items are displayed visually in Figure 20.1. The unidimensional line represents values of the unobservable theoretical concept of interest but the substantive meaning of the entities along the line differ because some are subjects and the others are the data generating objects (i.e., the item or test questions).



**Figure 20.1 Latent variables and item parameters**

Note: This plot displays latent traits for 5 units  $\theta_1 = -2, \theta_2 = -1, \theta_3 = 0, \theta_4 = 1, \theta_5 = 2$  and 3 items  $\alpha_1 = 1.5, \alpha_2 = 0, \alpha_3 = 1.5$  all arrayed along the same unidimensional line. The unidimensional line represents values of the unobservable theoretical concept of interest but the substantive meaning of the entities along the line differ because some are subjects and others are the data generating indicators (i.e., the item responses generated by the subjects). The subjects and items are comparable in this space however. In particular, the comparison of the distance between subject and object determines the observed binary item responses for each subject-object pairing.

The relationships displayed visually in Figure 20.1 are unobserved. What we actually observed are binary responses (e.g., the answers to questions generated by subjects or the categorical values created to compare country-year units). Our measurement goal is to create a test or categorization scheme that relates the observed data back to the unobserved latent traits. This is done by assuming a data generating process from the latent trait to the indicators. Here we will use a deterministic function for the relationship between each subject-item pairing, which is displayed

in Equation 1. Later on we will introduce a probability model for accomplishing this task.

$$y_{ik} = \begin{cases} 1 & \text{if } \theta_i > \alpha_k \\ 0 & \text{if } \theta_i \leq \alpha_k \end{cases} \quad (1)$$

Equation 1 represents the data generating function for the binary item responses produced for each subject-item pair. For the illustrative example,

$$y_i^+ = \sum_k^K (y_{ik}) \quad (2)$$

Equation 2 represents the observed additive scale value for each subject  $i$ , which is determined by the value of the logical proposition in equation 1. Table 20.1 presents the additive scale values for  $y_i^+$  based on the pairwise comparisons between the five subjects and the three items. The additive scale is a deterministic, continuous scale, which satisfies the conditions outlined by Guttman (e.g., Guttman, 1949; van Schuur, 2003). In words, the first subject's ability is always less than the value of the item. To reiterate, the values are substantively distinct but are comparable together on the same latent scale.

The additive scale can also be rewritten as a function of just the values of the latent trait and the difficulties. This is the function in Equation 3, where the additive value is found by checking the latent trait's value against the ordered alphas. This emphasizes that in additive scales there is an assumption that all items can be ordered in such a way that

**Table 20.1 Example of additive scale function**

Latent Trait	Items			Additive Scale
$\theta_i$	$\alpha_1 = -1.5$	$\alpha_2 = 0$	$\alpha_3 = 1.5$	$y_i^+$
$\theta_1 = -2$	$\theta_1 \leq \alpha_1 \Rightarrow +0$	$\theta_1 \leq \alpha_2 \Rightarrow +0$	$\theta_1 \leq \alpha_3 \Rightarrow +0$	$y_1^+ = 0$
$\theta_2 = -1$	$\theta_2 > \alpha_1 \Rightarrow +1$	$\theta_2 \leq \alpha_2 \Rightarrow +0$	$\theta_2 \leq \alpha_3 \Rightarrow +0$	$y_2^+ = 1$
$\theta_3 = 0$	$\theta_3 > \alpha_1 \Rightarrow +1$	$\theta_3 \leq \alpha_2 \Rightarrow +0$	$\theta_3 \leq \alpha_3 \Rightarrow +0$	$y_3^+ = 1$
$\theta_4 = 1$	$\theta_4 > \alpha_1 \Rightarrow +1$	$\theta_4 > \alpha_2 \Rightarrow +1$	$\theta_4 \leq \alpha_3 \Rightarrow +0$	$y_4^+ = 2$
$\theta_5 = 2$	$\theta_5 > \alpha_1 \Rightarrow +1$	$\theta_5 > \alpha_2 \Rightarrow +1$	$\theta_5 > \alpha_3 \Rightarrow +1$	$y_5^+ = 3$

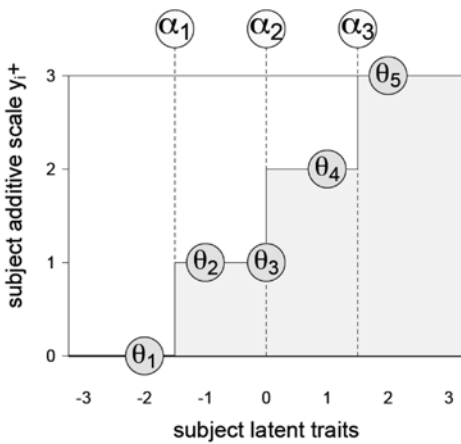
Note: The additive scale values are based on the status of the logical propositions for each subject-item comparison.

they are monotonically increasing in difficulty. Once ordered, a researcher can identify where a unit is on the additive scale based on when its indicators switch from 1 to 0.

$$y_i^+ = \begin{cases} 3 & \text{if } \theta_i > \alpha_3 \\ 2 & \text{if } \theta_i > \alpha_2 \quad \text{and } \theta_i \leq \alpha_3 \\ 1 & \text{if } \theta_i > \alpha_1 \quad \text{and } \theta_i \leq \alpha_2 \\ 0 & \text{if } \theta_i \leq \alpha_1 \end{cases} \quad (3)$$

We can visually represent the relationship between the values of the additive scale, the latent trait, and the items in Equation 3. We do this in Figure 20.2.

Up until now, we have assumed a deterministic model between the observed items and the latent trait, which are consistent with the assumptions from Guttman (1949). In later measurement research, Mokken (1971) developed a stochastic version under the assumptions of a unidimensional latent variable, latent monotonicity and local independence. Under these assumptions, the



**Figure 20.2 Example of additive scale function**

*Note:* This plot displays latent traits for 5 units  $\theta_1 = -2$ ,  $\theta_2 = -1$ ,  $\theta_3 = 0$ ,  $\theta_4 = 1$ ,  $\theta_5 = 2$  and 3 items ( $\alpha_1 = 1.5$ ,  $\alpha_2 = 0$ ,  $\alpha_3 = 1.5$ ) all arrayed along the same unidimensional line displayed in Figure 20.1. The additive scale values on the y-axis are based on the status of the logical propositions for each subject-item comparison in Table 20.1.

proportion of ‘correct’ answers by subject  $i$  to item  $k$  is nondecreasing in the sum of all the items. These assumptions also imply that all of the items are positively correlated across all subsets of subjects (Mokken, 1971). Under these assumptions the unweighted sum of the variables increase as  $\theta$  increases. Mokken Scaling Analysis (MSA) is simply a stochastic version of a Guttman scale, in which items measure a single latent construct and can be ordered by difficulty (Guttman, 1949) but are not assumed to be generated without error (van Schuur, 2003).

The assumptions made by Mokken (1971) are common across many latent variable models and so are worth exploring in more depth. The first assumption is that  $\theta$  is a *unidimensional latent variable*, which means that the values of the latent trait reside on a single axis. This assumption can be tested using parameters from the Mokken Scaling Analysis (MSA) model (van Schuur, 2003). If this assumption fails, it means that the latent trait cannot be collapsed into a single dimension but that units can be high in one dimension and low on another.

The second assumption is of *latent monotonicity*, which means that the item step response function is strictly increasing on  $\theta$ ;  $\theta_1 \leq \theta_2 \Rightarrow P(Y_{ik} \geq y_{ik} | \theta_1) \leq P(Y_{ik} \geq y_{ik} | \theta_2)$ . This implies that as a unit increases in the latent variable, the probability of observing a positive indicator also increases.

The third assumption is of *local independence*, which means that the item responses are not deterministically related to each other outside of their relationship to the latent trait. This implies that the probability of the set of each subject’s item responses is

$$P(Y_{i1} = y_{i1}, Y_{i2} = y_{i2}, \dots, Y_{iK} = x_{iK} | \theta_i) = \prod_{k=1}^K P(Y_{ik} = y_{ik} | \theta_i) \quad (\text{van Schuur, 2003}).$$

The only relationship between items is through their relationship with the latent variable. This can be violated in the testing environment when getting one answer correct depends on getting previous answers correct.

To summarize, additive scaling is a data generating procedure that maps the latent trait to an additive index. In order to estimate a stochastic additive scale, researchers must make assumptions about unidimensionality, monotonicity and local independence. As we discuss next, these assumptions are also present in more complicated latent variable models which also allow more variation in how the latent trait relates to the observed indicators.

### Identification Assumptions of Latent Variable Models

We now move to estimate  $\theta$  itself because, up until this point, this parameter has been entirely conceptual. We do this through the Item Response Theory (IRT) framework which allows us to estimate  $\theta$  as well as other parameters in the data generating process. In addition, using this framework we can add an additional layer of complexity of cross-sectional time-series data (i.e., country-year units) instead of the five hypothetical subjects from before.

In principal, IRT models are rooted in the same assumptions as the additive scale above; that is, we assume that  $\theta$  is a *unidimensional latent variable* and that its relationship with its associated items is characterized by *latent monotonicity* and *local independence*.

Under the IRT framework, the latent trait is  $\theta_i$  where the subscript  $i = 1, \dots, N$  indicates multiple units.  $y_{ik}$  is the observed value for item  $k$  for unit  $i$ . For each item  $\alpha_k$  and  $\beta_k$  are also estimated.  $\alpha_k$  continues to act as a ‘difficulty’ parameter, or a threshold that benchmarks how likely an indicator is to be observed relative to the values of the latent trait. In our formulation, this is analogous to an intercept in a traditional logistic regression model.  $\beta_k$  is often referred to as the ‘discrimination’ parameter and is the analogue of a slope coefficient.

The relationship between  $\theta_i$  and our indicator  $y_{ik}$  is:

$$P(y_{ik} = 1) = \Lambda(\alpha_k - \beta_k \theta_i) \quad (4)$$

where  $\Lambda$  is the logistic function. Unlike in the case of the additive scale, this is necessarily probabilistic.<sup>5</sup> The likelihood function encompassing the latent trait, realizations of the manifest indicators and item-specific parameters take the following form:

$$\mathcal{L} = \prod_{i=1}^N \prod_{k=1}^K \Lambda(\alpha_k - \beta_k \theta_i)^{y_{ik}} (1 - \Lambda(\alpha_k - \beta_k \theta_i))^{1-y_{ik}}$$

The model estimates the placement of one unit relative to all the other units based on the values of the observed items. Without additional information such models are not identified, which means that estimation is not possible because multiple sets of values for the parameter estimates will fit the data equally well. There are generally three types of identification problem that most applied researchers will encounter: additive, scale and rotational. In each of these cases the likelihood is invariant across multiple parameter estimates. To prevent this situation, the researcher must make several benign assumptions that provide additional information to the model and prevent invariance.

The issues of scale and additive invariance are often the easiest to solve. In the case of additive invariance,  $\theta + \delta$  and  $\alpha - \delta$  lead to equivalent likelihood for any  $\delta$ . Scale invariance is similar except is a result of multiplication:  $\delta \cdot \theta$  and  $\frac{\theta}{\delta}$  would again produce equivalent likelihoods. This invariance is commonly solved by providing information to  $\theta$  through a standard normal distribution as the prior. This is useful as it leads to estimates of  $\theta$  that are mean 0 with a standard deviation of 1.

Rotational invariance can be more complicated. Rotational invariance is the result of equivalent likelihoods that result when  $\theta$  is multiplied by  $-1$  or ‘flipped’. In the context

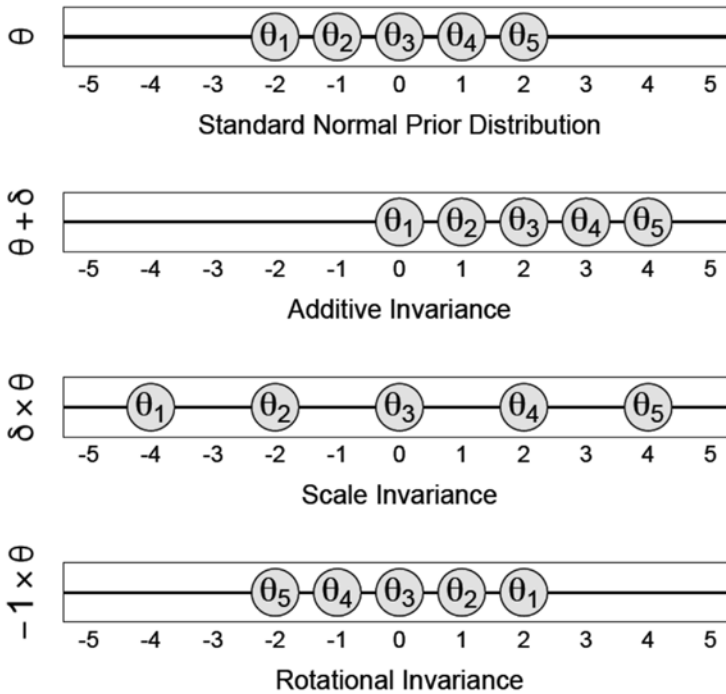


of a latent variable for ideology, estimates with negative numbers as conservative and positive numbers as liberal are the same as when negative numbers or liberal and positive numbers are conservative. Put differently, the model has no way of knowing whether to order the units from liberal to conservative, or from conservative to liberal ideologies.<sup>6</sup>

One simple strategy for resolving rotational invariance is to fix the values of the latent trait for two or more units. In the political ideology example, this could be achieved by assigning values of the latent trait for a very liberal and a very conservative individual. An alternative strategy imposes assumptions about the relationship between

the manifest indicators and the latent trait through the discrimination parameters,  $\beta_k$ . For example, Fariss (2014) relies on a series of indicators believed to positively correlate with respect for human rights, and therefore restricts the  $\beta$  parameters to take on positive values. In practice, this can be done through the use of truncated distributions (e.g., half-normal) or strictly positive distributions (e.g., gamma).

As a demonstration of issues of invariance, consider the simple single dimensional model for five units. We plot these five units along a single dimension in Figure 20.3. The first row shows the baseline, placing all five units in order. The second row shows



**Figure 20.3 Identification issues in latent variables**

*Note:* This plot displays latent traits from four idealized models. The top row displays the 5 units scaled so that the mean value is 0. The other rows show the consequence of the values of the latent trait when adding a constant (row 2), multiplying a constant (row 3), and multiplying by  $-1$  (row 4). These models each provide the same values for comparisons of the value of one unit relative to any other or to the mean value of all of the units. Since we do not know the true absolute value of the concept we wish to make inferences about, it is useful to constrain the values of the latent trait to occupy the standard normal density function. By constraining the model in this way, we ensure that we are not mixing and therefore comparing values from the other models represented in this visualization.

a rightward shift of all five units (additive invariance). Since the latent dimension is arbitrary, this move does not matter as long as all units move in a similar way and there are no assumptions made about where the center of the latent space is.

In row 3 we demonstrate the issue of scale invariance. Here, the latent trait has been multiplied by 2, expanding the latent scale. Again, because each unit moves equally the end result is no different from the initial placement in row 1 if there is no constraint placed on the scale of the latent trait. Finally, row 4 shows rotational invariance. The latent traits have been reversed so that  $\theta_1$  moves from 2 to  $-2$ . This is equivalent to the first row if there is no constraint placed on the direction of the scale.

In our running examples, we place normal priors on the latent trait and resolve the issues of location and scale invariance.<sup>7</sup> To resolve rotational invariance, we constrain  $\beta_k$  to be greater than zero, such that increasing values of each manifest indicator are associated with increasing values of the latent trait. Finally, we place weakly informative normal priors on the difficulty parameters. The prior assignments can therefore be expressed as:

$$\begin{aligned}\theta_i &\sim N(0,1) \quad \forall i = 1, \dots, N \\ \beta_k &\sim \text{HN}(0,3) \\ \alpha_k &\sim N(0,3)\end{aligned}$$

where HN is the half-normal distribution, with support on  $[0, \infty)$ .

### **Local Independence and Assumptions about Dynamics**

The model described above can be expanded to include units over multiple time periods. In the above equations, this is accommodated by replacing  $\theta_i$  with  $\theta_{it}$  where  $t$  indexes time periods from 1, ...,  $T$ . There is no requirement that all units must be observed over all time periods.

This does lead to some methodological questions. Latent variable models, including simple additive and cumulative scales, are built on the assumption that each observed variable for a unit is generated independently of the other observed pieces of information about that unit. This is the assumption of *local independence*. For the type of cross-sectional time-series data that we consider in this chapter, the assumption of local independence means that any two observed variables are *only* related because of the fact that they are each an observable outcome of the same latent variable.

There are three relevant local independence assumptions: (1) local independence of different indicators within the same country-year; (2) local independence of indicators across countries within years; and (3) local independence of indicators across years within countries. Priors are a useful and common means of addressing potential violations of the lattermost type of local independence violations. Applied researchers in international relations are likely to encounter problems where they are attempting to estimate a measure of multiple units observed over time. The dependencies within a unit across time can be modeled as part of the prior on the latent variable. In this section we discuss three broad approaches in the field. Two of these are relatively common, while the last has been recently introduced. In each case we discuss the assumptions that the model makes, the benefits of it and the costs.

#### **Static model**

The three modeling strategies we present are differentiated by the prior information assigned to the latent variable. We start here with the simplest model, the static model. The static model places a standard normal prior on all units for all time periods:

##### *Static model prior*

$$\theta_{it} \sim N(0,1) \quad \forall i = 1, \dots, N \quad \forall t = 1, \dots, T$$

The standard normal prior, as discussed above, prevents additive and scale invariance. Estimates for the latent trait for each unit in

each time period are differentiated exclusively by the values of the indicators for that unit at that time period. This model treats each unit-time period as independent, which is a bold assumption to make in most applied research. In addition, this limits the information that is being used to estimate the latent trait and so is likely to increase credible intervals. In the case where the indicator variables contain sufficient information on the latent trait, this modeling strategy may not be problematic. Unfortunately, this is seldom the case when using social science data, where indicators are often coarse or missing. As a result, these indicators often do not contain sufficient information to differentiate between theoretically distinct units. The benefit to this approach is that it does not force any atheoretical ‘memory’ on the latent trait allowing sudden changes in the latent trait across time-periods.

### *Standard dynamic model*

To address temporal non-independence in the data, many researchers have used a dynamic prior for the latent trait, where the latent trait for unit  $i$  in time  $t$  is related directly to the latent trait for unit  $i$  at time  $t - 1$  (Martin and Quinn, 2002; Schnakenberg and Fariss, 2014; Fariss, 2014; Caughey and Warshaw, 2015; Kōnig et al., 2013). The choice of a ‘random walk’ prior on the latent variable is particularly common.

The random walk approach begins with the use of a standard normal prior on the latent trait in the first observation period for every unit. Then for each subsequent time period, the prior is normally distributed with mean  $\theta_{i(t-1)}$ , and a standard deviation  $\sigma$  which is either assigned by the researcher or, more commonly, estimated from the data.<sup>8</sup> Here, we assign a weakly informative prior to  $\sigma$  by using a half-normal distribution with standard deviation of 3 and mean 0.

#### *Standard dynamic model priors*

$$\begin{aligned} \theta_{i1} &\sim N(0,1) \quad \forall i = 1, \dots, N \\ \theta_{it} &\sim N(\theta_{i(t-1)}, \sigma) \quad \forall i = 1, \dots, N \quad \forall t = 2, \dots, T \\ \sigma &\sim \text{HN}(0,3) \end{aligned}$$

This strategy trades the assumption that observations are independent with the assumption that the latent trait will be correlated over time and will follow a random walk. As a result, estimates from dynamic models typically have less uncertainty because more information is used to estimate each latent variable. This also induces smoothing over time because changes between time periods are constrained. When researchers have theoretical reasons to expect that the latent trait is relatively slow-moving over time, both modeling features can be desirable. If, however, the latent trait is subject to rapid fluctuations or state changes between time periods, this temporal smoothing can produce biased estimates. The modeling strategy we introduce below is designed to address this problem while still accounting for temporal dynamics.

### *Robust dynamic modeling*

We recently proposed an alternative strategy that drew on the robust modeling literature to implement a robust version of the dynamic modeling (Reuning et al., 2019). In the Bayesian framework, robust models alternate normal distributions with the Student’s  $t$ -distribution to account for outliers (Gelman et al., 2014; Lange and Sinsheimer, 1993; Lange et al., 1989; Geweke, 1993; Fonseca et al., 2008). In the context of dynamic latent variables, potential outliers are the ‘shocks’ where values of the true latent variable change suddenly within a unit’s time series.

The robust dynamic model continues to use a standard normal distribution for the first observation in a unit’s time series.<sup>9</sup> In subsequent years, the prior follows a Student’s  $t$ -distribution with four degrees of freedom. Setting the degrees of freedom to a relatively low value increases the density of the tails of the distribution, which allows ‘extreme values’ to be estimated from time period to time period. Thus, the model smooths estimates across time during periods of stability, but also allows for rapid changes in the latent

trait during periods of volatility. It is possible to estimate the degrees of freedom, but this can lead to identification problems, which we explore in more detail in the appendix to Reuning et al. (2019). Setting a low degree of freedom of 4 has been recommended in other contexts (Gelman et al., 2014) and so we believe that it will be useful in most latent variable cases.

#### *Robust dynamic model priors*

$$\begin{aligned}\theta_{i1} &\sim N(0,1) \quad \forall i = 1, \dots, N \\ \theta_{it} &\sim T_3(\theta_{i(t-1)}, \sigma) \quad \forall i = 1, \dots, N \quad \forall t = 2, \dots, T \\ \sigma &\sim \text{HN}(0,3)\end{aligned}$$

## **EXTENSIONS OF LATENT VARIABLE MODELS AND SUGGESTIONS FOR FUTURE RESEARCH**

In this final section, we highlight different fruitful paths for research using latent variable models. We discuss new ways that theory has informed particular modeling strategies and how this can provide new insights. We then present Multi-Rater/Aldrich-McKelvey Scaling models, which allow researchers to use latent variable models to reduce the impact of rater preferences when trying to develop uniform scales from expert surveys. We go on to introduce Multiple-Indicator Multiple-Causes models. These models are relatively common in psychology but are rarely used in published political science research, even though they provide a principled way to test what drives change in a latent variable. We then discuss problems with different model fit statistics. We close with a set of best practices useful for guiding future research.

### ***The Seriousness with Which One Must Take Time***

The modeling structures outlined above identify only a few ways in which researchers

may care to model temporal dynamics. In practice, researchers are beginning to identify a variety of new strategies to address different forms of temporal non-independence. At times, for example, researchers have reason to suspect that the relationship between a manifest indicator and the latent trait may change over time. Kenwick (2018), for example, is interested in civilian control of regime institutions and argues that the strength of this control increases over time, with civilian control expected to be higher in a state where civilians have ruled for several decades than in one that had previously experienced a military take-over. He therefore structures the prior distribution on the latent trait for civilian regimes as a random walk with drift, allowing the values of the latent trait to systematically increase (or decrease) over time. Fariss (2014) faces a different type of temporal non-independence in the study of human rights violations, and argues that the standards with which human rights reports are written has changed over time. To accommodate these potential biases, Fariss (2018b) allows the item discrimination parameters linking standards based indicators to latent trait to vary over time to mitigate temporal biases.

In each case, the specific modeling structure used to generate estimates of the latent trait was informed by prior theory and the results are empirically validated against competing models. These examples demonstrate how the choice of modeling structure can fundamentally alter the estimates of the latent trait itself, and the theoretical inferences one draws from the measurement analysis. These insights are often non-trivial and must be treated with the same care with which other forms of hypothesis testing are conducted. Nevertheless, these examples demonstrate how the proliferation of dynamic variable modeling techniques offers fertile new testing grounds for the theoretical evaluation of concepts of interest.

### ***Models of Other Unit Dependences: Multi-Rater/Aldrich–McKelvey Scaling***

Latent variable approaches can also be useful in the context of expert and non-expert survey when there is concern over how individuals will respond to survey items. This question was first approached in research on surveys of voters in the United States (Aldrich and McKelvey, 1977; Hare et al., 2015), but has also recently been used in the context of expert surveys to quantify country level attributes (Marquardt and Pemstein, 2018). The benefits of these approaches, which we will refer to here as multi-rater IRT, is that in using them, researchers can place answers from survey participants that might view underlying concepts on different scales onto a single scale.

As an example, take the work of Marquardt and Pemstein (2018), in which the authors use a multi-rater model to place expert surveys about democratic practices within a country on a single scale. They start with a survey of experts, asking them to rate several countries on a variety of democratic attributes. The problem with using these ratings directly is that different experts might have different opinions about how democratic a country must be to be considered the most democratic, and may also vary in their general understanding of the question. This is a form of *differential item functioning* where the relationship between an item (a response to a particular survey question) and the latent trait varies.

To account for differential item functioning the  $\beta$  (discrimination) and  $\alpha$  (difficulty) parameters are estimated for each survey participant but held constant across the countries that they rated. For example, if  $Y_{ic}$  is expert  $i$ 's response to a question on country  $c$  then it would be estimated as a function of  $\alpha_i + \beta_i \theta_c$ .

This technique is fruitful not only in the context of expert surveys but also for non-expert surveys where there are varying perceptions. Hare et al. (2015) use this to

identify ideological placement of US senators from a survey of voters. The multi-rater method accounts for the fact that more liberal voters are likely to see the same senator as being more conservative than a moderate voter.

Nevertheless, in order for measures to be made comparable, there has to be a degree of overlap in the units that survey participants rate. This returns to the problem of bridging discussed above. Without overlap, the latent estimates will not be comparable across units. Overlap allows us to identify the degree of differential item functioning and so provide estimates of latent variables that are comparable when there is significant differential item functioning.

### ***Adding Even More Structure: MIMIC Models***

The final extension we consider is less focused on particular latent models and more on the use of estimates from the latent models. Latent models produce estimates of the latent traits that include error. The error needs to be a part of any future models that use the latent variable. When the latent variable estimates are used as an independent variable, estimation that incorporates error can be achieved relatively easily. All that is necessary is to take  $N$  draws from the posterior of the latent variable, estimate  $N$  models that use the latent variable as an IV and then combine those estimates using the same process that is used to combine multiple imputations.<sup>10</sup>

Estimating models where the latent variable is the dependent variable requires more care, but there are methods that are commonly used outside of political science that can accomplish this goal. Multiple-Indicator Multiple-Causes (MIMIC) models were developed starting in the 1970s to allow researchers to use multiple measures of a trait when estimating the impacts of exogenous variables on that trait (Jöreskog

and Goldberger, 1975; Muthén, 1989). The MIMIC model approach is commonly employed in psychology (Krishnakumar and Nagar, 2008) and was more recently introduced to political science in the context of political psychology (Pérez, 2011).

In brief, MIMIC models include covariates for the latent variable that is being estimated. These covariates are included in the initial estimation process and so capture the error that is inherent in measuring a latent variable. Covariates are included by modeling  $\theta$  directly as a function of the covariates instead of just setting a simple prior on it.<sup>11</sup> In addition to providing better estimates of the covariates on the underlying latent trait, MIMIC models can be modified to identify differential item functioning that is correlated with one of the covariates (Pérez, 2011).

One caveat for MIMIC models is that we are unaware of anyone who has connected the MIMIC approach to the dynamic latent variable approaches discussed here. Both approaches involve modifying the modeling of the latent variable (either through an informative prior or a regression setup) and so connecting the two will require additional work.

### **Assessing Model Fit: WAIC for Hierarchical and IRT Models**

WAIC (the Watanabe–Akaike or widely applicable information criterion) is currently one of the more preferred model diagnostics for Bayesian models (e.g., Gelman et al., 2014). However, several open research questions remain under-explored when using WAIC with hierarchical or IRT models.

WAIC is an approximation of leave-one-out validation, but approximating leave-one-out validation leads to a problem in IRT data over what ought to be ‘left out’ when validating models. That is, should individual items be left out for all unit-time periods, for units from a panel or for all unit-years? Or should all the items be left out for one of these unit

structures? Newly published research extends WAIC to cases in which items are clustered within an observation (Furr, 2017) as well as other work incorporating time dynamics (Li et al., 2016). Another recent area of work is diagnostics, and best practices for WAIC and other models (Vehtari et al., 2017).

When there is concern over the validity of WAIC statistics, it is useful to also estimate a K-fold cross validation. This of course also requires removing a set of data and estimating the model. We suggest that researchers randomly sample indicators to remove so that each unit-time is still in the model. This allows estimates of latent traits for each unit-time and those estimates can be used to calculate a held-out log-likelihood.

We suggest that while this area of research continues, researchers should provide multiple checks of model fit. Posterior predictive checks are another very powerful way to test how well an IRT model fits data (Gelman and Hill, 2007). Overall, fit statistics, posterior predictive checks and visual analysis of the temporal patterns of well-known cases allow for the evaluation of competing models without relying on a single statistical tool.

### **Best Practices for Applied Measurement Research**

Finally, as researchers use these methodologies, we propose a few useful suggestions on how to best approach modeling latent variables. It is our intention that these suggestions are consistent with the statistical modeling choices made when selecting the component parts of latent variable models, and that these choices will be made with reference to the two main types of construct validity also discussed. Recall that the process of measurement occurs in three iterative stages: *conceptualization* of the sociological or physical system being studied; *operationalization* of the data generating process that approximates the system; and *empirical analysis* of the data. The specific terms we use for each

of these three stages is *concept, construct, data*. Construct validity is an overarching term for assessing the relationship between one or more of the entities represented in each of these steps.<sup>12</sup>

- **Validate by letting the theoretical concept drive the measurement specification:** We have referred to this type of validation as translation validity and it is concerned with the link between the theoretical concept and the operationalized construct. It is not possible to consider a measure of an unobserved concept without referencing a theoretical concept. For a construct to be valid, it needs to translate the theoretical concept into an operational procedure that will generate data consistent with the theory. Thus, the first step for any research on latent variables is to outline the assumed relationships between the data generating process and the concept to be measured. Will the data generating process produce indicators that reflect the underlying concept of interest? Are the proposed items manifest of the underlying concept? Are the proposed items substitutes for each other? How are proposed items measured over time?
- **Validate by assessing the assumptions of the measurement model as they relate to theoretical concept of interest.** This is also a suggestion about translation validity. How does the specification of the measurement model translate the theoretical concept into the operational procedure that generates the observed data? Every measurement model has underlying assumptions and it is important that any empirical patterns are the result of the underlying data and not of the assumptions. In the case of latent measurement models, researchers must pay close attention to any parameters that are set without reference to theory of their concept of interest.
- **Validate the fit of the measurement model as it relates to the observed data.** How does the model of the data generating process, the latent variable, fit the observed data? This is an assessment of measurement validity. Measurement validity is an evaluation of the fit between the proposed data generating procedure and the actual data obtained from it. WAIC (the Watanabe–Akaike or widely applicable information criterion) and other statistical tools are useful ways to test model fit,

but researchers should not just select a model based on a single statistical tool. One useful way to test competing models is to focus on divergent estimates and use *a priori* knowledge about the world to validate which one is the best.

There is no guarantee that any single modeling strategy will be equally well-suited for use with all data types or for estimating all types of latent concepts. The assumptions of the measurement model will influence the conclusions researchers draw about the underlying theoretical concept of interest, as well as the empirical linkages between these concepts and other political phenomena.

## CONCLUSION

The assessment of theories about political institutions and behaviors often requires measuring concepts that are not directly observable. Thus, for science to proceed, measurement is essential, because without a clearly articulated link between the empirical content of a study and the theoretical structure that gives rise to that content, it is not possible to make claims about the relationship between data and the world. Yet, despite the necessity for valid measurement, research in the social sciences still often tends to ignore the construct validity of most measures and usually takes existing data, especially experimental data, for granted or at least as good enough. Thus, one of the critical steps in evaluating theoretical concepts is the development, formalization and validation of measurement models. This is because there is no model-free way to measure unobservable or difficult to observe concepts. And many of the concepts of interest to the political science community are often by definition difficult to observe. As we have discussed in this chapter, construct validity – and measurement models in general, and latent variable models in particular – are tools which are useful for systematically evaluating the

relationship between concepts, operational procedures (e.g., the data generating process) and data.

## Notes

- 1 For the purposes of this chapter, we focus exclusively on unidimensional measurement models that are explicitly created in an effort to link observed data to an unobservable concept.
- 2 The development of the concept of construct validity has occurred over many decades. Primary contributors include: Campbell and Fiske (1959); Campbell (1960); Campbell and Ross (1968); Cook and Campbell (1979); Shadish (2010); Shadish et al. (2001). However, the conceptual meaning of the terms used in these article have evolved over time. As Jackman (2008) notes, 'there are several species of measurement validity. But at least in the context of latent variables, the term "construct validity" has lost much of the specificity it once had, and today is an umbrella term of sorts' (122). We use the term construct validity in this way and point out specific subtypes where appropriate. We note further that different fields and subfields use the various construct validity terms in different ways, which has led to some confusion when translating across terms. Adcock and Collier (2001) review this issue in brief, but like them, we leave a full accounting for the agreement and disagreement of overlapping validity concepts to future work.
- 3 Reuning et al. (2018) provide a complete and detailed set of replication files that demonstrate how to use these particular latent variable models using both applied examples and a set of simulation-based models: <https://doi.org/10.7910/DVN/SSLCHF>.
- 4 We build on ideas covered in Adcock and Collier (2001) and elsewhere (e.g., Jackman, 2008; Shadish, 2010; Shadish et al., 2001).
- 5 The additive scale can be seen as a result of rewriting this to  $\beta_k (\theta - \alpha_k)$  and fixing  $\beta = \infty$ . This creates the step function that can be seen in Figure 20.2.
- 6 As the number of dimensions for the latent variable increases there is an increasing number of invariant rotations. For one dimension there are only two equivalent estimates; with two dimensions that number increases to eight (e.g., Jackman, 2001).
- 7 In the following section we will continue to leverage the normal prior for identification constraints, but we will introduce modifications to accommodate temporal dynamics.
- 8 The  $\sigma$  parameter is sometimes referred to as the innovation parameter.
- 9 In practice, one can also substitute a Student's t-distribution with a very high degree of freedom (e.g., 1,000), which closely approximates the normal distribution.
- 10 Mislevy (1991), Bolck et al. (2004) and Schnakenberg and Fariss (2014) each provide arguments and detailed suggestions on how to incorporate the uncertainty from latent variable estimate using the multiple imputation equation formula from Rubin (1987).
- 11 For more detailed discussion of estimations of MIMIC models see Fahrmeir and Raach (2007).
- 12 Two important parts of construct validity are translation validity and measurement validity. Translation validity is an evaluation of the match between the theoretical construct and the proposed data generating procedure which generates the observed pieces of information. Measurement validity is an evaluation of the fit between the proposed data generating procedure and the actual data obtained from it (Fariss and Dancy, 2017).

## REFERENCES

- Adcock, Robert and David Collier. 2001. 'Measurement Validity: A Shared Standard for Qualitative and Quantitative Research.' *American Political Science Review* 95(3):529–546.
- Aldrich, John H. and Richard D. McKelvey. 1977. 'A Method of Scaling with Applications to the 1968 and 1972 Presidential Elections.' *American Political Science Review* 71(1):111–130.
- Anders, Therese, Christopher J. Fariss, and Jonathan Markowitz. Forthcoming. 'Bread before guns or butter: Introducing Surplus Domestic Product (SDP)' *International Studies Quarterly*.
- Barbera, Pablo. 2015. 'Birds of the Same Feather Tweet Together: Bayesian Ideal Point Estimation Using Twitter Data.' *Political Analysis* 23(1):76–91.
- Blaydes, Lisa and Drew A. Linzer. 2008. 'The Political Economy of Women's Support for Fundamentalist Islam.' *World Politics* 60(July):576–609.
- Bolck, Annabel, Marcel Croon and Jacques Hagnaars. 2004. 'Estimating Latent



- Structure Models with Categorical Variables: One-Step versus Three-Step Estimators.' *Political Analysis* 12(1):3–27.
- Bond, Robert M. and Solomon Messing. 2015. 'Quantifying Social Media's Political Space: Estimating Ideology from Publicly Revealed Preferences on Facebook.' *American Political Science Review* 109(1):62–78.
- Campbell, Donald T. and Donald W. Fiske. 1959. 'Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix.' *Psychological Bulletin* 56(2):81–105.
- Campbell, Donald T. 1960. 'Recommendations for APA Test Standards Regarding Construct, Trait, or Discriminant Validity.' *American Psychologist* 15(August):546–553.
- Campbell, Donald T. and H. Laurence Ross. 1968. 'Analysis of Data on the Connecticut Speeding Crackdown as a Time-Series Quasi-Experiment.' *Law and Society Review* 3(1):33–54.
- Carpenter, Bob, Andrew Gelman, Matt Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus A. Brubaker, Jiqiang Guo, Peter Li and Allen Riddell. 2017. 'Stan: A Probabilistic Programming Language.' *Journal of Statistical Software* 76(1):1–32.
- Caughey, Devin and Christopher Warshaw. 2015. 'Dynamic Estimation of Latent Opinion Using a Hierarchical Group-Level IRT Model.' *Political Analysis* 23(2):197–211.
- Cook, Thomas D. and Donald T. Campbell. 1979. *Quasi-Experimentation: Design and Analysis for Field Settings*. Boston: Houghton Mifflin.
- Fahrmeir, Ludwig and Alexander Raach. 2007. 'A Bayesian Semiparametric Latent Variable Model for Mixed Responses.' *Psychometrika* 72(3):327. URL: <https://doi.org/10.1007/s11336-007-9010-7>
- Fariss, Christopher J. 2014. 'Respect for Human Rights Has Improved over Time: Modeling the Changing Standard of Accountability.' *American Political Science Review* 108(2):297–318.
- Fariss, Christopher J. 2018a. 'Are Things Really Getting Better? How to Validate Latent Variable Models of Human Rights.' *British Journal of Political Science* 48(1):275–282.
- Fariss, Christopher J. 2018b. 'The Changing Standard of Accountability and the Positive Relationship between Human Rights Treaty Ratification and Compliance.' *British Journal of Political Science* 48(1):239–272.
- Fariss, Christopher J. 2019. 'Yes, Human Rights Practices Are Improving over Time.' *American Political Science Review* 113(3):868–881.
- Fariss, Christopher J. and Geoff Dancy. 2017. 'Measuring the Impact of Human Rights: Conceptual and Methodological Debates.' *Annual Review of Law and Social Science* 13:273–294.
- Fariss, Christopher J., Michael R. Kenwick and Kevin Reuning. 2020. 'Estimating One-Sided Killings from a Robust Measurement Model of Human Rights.'
- Fonseca, Thaís C. O., Marco A. R. Ferreira and Helio S. Migon. 2008. 'Objective Bayesian Analysis for the Student-*t* Regression Model.' *Biometrika* 95(2):325–333.
- Furr, Daniel C. 2017. Bayesian and frequentist cross-validation methods for explanatory item response models. PhD thesis, University of California, Berkeley.
- Gandhi, Jennifer and Jane Lawrence Sumner. 2019. 'Measuring the Consolidation of Power in Non-Democracies'. *Journal of Politics*. Forthcoming.
- Gelman, Andrew, and Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press.
- Gelman, Andrew, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari and Donald B. Rubin. 2014. *Bayesian Data Analysis*. 3rd ed. Boca Raton, FL: CRC Press.
- Geweke, John. 1993. 'Bayesian Treatment of the Independent Student-*t* Linear Model.' *Journal of Applied Econometrics* 8(S1):S19–S40.
- Guttman, Louis. 1949. *The Basis for Scalogram Analysis*. Indianapolis: Bobbs-Merrill.
- Hare, Christopher, David A. Armstrong, Ryan Bakker, Royce Carroll and Keith T. Poole. 2015. 'Using Bayesian Aldrich-McKelvey Scaling to Study Citizens' Ideological Preferences and Perceptions.' *American Journal of Political Science* 59(3):759–774.
- Hollyer, James R., B. Peter Rosendorff and James Raymond Vreeland. 2014. 'Measuring Transparency.' *Political Analysis* 22(4):413–434.
- Imai, Kosuke, James Lo and Jonathan Olmsted. 2016. 'Fast Estimation of Ideal Points with Massive Data.' *American Political Science Review* 110(4):631–656.

- Jackman, Simon. 2000. 'Estimation and Inference via Bayesian Simulation: An Introduction to Markov Chain Monte Carlo.' *American Journal of Political Science* 44(2):375–404.
- Jackman, Simon. 2001. 'Multidimensional Analysis of Roll Call Data via Bayesian Simulation: Identification, Estimation, Inference, and Model Checking.' *Political Analysis* 9(3):227–241.
- Jackman, Simon. 2008. Measurement. In *The Oxford Handbook of Political Methodology*, ed. Janet M. Box-Steffensmeier, Henry E. Brady and David Collier, 119–151. Oxford University Press.
- Jessee, Stephen A. 2017. "'Don't Know" Responses, Personality and the Measurement of Political Knowledge.' *Political Science Research and Methods* 5(4):711–731.
- Jöreskog, Karl G. and Arthur S. Goldberger. 1975. 'Estimation of a Model with Multiple Indicators and Multiple Causes of a Single Latent Variable.' *Journal of the American Statistical Association* 70(351a):631–639.
- Kenwick, Michael R. 2018. 'Self-Reinforcing Civilian Control: A Measurement-Based Analysis of Civil-Military Relations.' *Working Paper*.
- König, Thomas, Moritz Marbach and Moritz Osnabrügge. 2013. 'Estimating Party Positions Across Countries and Time – A Dynamic Latent Variable Model for Manifestos Data.' *Political Analysis* 21(4):468–491.
- Krishnakumar, Jaya and A. Nagar. 2008. 'On Exact Statistical Properties of Multidimensional Indices based on Principal Components, Factor Analysis, MIMIC and Structural Equation Models.' *Social Indicators Research* 87:481–496. ID: UNIGE:41664. URL: <https://archive-ouverte.unige.ch/unige:41664>
- Lange, Kenneth and Janet S. Sinsheimer. 1993. 'Normal/Independent Distributions and Their Applications in Robust Regression.' *Journal of Computational and Graphical Statistics* 2(2):175–198.
- Lange, Kenneth L., Roderick J. A. Little and Jeremy M. G. Taylor. 1989. 'Robust Statistical Modeling Using the  $t$  Distribution.' *Journal of the American Statistical Association* 84(408):881–896.
- Li, Longhai, Shi Qiu, Bei Zhang and Cindy X. Feng. 2016. 'Approximating Cross-Validatory Predictive Evaluation in Bayesian Latent Variable Models with Integrated IS and WAIC.' *Statistics and Computing* 26(4):881–897.
- Linzer, Drew A. and Jeffrey K. Staton. 2016. 'A Global Measure of Judicial Independence, 1948–2012.' *Journal of Law and Courts* 3(2):223–256.
- Marquardt, Kyle L. and Daniel Pemstein. 2018. 'IRT Models for Expert-Coded Panel Data.' *Political Analysis* 26(4):431–456.
- Martin, Andrew D. and Kevin M. Quinn. 2002. 'Dynamic Ideal Point Estimation via Markov Chain Monte Carlo for the U.S. Supreme Court, 1953–1999.' *Political Analysis* 10(2):134–153.
- Martin, Andrew D., Kevin M. Quinn, and Lee Epstein. 2005. 'The Median Justice on the U.S. Supreme Court.' *North Carolina Law Review* 83(5):1275–1321.
- Mislevy, Robert J. 1991. 'Randomization-Based Inference about Latent Variables from Complex Samples.' *Psychometrika* 56(2):177–196.
- Mokken, R. J. 1971. *A Theory and Procedure of Scale Analysis*. Mouton.
- Muthén, Bengt O. 1989. 'Latent Variable Modeling in Heterogeneous Populations.' *Psychometrika* 54(4):557–585. URL: <https://doi.org/10.1007/BF02296397>
- Pan, Jennifer and Yiqing Xu. 2018. 'China's Ideological Spectrum.' *Journal of Politics* 80(1):254–273.
- Pemstein, Daniel, Stephen A. Meserve and James Melton. 2010. 'Democratic Compromise: A Latent Variable Analysis of Ten Measures of Regime Type.' *Political Analysis* 18(4):426–449.
- Pérez, Efrén O. 2011. 'The Origins and Implications of Language Effects in Multilingual Surveys: A MIMIC Approach with Application to Latino Political Attitudes.' *Political Analysis* 19(4):434–454.
- Plummer, Martyn. 2017. JAGS (Just Another Gibbs Sampler) 4.3.0. URL: <http://mcmc-jags.sourceforge.net/>
- Reuning, Kevin, Michael R. Kenwick and Christopher J. Fariss. 2018. 'Replication Data for: Exploring the Dynamics of Latent Variable Models.' URL: <https://doi.org/10.7910/DVNI/SSLCF>
- Reuning, Kevin, Michael R. Kenwick and Christopher J. Fariss. 2019. 'Exploring the Dynamics of Latent Variable Models.' *Political Analysis* 27(4):503–517.

- Rubin, Donald B. 1987. *Multiple Imputation for Nonresponse in Surveys*. Hoboken, NJ: J. Wiley & Sons.
- Schnakenberg, Keith E. and Christopher J. Fariss. 2014. 'Dynamic Patterns of Human Rights Practices.' *Political Science Research and Methods* 2(1):1–31.
- Shadish, William R. 2010. 'Campbell and Rubin: A Primer and Comparison of Their Approaches to Causal Inference in Field Settings.' *Psychological Methods* 15(1):3–17.
- Shadish, William R., Thomas D. Cook and Donald T. Campbell. 2001. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Belmont, CA: Wadsworth.
- Stegmueller, Daniel. 2011. 'Apples and Oranges? The Problem of Equivalence in Comparative Research.' *Political Analysis* 19(4):471–487.
- Stegmueller, Daniel. 2013. 'Modeling Dynamic Preferences: A Bayesian Robust Dynamic Latent Ordered Probit Model.' *Political Analysis* 21(3):314–333.
- Treier, Shawn and D. Sunshine Hillygus. 2009. 'The Nature of Political Ideology in the Contemporary Electorate.' *Public Opinion Quarterly* 73(4):679–703.
- Treier, Shawn and Simon Jackman. 2008. 'Democracy as a Latent Variable.' *American Journal of Political Science* 52(1):201–217.
- van Schuur, Wijnbrandt H. 2003. 'Mokken Scale Analysis: Between the Guttman Scale and Parametric Item Response Theory.' *Political Analysis* 11(2):139–163.
- Vehtari, Aki, Andrew Gelman and Jonah Gabry. 2017. 'Practical Bayesian Model Evaluation Using Leave-One-Out Cross-Validation and WAIC.' *Statistics and Computing* 27(5): 1413–1432.
- Voeten, Erik. 2000. 'Clashes in the Assembly.' *International Organization* 54(2):185–215.
- Windett, Jason H., Jeffrey R. Harden and Matthew E. K. Hall. 2015. 'Estimating Dynamic Ideal Points for State Supreme Courts.' *Political Analysis* 23(3):461–469.

# Measuring Attitudes – Multilevel Modeling with Post-Stratification (MrP)\*

Lucas Leemann and Fabio Wasserfallen

## INTRODUCTION

Reliable information on public opinion is, for the empirical analysis of many political science questions, a *conditio sine qua non* – not only for the whole sub-field of political behavior, but also for researchers interested in, for example, the congruence of public opinion with the legislator or government action (and how that is shaped by institutions). The basic methodological challenge for research on public opinion is to make valid inferences from a collected survey sample to the larger (underlying) population. Also, there may be several (sub-)populations of interests if we want to compare, for example, public opinion across the US states with national polling data that only includes a small number of respondents for certain states. We will present and discuss in this chapter multilevel regression and post-stratification, so-called MrP, which has made a seminal contribution to the estimation of

public opinion for subnational units, and also advances public opinion research in several other respects.

Often, we collect national polls, but may be interested in the estimation (and comparison) of public opinion on lower constituency levels (such as states or congressional districts in the US, Bundesländer or Wahlkreise in Germany and cantons and municipalities in Switzerland). The problem is that the sub-samples of respondents in the poll for some units – particularly the smaller ones – are too small. Small samples will increase the estimates' mean squared error – these estimates remain unbiased but their variance is large in small samples. In addition (and related to) this small-n problem, samples are often not representative for the larger population. The usual fixes for these caveats have been that researchers pool data from several surveys (to increase the sample) and that they use raking or post-stratification for calibration, which, in

effect, induces the structure of the underlying population on the sample (Miller and Stokes, 1963; Erikson et al., 1993; Zhang, 2000). The more recent literature has established MrP as a superior solution to these methodological problems, with several studies testing, validating and extending the method (Gelman and Little, 1997; Lax and Phillips, 2009b; Warshaw and Rodden, 2012; Leemann and Wasserfallen, 2017).

MrP is powerful because it combines the multilevel modeling of the survey data with post-stratification, using information on the population structures for which we aim to derive estimates. In short, researchers using MrP (a) organize the data in fine-grained ideal types, (b) make predictions for each ideal type using (national) survey data and (c) use post-stratification for the estimation of public opinion of the (subnational) constituencies of interest. This approach provides more precise estimates in case researchers face the small-*n* problem discussed above and/or work with skewed samples. Equally (or even more) important is that MrP provides a flexible and general framework that has opened avenues for constant improvements, fruitful combinations with other methods and sophisticated applications that are nicely tailored to specific data and research design challenges.

In the next section we present the basics of MrP and discuss technical issues, methodological limitations and extensions of the method. In essence, MrP is a stepwise framework that is conducive to flexible adaptations. Many different substantive political science puzzles have been empirically addressed with different versions of MrP. The contributions to various political science literatures in that respect are quite impressive, considering that MrP is still a young method. Among others, MrP has contributed to the literatures on political behavior, American politics, federalism, comparative politics and institutions. More specifically, several MrP studies generated important

insights into the study of partisanship, ideology, the responsiveness of institutions to public opinion, congruence of elite and voter preferences and polarization (e.g., Lax and Phillips, 2012; Tausanovitch and Warshaw, 2013; Leemann and Wasserfallen, 2016; McCarty et al., 2019). The impressive track record of MrP and its broad scope of substantive applications (also beyond political science, e.g., in sociology) may be considered to be the greatest contribution of the still rather young method. In that sense, MrP is much more than the successor of raking or post-stratification – with substantial potential for future research.

As far as methodological advances are concerned, Ghitza and Gelman (2013) elaborated a method for the analysis of deeply interacted subgroups; Leemann and Wasserfallen (2017) developed a MrP approach that relaxes the data requirement and increases the prediction precision; and Kousser et al. (2018) and Tausanovitch and Warshaw (2013) combine MrP with joint scaling and IRT. These are just a few examples of some recent methodological improvements that are, in all these cases, sophisticated answers to specific research design challenges. In addition to making academic contributions, MrP is also widely applied in the real world. This comes as no surprise, given that political actors are eager to get precise estimates of public opinion in election and referendum campaigns (Hanretty et al., 2016a). In addition, MrP's logic of ideal types lends itself straightforwardly to the analysis of public opinion in specific subgroups, which is of importance for political actors that want to develop campaigning efforts targeting specific subgroups of the population. A further attractive feature of MrP is that it allows leveraging of large datasets, also from samples with unequal participation (Wang et al., 2015; Downes et al., 2018). In the digital age, such data become the norm.

In this chapter we illustrate the basics of MrP with a running example, before we

discuss several technical issues, extensions and advances of the method. Moreover, we provide a broad overview of MrP, as it is applied in several (sub-)fields of the political science literature, by also discussing some (selected) substantive insights that we could gain thanks to several sophisticated studies. The broad spectrum of MrP applications is quite impressive and shows that MrP is, next to its methodological appeal, an approach that has great potential to further contribute important insights to various puzzles of political science. Finally, we elaborate on the challenges and limitations of MrP and speculate on future developments.

## MULTILEVEL REGRESSION WITH POST-STRATIFICATION (MRP)

Raking and post-stratification are two ways to generate weights for each individual observation. The idea of both methods is to rely on some stratifying variables (e.g., education, gender, age) to generate weights such that the weighted sample has the same marginal or joint distribution (with respect to these variables) as the target population.

Let's take a step back and look at the estimates of the outcome variable rather than how to generate weights for the observation. In both raking and post-stratification, we generate weights for each individual observation. Another way to think about this is that we organize the data in ideal types. If we were to post-stratify by education (six levels), gender (two levels) and age (four categories), we would think of the population as consisting of 48 different ideal types. Each individual in the sample can be assigned to a specific ideal type that is defined by the individual's education, gender and age. To generate weights, we can look at the magnitude of the share of respondents that belong to a specific ideal type and how large that share should be in the target population. The

weight is chosen such that the weighted share in the sample is equal to the actual share in the target population.

For raking or post-stratification, one just takes the mean of the outcome variable among all people belonging to the same ideal type. That estimate is then the best guess for the outcome variable for people of that ideal type. If the outcome variable is the support for a specific politician, the estimate will be between 0 and 1 and be interpreted as the share of people of that ideal type that support a specific politician.

What multilevel regression with post-stratification (MrP) does is different in the way we determine the estimate of the outcome variable for a specific ideal type. Rather than using this estimate for the ideal type's average support, MrP relies on a model to estimate the support among all survey respondents (Gelman and Little, 1997; Park et al., 2004). Based on this model, we can generate different predictions for different ideal types. This is the main technical difference from simpler methods, such as raking or post-stratification. As we show later, relying on a model rather than just a simple mean opens up many doors for analysis.

### *An Example: Voting on a Minaret Ban in Switzerland*

An example will help to illustrate MrP and its extensions discussed later. The example here is an (in)famous vote in Switzerland in 2009, when a slight majority of the Swiss population supported an initiative to ban the future construction of minarets. Surveys in advance of the vote indicated that it would be rejected but to the surprise of many it was accepted. The example here will rely on a dataset from a post-vote CATI carried out by universities in cooperation with a private company.

We start the illustration with a fairly simple MrP analysis which only includes

individual-level variables (education, gender, age). We thus think of the sample and also the population of consisting of 48 different ideal types.<sup>1</sup> At the core of MrP is a multi-level (or hierarchical) response model with the variable of interest as the outcome (in this example a yes or no vote for the minaret ban). It can be modeled as a probit or a logit – we opt for the probit. Education, gender and age are added to the model as random effects such that each education group has its own draw from a common normal distribution as well as the groups for gender and age. In addition, we add a random effect for the subnational units (here, cantons) to the model. If attitudes vary regionally (beyond variation due to different demographics), we account for this in the subnational (here, cantonal) random effects. There are 26 cantons in Switzerland and this leads to 1248 (26 × 48) different ideal types.

$$\begin{aligned}
 Pr(y_i = 1) &= \Phi\left(\beta_0 + \alpha_{k[i]}^{education} + \alpha_{j[i]}^{gender} + \alpha_{m[i]}^{age} + \alpha_{n[i]}^{canton}\right) \\
 \alpha_k^{education} &\sim N(0, \sigma_{education}^2), \text{ for } k = 1, \dots, 6 \\
 \alpha_j^{gender} &\sim N(0, \sigma_{gender}^2), \text{ for } j = 1, 2 \\
 \alpha_m^{age} &\sim N(0, \sigma_{age}^2), \text{ for } m = 1, \dots, 4 \\
 \alpha_n^{canton} &\sim N(0, \sigma_{canton}^2), \text{ for } n = 1, \dots, 26
 \end{aligned} \tag{1}$$

This model yields estimates for the parameter  $\hat{\beta}_0$  as well as for the realizations of the random effects (e.g., for the first education group it is  $\hat{\alpha}_{k=1}^{education}$ ). This allows us to create a prediction for any of the 48 ideal types by adding the respective random effect realization of that ideal type. For any ideal type, we can add the constant to the four chosen random effect realizations and have an ideal type's score on the latent variable. After transforming that – via the cumulative standard normal distribution – into a probability, we derive an estimate of the share of a specific ideal type that is expected to support the minaret ban. This is markedly

different from what one does when relying on raking or post-stratification, where we would have just taken the average response (among all respondents of that ideal type). In this example, there are actually less observations than ideal types, but since the predictions are model-based this does not pose a problem.

The second step of MrP is post-stratification. We generate an average support for each ideal type and then weigh this by the relative share of a type in the target population ( $N_{ng}$  denotes the number of people of type  $g$  living in subnational unit  $n$ ). This is only possible with precise information on the structure of the target population. Often used variables such as age, education, gender and (in the US) race are known due to the census. Here, we need to know exactly how many people of a specific age group, education level and gender live in a specific subnational unit. This allows us to determine the average support in a subnational unit by weighing the support per ideal type (denoted as  $g$ ) by the share of that type in the unit:

$$\hat{\pi}_n = \frac{\sum_{g \in n} \hat{\pi}_{ng} N_{ng}}{N_n} \tag{2}$$

The running example here is the vote on the minaret ban in Switzerland in 2009. The raw data of the exit poll suggests that a majority of 51.5% of the electorate was against the ban. In other surveys leading up to the vote, an even larger majority seemed to oppose the ban. It was a great surprise to many observers when, on the day of the vote, the official results were announced: 57.5% of voters supported the ban. This triggered a larger debate about the values of public opinion polling in Switzerland. Would MrP have helped here? Yes, the simple MrP model described above provides an estimate of 58.0%, which is very close to the true value.

### Using MrP to Generate Subnational Preference Estimates

As mentioned above, MrP has received a lot of scholarly attention mostly due to its ability to generate credible estimates of subnational preferences (Lax and Phillips, 2009a,b). So far, we have only used MrP to correct for the unequal make-up of the sample’s structure (non-response bias). In a next step, we will use additional context-level variables to generate subnational estimates. We can add context-level explanatory variables by redefining the distribution of the cantonal random effect to be  $\alpha_n^{canton} \sim N(\beta X_n, \sigma_{canton}^2)$  whereas  $X$  is a matrix with a leading column of 1’s. Context-level variables were not included in the original MrP paper (Gelman and Little, 1997), but from Park et al. (2004) onwards many published MrP models include context-level variables.

Warshaw and Rodden (2012) show – looking at US data – that using context-level variables improves the district-level estimates. Since the estimated support for an ideal type is based on a model prediction rather than a simple sample average, one can easily include additional information into the model. With

US data, the most common variable used is the presidential vote share in the preceding election, but other frequently used variables are median income, share of veterans or share of evangelical Protestants and Mormons.

In the example here, we can also add a context-level variable. We are looking for a variable that only varies across cantons and that is likely correlated with the collective part of the voting decision. Variables picking up on political culture or more structural variables such as unemployment or income levels could be used as well. But since Swiss voters vote frequently on issues, we can actually use a past vote on a related issue. In 2008, an initiative wanted to change the constitution – to de facto reverse a ruling by the highest court – to make it easier for municipalities to vote on naturalization cases. The issue at hand and the likely motivating factors as well as the partisan vote recommendations were similar to those in the minaret ban vote. We include the share of people voting in favor of that initiative (which was eventually rejected) to the response model.

Figure 21.1 shows the estimates based on the raw data and two different MrP models. The estimates based on the raw data take the

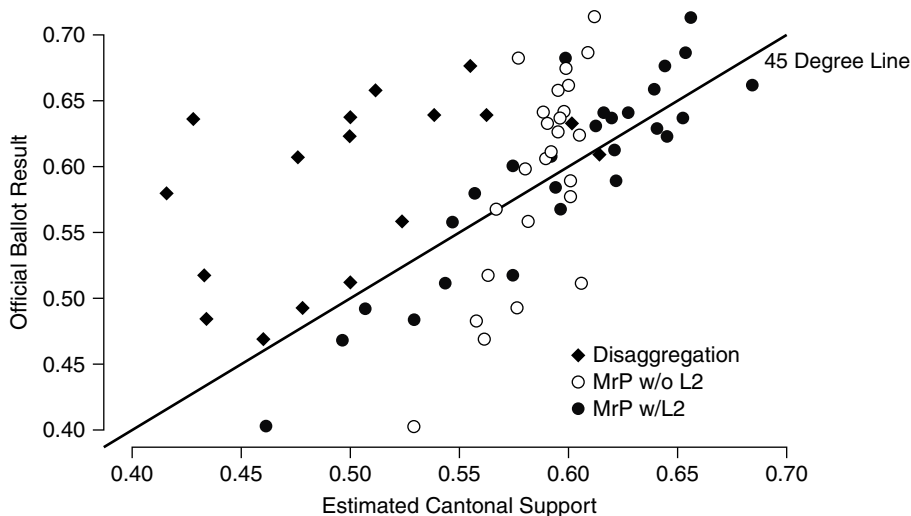
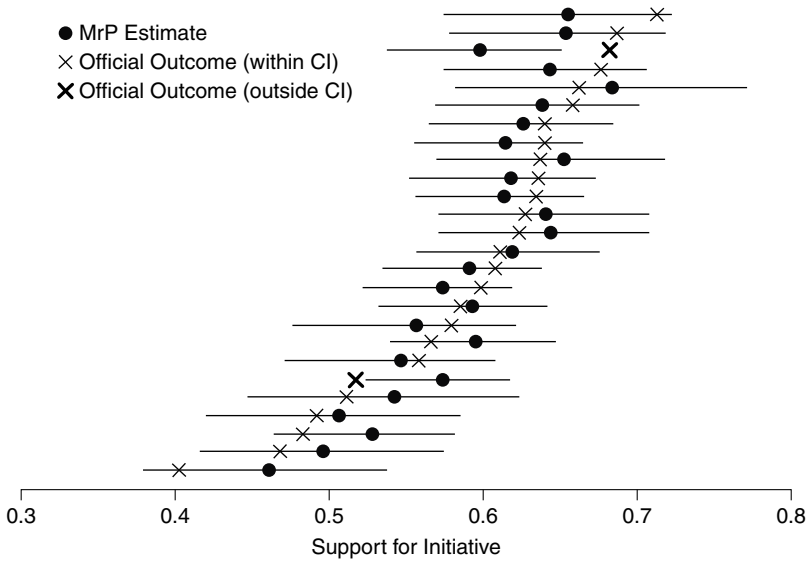


Figure 21.1 Cantonal estimates and true approval rate





**Figure 21.2 Cantonal estimates and uncertainty**

average response per canton. The MrP without any context-level variable is what was defined in Equation 1, and the third model includes the vote share in the earlier vote. The disaggregation estimates do not perform well due to the low sample size per canton. This is in line with prior research (Lax and Phillips, 2009b; Warshaw and Rodden, 2012; Leemann and Wasserfallen, 2017). In addition, we see that the MrP estimates based on a model with no context-level variable yields estimates biased toward the sample mean (here 48.5%). We discuss this below in more detail. Finally, adding a context-level predictor increases prediction accuracy and the estimates based on MrP with a context-level variable are neatly clustered around the 45 degree line.

### ***Some Technical Remarks***

Why is it that MrP performs so much better than raw data in the estimation of sub-national public opinion? Part of the answer is that the model is estimated on all observations in the sample, while the disaggregation

results only rely on responses from a specific canton. The key is in the multilevel model used here that allows for partial pooling. When estimating the realizations of the random effects, for example, for the four age groups, all observations are used. When then generating predictions (e.g., for a young well-educated woman in the canton of Zurich), the model also benefits from men in other cantons in different education groups.

Partial pooling comes into the model by incorporating random effects. Random effects can be thought of as a weighted average between a global estimate and a local estimate, whereas the specific weights are based on the entire variation and the local variation (see Gelman and Hill, 2007: 253–4 for a more thorough discussion). Hence, the multilevel model allows for a more efficient use of the data.

A consequence of partial pooling is that the cantonal random effects can be biased toward 0 and this leads in the predictions to a bias toward the sample mean. The sample mean within a unit is less influential on that unit's random effect realization if there are few observations within the unit (Gelman and

Hill, 2007: 253). As a consequence, the partial pooling will be stronger for smaller units with fewer respondents (Buttice and Highton, 2013: 5). This can be nicely seen in Figure 21.2, where the estimates of MrP without a context-level variable are too steep and show little variation across cantons.

Once we add context-level variables the subnational estimates vary much more, since they enter the model as a regular variable and we estimate a fixed coefficient.<sup>2</sup> Almost all MrP applications now rely on a multilevel model with context-level variables and it is good practice to include them.

Finally, it is also worth considering that MrP is partly based on post-stratification. And post-stratification will perform best when there is little variation within a cell (here: specific combination of gender, age, education and canton) but large variation across cells. That is the case for any procedure that relies on strata, cells or ideal types. MrP will perform better the more homogeneous preferences are within an ideal type and the more variation there is among different ideal types.

One issue we have not yet touched upon is uncertainty. We presented the estimates as point estimates and did not discuss any strategy to generate the appropriate confidence intervals. Since the estimates are based on model predictions, we can easily generate the uncertainty estimates by relying on simulation (Herron, 1999; Gelman and Hill, 2007). This approach works whether one relies on `stan` or `lme4` to estimate the response model. We can illustrate how certain or uncertain we are by displaying the 95% confidence interval for the minaret vote.

Most confidence intervals have a width of about 12 percentage points and 24 of the 26 cantons had an actual result that was in line with the MrP estimate, that is, the official outcome was within the confidence interval.

This section described the standard MrP model and used an example to illustrate that MrP can help with two different – but frequent – challenges encountered when

working with survey data. On the one hand, MrP can deal well with data sets that suffer from non-response bias as it is found in many standard surveys used by academics. On the other, it enables researchers to generate subnational estimates that are by far superior to disaggregation.

Finally, a last point that has not yet received a lot of attention is how to select context-level variables. In the running example here it is easy to come up with a related but older ballot vote and to use those results as the context-level variable. But what do we do if we want to estimate support or opposition to a new EU treaty in a country that does not hold regular ballot votes on issues? It is worth stressing that the selection of context-level variables is important for two reasons. First, those are the only variables that enter the model as ‘fixed effect’ (in the hierarchical meaning and not meant to describe the within-estimator). Since MrP engages in prediction we should worry about overfitting (James et al., 2013). Second, it has been shown that using optimal context-level variables can greatly increase the prediction precision of MrP (Warshaw and Rodden, 2012). While authors discuss the variables included in the model, there is often no systematic justification provided for the specific choice.<sup>3</sup> We return to this issue below.

### ***Methodological Limitations and Extensions of MrP***

MrP has proven to be a valuable method in the political science toolbox, but it also comes with limitations and downsides. For one, MrP may work well to estimate district-level preferences when we have at least some responses in every district. However, MrP will not work to one’s satisfaction at increasing degrees of granularity – at some point, the identified small areas are too small and one will have hardly any respondents in the sample.

But beside this general limitation there are also more specific points. Buttice and

Highton (2013) published a paper that shows that the performance of MrP can vary and that there is no guarantee that MrP estimates are precise. They argue that MrP performs well for cultural political attitudes in the US but that this does not extend to all areas. The strength of this paper is that they work through various aspects that are important.

One relevant aspect to assess whether MrP can do well or not is the intra-class correlation (ICC). The ICC accounts for how strongly opinions vary across states and how strongly they vary within states. Buttice and Highton (2013) argue that MrP performs most strongly when almost all variance is on the context level (for them, the state level). In addition, they underline the value of having context-level variables in the multilevel model.

Another point made by Buttice and Highton (2013) is that researchers cannot rely on a canonical response model when analyzing different issues. More practically expressed, the hierarchical response model should be differently specified when researchers analyze public opinion on cultural issues (e.g., same sex marriage) or economic questions (e.g., tax levels). This echoes the point made by Warshaw and Rodden (2012), who advise including variables that are strong predictors of variation across units in the response model. The question of how to choose optimal context-level variables remains difficult. As we discuss later, machine learning applications of MrP promise improvements in that respect.

Some of the discussed shortcomings have already been addressed in the literature. First, the problem of *too many* subnational units was encountered early on by researchers working on German elections, where there are 299 districts. This is the problem that Selb and Munzert (2011) tackled: they wanted to estimate constituency-level estimates for 299 German Wahlkreise. Technically, nothing is wrong and one can estimate an MrP model. Realizations of the context-level random effect are 0 for those units that have no

respondent in the sample. But this approach might be inefficient as it does not exploit all information we have. While we may not have any respondents from a specific unit, we might have respondents from neighboring units, and could exploit that.

Another problem in such a situation is that we may not have socio-economic data broken down by unit, as electoral districts need not perfectly overlap with administrative boundaries, which are usually the standard units of census data. Hence, it is not immediately clear how one would carry out the post-stratification step. The major contribution in Selb and Munzert (2011) is to formulate a model where the unit-level random effect is spatially autocorrelated – this then allows us to exploit that we know which respondents live closer or further away from a specific district. They show how information can be added in such a model and also how this can be used within an MrP model (given that one does have administrative data on the electoral districts to be able to post-stratify). They replicate the empirical example of Park et al. (2004), which only has 50 units, and show that by adding a spatially autocorrelated random effect the estimates do not improve by very much. With a well-specified MrP model there is not that much unexplained variation that is spatially correlated and can be absorbed. But that just shows that it is not always necessary to rely on their extension – nevertheless, when facing too many districts, Selb and Munzert (2011) have shown a reliable and feasible procedure to produce estimates.

A second extension relates to the ability of the researcher to build a strong response model. While Buttice and Highton (2013) and Warshaw and Rodden (2012) advocate improving model quality by selecting better context-level variables, there is also another complementary strategy. It is also possible to add additional individual-level variables that will greatly improve the estimates. Variables such as party affiliation and income might be powerful predictors but are not part of the

census. If a variable is not part of the census, we cannot get the joint distribution, and hence it is not possible to post-stratify by such additional variables. This is unfortunate since one usually relies on less strong models due to the data constraints following in the post-stratification step. This was the motivating problem that Leemann and Wasserfallen (2017) address.

The first part of their argument is that the need for the full joint distribution stems from the binary part of the response model. When generating predicted probabilities, the impact of the realization of, e.g., the gender random effect will vary depending on where an ideal type's score on the latent variable is. Simply put, if the response model was a linearly additive model (as with, e.g., OLS) there would be no problem. But since we rely on binary models, we are confined to using a joint distribution. The second part is that they show how one can generate a synthetic joint distribution whereby one assumes independence between gender/education/age and, e.g., party ID (simple synthetic) or learns about the correlational structure from the survey data itself (adjusted synthetic). Following their approach allows reliance on powerful predictors on the individual level as long as one can find marginal distributions for those variables at the subnational level. This leads to multilevel regression with synthetic post-stratification (MrsP). For the authors, MrsP is MrP's better half. They show in simulations and replications that MrsP can outperform MrP and reduce the estimation error by as much as MrP does over disaggregation.

## THE USE OF MrP IN THE LITERATURE

The chapter has so far presented the underlying methodology of MrP by illustrating the method with a running example and has discussed several methodological limitations and extensions of MrP. What has become clear from this discussion is that MrP

provides a more precise and sophisticated method for subnational unit estimation as compared to disaggregation or raking. In addition, the discussion has shown that MrP builds on a distinct methodological structure, which provides the ground for ongoing and future innovations, adaptations and combinations of MrP with other methods. Accordingly, MrP is not simply the more precise successor of disaggregation or raking. Rather, the advances of MrP allow for the study of a new set of substantively interesting political science research questions.

It is noteworthy that the development of the method is not (exclusively) out-sourced to specialized method journals, but has also been advanced by substantive political science research. This combination of methodological development and substantive advances of various sub-field literatures is one of the most exciting features of the recent MrP literature. In the following, we discuss the state-of-the-art literature by focusing on three themes: testing and validation in different contexts, combinations of MrP with other methods and the broad substantive applications of MrP.

Multiple articles using MrP provide a rigorous methodological testing of the method. The previous section already discussed important methodological limitations and extensions of MrP in that respect. Given that MrP is still a relatively young method, the analyses of both the estimation precision and the scope conditions under which it performs well are critical for its establishment as 'gold standard' (Selb and Munzert, 2011: 456). An important factor for the quick establishment of MrP is that several substantive articles using MrP test and validate the precision of MrP in various different settings and contexts before they apply the method. Situated in the scholarship on American politics, Enns and Koch (2013), for example, estimate state public opinion on the level of US states and validate their application of MrP with specialized polls that were conducted for a subset of six states.

Going one level deeper, Warshaw and Rodden (2012) made a seminal contribution in popularizing MrP by showing how the method can be applied for the estimation of public opinion on congressional and state legislative districts in the United States. Among others, they validate the method by comparing MrP estimates with results of same sex marriage referendums in Arizona, California, Michigan, Ohio and Wisconsin. Also, other cross-validation methods have been applied, such as randomly splitting data from the whole sample (using only part of the sample for predictions). However, the validation with voting data is particularly stringent, as it uses the real world as a benchmark (as the example illustrated in the previous section does). Leemann and Wasserfallen (2016) make a similar test of their Swiss MrP models by comparing cantonal MrP estimates with results of 186 direct democratic votes, whereas Tausanovitch and Warshaw (2013) use, as real-world comparison, 2008 presidential vote shares.

Besides showing that MrP outperforms other approaches, several of these studies also advance MrP in some respect, often by combining it with other methods. For example, Enns and Koch (2013) estimate state-level policy preferences from 1956 to 2010, extending the use of MrP to the analysis of long time series. Also, Pacheco (2011, 2012) estimates time-series data for her analysis of change in US state public opinion, using imputation techniques to improve the estimates. A classic question of MrP studies is the study of congruence between public opinion and the views of their political representatives, and between public opinion and policy outcomes. The MrP literature has been careful to make sure that the opinions of voters and representatives are measured on the same scale. To that end, Kousser et al. (2018) combine MrP with joint scaling; Tausanovitch and Warshaw (2013) scale survey items with IRT; and Leemann and Wasserfallen (2016) field their own survey, accounting for the joint scaling in the design of the questionnaire.

The availability of precise time-series public opinion data and consistent estimates of elite and voter opinions on the same issues produced important empirical insights for the literatures on public opinion, representation and democratic performance.

The previous section discussed the pertinent challenges of the specification of the hierarchical response model as well as the methodological contributions by Selb and Munzert (2011) and Leemann and Wasserfallen (2017). More generally speaking, the literature on the limitations and extensions of MrP highlights that researchers ought to be familiar with the basics of MrP. They should develop an application of the method that is carefully tailored to their research design by taking into account all available data sources for their units of analysis. In that respect, also the work by Ghitzza and Gelman (2013) makes an important contribution by reducing model-dependency with the analysis of deeply interacted subgroups, which are subsets of the population that are defined by multiple demographic and geographic characteristics.

This line of research is of interest for political scientists that are interested in very specific subgroups of the population. Also, for political actors that are seeking information on turnout and vote choice – preferably on a very specific level in terms of voter characteristics that allow for targeted campaigning efforts. Accordingly, it does not come as a surprise that MrP has become a standard tool in the field of campaigning and forecasting. A further powerful feature is that MrP can leverage large data sources that nowadays can be collected through online surveys. We name just a few examples here (from a long list): Hanretty et al. (2016a) present a forecast model for the 2015 British general election, and, among many others, CBS (in the US), YouGov (in the UK) and LeeWas (in Switzerland) conduct regular surveys using MrP. A further, rather general, attractive feature of MrP is that it can be applied to the analysis of large datasets collected online, including from samples

with unequal participation (Wang et al., 2015; Downes et al., 2018).

In the digital age, large datasets that are not random samples of the target population have become the norm. MrP has been successfully used in the analysis of such non-random samples, even in a case in which Xbox players were used to predict the 2012 presidential election in the US (Wang et al., 2015).<sup>4</sup> While MrP has received some attention for its seeming ability to work well with non-probability polls, there is no secret magic entailed in the use of a regression model and a post-stratification procedure. When non-probability samples can yield valuable insights and when they cannot, will likely be a question that will attract much more research, particularly from polling companies that have an economic interests in further developing such approaches. For now, we refer to Ansolabehere and Rivers (2013) for a useful overview and discussion of the necessary assumptions.

Coming back to the use of MrP in the political science literature, it has been applied to a broad spectrum of substantive research. Among others, MrP has advanced literatures on partisanship, ideology, the responsiveness of institutions to public opinion, congruence of elite and voter preferences and polarization. For example, Kestel et al. (2010) analyze how public opinion matters for senate vote confirming supreme court nominees, McCarty et al. (2019) analyze how polarization differs between voters and legislators, while a series of studies investigates the congruence between public opinion, representatives and policy outcomes – typically as a function of different institutional and electoral arrangements (Kousser et al., 2018; Lax and Phillips, 2012; Tausanovitch and Warsaw, 2013; Leemann and Wasserfallen, 2016). For a still rather young method, MrP has an impressive track record in providing substantive empirical insights to various (sub-)disciplines of political science. This may be its greatest success (and promise for the future).

Beyond political science, MrP has also been used, for example, in sociology. Claassen and Traunmüller (2017) use MrP as a tool for the analysis of hard to measure populations (not as a method for estimating public opinion). More specifically, they estimate the demographic structures of Hindus, Muslims and Jews in Great Britain using survey data. Since there is good census data for these religious minorities in Great Britain, they can validate the accuracy of their estimates. Thus, methodological advances and extensions of MrP (that are motivated by substantive research) are not restricted to the political science literature.

## CONCLUSION

MrP has, in a rather short time, established itself as the standard method for the estimation of public opinion for subnational units (and other subgroups of populations and samples). MrP provides a distinct and flexible framework in three steps, with the specification of the hierarchical response model, the prediction for ideal types and the post-stratification step. This general methodological structure allows for tailored applications, combinations with other methods and innovative extensions. In this chapter, we presented all steps of MrP with a running example, before discussing in greater detail technical issues, methodological limitations and extensions.

We have also discussed multiple advances of MrP and expect much more to come. As discussed, the selection of optimal context-level variables for the response model is critical, yet not straightforward. With the increasing use of machine learning approaches in political science (see e.g., Montgomery and Olivella, 2018), there are promising ideas and projects that are likely to provide further useful solutions for the challenge of selecting the response model in MrP applications. Machine learning is

an umbrella term for a lot of methodologies that improve model-specification and prediction accuracy with a disciplined approach. Several working papers combine classifiers with post-stratification and aim to bring the promise of statistical learning to MrP (Goplerud et al., 2018; Ornstein, 2017; Broniecki et al., 2018). The combination of MrP with other methods has already shown it to be very productive, and semi-automated procedures will likely help further improve MrP for certain applications.

However, most important for the continuing success story of MrP will be that methodological innovations are not an end in itself. Rather, future advances of MrP have to provide solutions for technical problems that are motivated by substantive research puzzles and are powerful enough to solve them. To the extent that this will continue to shape the development and applications of MrP, the importance of MrP for the political science literature will continue to grow. As the discussion of the broad substantive applications of MrP in the literature has shown, its track record in that respect is already quite impressive.

## Notes

- \* A complete replication file with an illustrative example can be found here: [https://github.com/lleemann/MrP\\_chapter](https://github.com/lleemann/MrP_chapter).
- 1 Gender has two categories, education has six categories (mandatory schooling or no response, apprenticeship, university-entrance diploma (Matura) and teachers college, additional job training, advanced training, university degree including universities of applied sciences), and age has four categories (18–34, 35–49, 50–64, 75–).
- 2 Fixed refers here to a coefficient that is estimated as a specific value, unlike random effects, where we estimate the variance part. This is not what is often referred to as within-estimator. See Gelman and Hill (2007: 245–6) for a more detailed discussion.
- 3 There is one exception: Leemann and Wasserfallen (2016) touch on this issue in the appen-

dix and show that they try to pick context-level variables based on AIC and BIC measures of the estimated models.

- 4 See also the work of Hanretty et al. (2016a,b) on UK YouGov samples for various elections and votes.

## REFERENCES

- Ansolabehere, Stephen, and Douglas Rivers. 2013. 'Cooperative survey research.' *Annual Review of Political Science* 16: 307–329.
- Broniecki, Philipp, Lucas Leemann, and Reto Wüest. 2018. 'Improved Multilevel Regression with Post-Stratification Through Machine Learning (autoMrP).' *Working Paper*.
- Buttice, Matthew K., and Benjamin Highton. 2013. 'How does multilevel regression and poststratification perform with conventional national surveys?' *Political Analysis* 21(4): 449–467.
- Claassen, Christopher, and Richard Traunmüller. 2017. 'Improving and validating survey estimates of religious demography using Bayesian multilevel models and poststratification.' *Sociological Methods & Research*. doi:10.1177/0049124118769086.
- Downes, Marnie, Lyle C. Gurrin, Dallas R. English, Jane Pirkis, Dianne Currier, Matthew J. Spittal and John B. Carlin. 2018. 'Multilevel regression and poststratification: a modelling approach to estimating population quantities from highly selected survey samples.' *American Journal of Epidemiology*. doi:10.1093/aje/kwy070.
- Enns, Peter K., and Julianna Koch. 2013. 'Public opinion in the US states: 1956 to 2010.' *State Politics & Policy Quarterly* 13(3): 349–372.
- Erikson, Robert S., Gerald C. Wright and John P. McIver. 1993. *Statehouse Democracy: Public Opinion and Policy in the American States*. Cambridge: Cambridge University Press.
- Gelman, Andrew, and Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press.

- Gelman, Andrew, and Thomas C. Little. 1997. 'Poststratification into many categories using hierarchical logistic regression.' *Survey Research* 23: 127–135.
- Ghitza, Yair, and Andrew Gelman. 2013. 'Deep interactions with MRP: election turnout and voting patterns among small electoral subgroups.' *American Journal of Political Science* 57(3): 762–776.
- Goplerud, Max, Shiro Kuriwaki, Marc Ratkovic and Dustin Tingley. 2018. 'Sparse Multilevel Regression (and Poststratification (sMRP)).' *Working Paper*, Harvard University.
- Hanretty, Chris, Ben Lauderdale and Nick Vivyan. 2016a. 'Combining national and constituency polling for forecasting.' *Electoral Studies* 41: 239–243.
- Hanretty, Chris, Benjamin E. Lauderdale, and Nick Vivyan. 2016b. 'Comparing strategies for estimating constituency opinion from national survey samples.' *Political Science Research and Methods* 6(3): 571–591
- Herron, Michael C. 1999. 'Postestimation uncertainty in limited dependent variable models.' *Political Analysis* 8(1): 83–98.
- James, Gareth, Daniela Witten, Trevor Hastie and Robert Tibshirani. 2013. *An Introduction to Statistical Learning*. Vol. 112. Springer.
- Kastellec, Jonathan P., Jeffrey R. Lax and Justin H. Phillips. 2010. 'Public opinion and Senate confirmation of Supreme Court nominees.' *Journal of Politics* 72(3): 767–784.
- Kousser, Thad, Justin Phillips and Boris Shor. 2018. 'Reform and representation: a new method applied to recent electoral changes.' *Political Science Research and Methods* 6(4): 809–827
- Lax, Jeffrey R., and Justin H. Phillips. 2009a. 'Gay rights in the States: public opinion and policy responsiveness.' *American Political Science Review* 103(3): 367–386.
- Lax, Jeffrey R., and Justin H. Phillips. 2009b. 'How should we estimate public opinion in the States?' *American Journal of Political Science* 53(1): 107–121.
- Lax, Jeffrey R., and Justin H. Phillips. 2012. 'The democratic deficit in states.' *American Journal of Political Science* 56(1): 148–166.
- Leemann, Lucas, and Fabio Wasserfallen. 2016. 'The democratic effect of direct democracy.' *American Political Science Review* 110(4): 750–762.
- Leemann, Lucas, and Fabio Wasserfallen. 2017. 'Extending the use and prediction precision of subnational public opinion estimation.' *American Journal of Political Science* 61(4): 1003–1022.
- McCarty, Nolan, Jonathan Rodden, Boris Shor, Chris Tausanovitch and Christopher Warshaw. 2019. 'Geography, uncertainty, and polarization.' *Political Science Research and Methods* 7(4): 775–794.
- Miller, Warren E., and Donald E. Stokes. 1963. 'Constituency influence in Congress.' *American Political Science Review* 57(1): 45–46.
- Montgomery, Jacob M., and Santiago Olivella. 2018. 'Tree-based models for political science data.' *American Journal of Political Science* 62(3): 729–744.
- Ornstein, Joseph T. 2017. 'Machine Learning and Poststratification.'
- Pacheco, Julianna. 2011. 'Using national surveys to measure dynamic U.S. state public opinion: a guideline for scholars and an application.' *State Politics & Policy Quarterly* 11(4): 415–439.
- Pacheco, Julianna. 2012. 'The social contagion model: exploring the role of public opinion on the diffusion of antismoking legislation across the American states.' *Journal of Politics* 74(1): 187–202.
- Park, David K., Andrew Gelman and Joseph Bafumi. 2004. 'Bayesian multilevel estimation with poststratification: state-level estimates from national polls.' *Political Analysis* 12: 375–385.
- Selb, Peter, and Simon Munzert. 2011. 'Estimating constituency preferences from sparse survey data using auxiliary geographic information.' *Political Analysis* 19(4): 455–470.
- Tausanovitch, Chris, and Christopher Warshaw. 2013. 'Measuring constituent policy preferences in Congress, state legislatures, and cities.' *The Journal of Politics* 75(2): 330–342.
- Wang, Wei, David Rothschild, Sharad Goel and Andrew Gelman. 2015. 'Forecasting



- elections with non-representative polls.' *International Journal of Forecasting* 31(3): 980–991.
- Warshaw, Christopher, and Jontahan Rodden. 2012. 'How should we measure district-level public opinion on individual issues?' *Journal of Politics* 74(1): 203–219.
- Zhang, Li-Chun. 2000. 'Post-stratification and calibration – a synthesis.' *The American Statistician* 54(3): 178–184.

PART IV

# Large-Scale Data Collection and Representation Methods



*This page intentionally left blank*

# Web Data Collection: Potentials and Challenges

Dominic Nyhuis

## INTRODUCTION

The universal adoption of the world wide web has brought about tremendous opportunities for the social sciences. Scores of data on every conceivable subject are often just one click away, making this the golden age of data. Indeed, one might reasonably argue that the main challenge in empirical social research has shifted from the collection of data to drawing value from the vast datasets. These changes are exemplified by the establishment of entirely new research fields, often summarized with the shorthand of computational social science (Alvarez, 2016; Behnke et al., 2018), and new professions such as the fabled data scientist – part statistician, part computer scientist, part social scientist.

One important aspect of these changes is the development of new methods for analyzing large datasets (see, e.g., Chapters 28–30 on text analysis and 55–56 on machine learning in this *Handbook*). In this chapter, we will

take a step back and consider an element of the research process in the era of big data that is less frequently talked about, but equally important – the question of how to assemble web data in the first place. The value of these skills hardly needs elaboration. If researchers know how to automatically collect data from the web, they are able to amass enormous datasets with little human input and little to no cost, putting even junior researchers with no research budget in a position to make valuable research contributions. What is more, while this skill set is highly sought after both inside and outside of academia, it is still surprisingly rare, such that mastering the art of web data collection allows scholars to push the boundaries of their particular field.

To be sure, even though web data collection is much less costly than, say, survey research, it can be a fairly cumbersome exercise nonetheless. Consider the workload in an ordinary survey research project. Survey research requires great care in preparing the

questionnaire and effort in surveying the respondents. Once the field work is done, however, the researcher can typically expect to hold a fairly clean spreadsheet of responses to proceed with the analysis. By contrast, in many web data collection projects, most time and effort is needed when the data collection in a strict sense is already completed. Data cleaning (e.g., checking for duplicate entries and correcting errors of various kinds) has become an integral part of the data collection exercise. This and other challenges related to web data collection are important considerations when deciding upon the best data collection strategy.

To elaborate the potentials and challenges of web data collection, we begin the remainder of this chapter with some technical backgrounds. Specifically, in the next section, we elaborate the basic ideas about gathering data from application programming interfaces (APIs) and how to scrape data from websites. We should caution readers that while the techniques are simple enough, there are quite a few fundamentals to cover before these methods can be pulled together for a practical application. The aim of this introduction is to provide you with a sense of the fundamental ideas in web data collection, not to equip you with the necessary tools to apply these ideas in practice. After reading this introduction, you should understand the basic building blocks of websites and web interactions, which suffices to achieve the majority of use cases in web data collection. (The final section points to some further readings to apply the techniques in practice.)

The third section discusses some current challenges in web data collection, where we distinguish between technical and conceptual challenges and try to highlight some potential remedies. The fourth section introduces a practical political science application that operates with web data. The aim here is not to discuss the substantive findings, but rather to highlight how the authors employ the tools we describe to make a valuable contribution to the political science literature.

## THE FUNDAMENTALS OF WEB DATA COLLECTION

### *Application Programming Interfaces (APIs)*

We can distinguish between two main strategies for web data collection: collecting data by scraping websites and gathering data from application programming interfaces (APIs). By web scraping, we mean accessing a website and extracting those pieces of information from the page that we are interested in. The relevant information is typically embedded in a larger context that we do not care about. Therefore, the greatest challenge in web scraping often lies in discarding the irrelevant parts. This step is mostly done away with when we collect data from programming interfaces. Such interfaces are explicitly set up by the providers of web services to streamline, among other things, the exchange of information. Simply put, web services define a set of valid operations and data that can be retrieved from these access points.<sup>1</sup>

The main difference between data collection via web scraping and programming interfaces is the format in which we receive the data. When we access an ordinary online resource, we receive an HTML document from the server, which contains a lot of superfluous information. APIs, by contrast, provide the sought-after information in much more useful data formats that can easily be converted into conventional data matrices.

Let's consider a practical example to highlight the difference between web scraping and APIs, but also to show how both strategies exhibit clear similarities. We employ the Wikipedia API which allows us to access various pieces of information from the service.<sup>2</sup> In this case, we access the daily page views for the Mannheim entry with the following query:

```
https://en.wikipedia.org/w/api.php?action=query&
titles=Mannheim&prop=pageviews&pvipdays=7&
format=xml
```

This small example can teach us a lot about the art of gathering data from the web. Consider first the structure of the query. It looks very similar to ordinary URLs. The specific query consists of four components that you will encounter all over the web. The query begins by defining a scheme, which describes the protocol that is used to exchange messages between the user and the server, in this case the HTTPS scheme. HTTP and HTTPS are the most common protocols in web interactions and while we will spare you the technical details, they prescribe the structure of the messages that are exchanged between the user and the server. The response from the server will often contain some requested file, for example, an HTML document, a JPEG image, a PDF file or, in the present case, an XML file.

After the scheme, we specify the domain where the resource is located, in this case ‘en.wikipedia.org’. Effectively, we are looking to retrieve information from the Wikipedia server. Next, we define the path to the specific resource on the server. Here, the individual elements of the path are comparable to nested folders on your computer, such that ‘api.php’ resides within the ‘w’ folder. The last component of the URL contains the specific query, which is preceded by a question mark. The elements of the query consist of key–value pairs where the individual keys (‘action’, ‘titles’, etc.) define the parameters. Our choices are specified in the values that are linked to the keys with equality signs. The key–value pairs are separated from one another with ampersands. A summary of the four most common components of URLs are presented in Table 22.1 with examples.<sup>3</sup>

Note that while the components are identical, the keys for the search queries differ between the three examples. This is what we mean when we say that each web service defines a set of valid operations. Indeed, while both Twitter and Google Scholar place the query string in a ‘q’ key, this is a mere convention and they could just as easily have named the query parameter something else.

**Table 22.1 Common components of URLs**

<i>Component</i>	<i>Example</i>
Scheme	https:// https:// https://
Domain	en.wikipedia.org/ twitter.com/ scholar.google.com/
Path	w/api.php search scholar
Query	?action=query&titles=Mannheim&prop=pageviews&pviptime=7&format=xml ?q=metoo&src=typd ?hl=en&q=anthony+downs

Note further that we typically do not use every possible parameter, but only the ones that we need to formulate our query, while other parameters might operate as defaults in the background or they might even be appended by the server as needed. For example, our results do not change when we drop the ‘src’ parameter from the Twitter URL.

An overview of the possible parameters for the Wikipedia API can be accessed at ‘https://en.wikipedia.org/w/api.php’. Along with defining the set of valid parameters, the web services also limit the possible parameter values. For instance, the ‘format’ parameter in the initial example is likely to only have a small set of possible values. Finally, note that not all of the four basic building blocks need to be present for constructing a valid URL. For example, ‘https://stackoverflow.com/’ describes a perfectly valid resource location that only consists of the scheme and the domain, and which will retrieve the main HTML document of the Stackoverflow site.

After inspecting the basic components of URLs, let’s return to the query we made to the Wikipedia API to collect the daily page views for the Mannheim Wikipedia page. The five key–value pairs of the request are listed in Table 22.2. The five elements are quite descriptive and we can easily infer the function of each. The ‘action’ parameter specifies that we are looking to make a query to extract

**Table 22.2 Parameters of the Wikipedia query**

Key	Value
action	query
titles	Mannheim
prop	pageviews
pvipdays	7
format	xml

some kind of information. The ‘titles’ parameter defines the specific page. The ‘prop’ parameter defines the type of information we are looking for, the daily page views. The parameter ‘pvipdays’ specifies the range for the request, in this case the last seven days. Finally, the ‘format’ parameter defines the format of the response, in this case an XML-formatted document, which we will discuss in a second.

If you type in the full URL from the beginning of this section into your browser, you should receive a response that looks similar to the one that we have displayed in Figure 22.1. Before we take the response apart, let’s stick with the URL for just a moment longer. One common solution to many problems in web data collection is the systematic manipulation of the URL to access not just one page

but all pages of interest. For instance, if we were interested in the daily page views of not just one but a set of Wikipedia articles, we could simply replace the value ‘Mannheim’ in the ‘titles’ parameter with other values, e.g., ‘Konstanz’, to retrieve the page views of that page, and so on (for examples see Kämpf et al., 2015; Roll et al., 2016).

APIs typically provide a variety of information. If we dig deeper into the documentation of the Wikipedia API, we find that there are a number of possible values for the ‘prop’ parameter, associated with the ‘query’ value of the ‘action’ parameter.<sup>4</sup> For instance, many Wikipedia pages are geolocated and we can retrieve this information using the ‘coordinates’ value for the ‘prop’ parameter. Dropping the ‘pvipdays’ parameter, which we only needed to specify the time frame for the ‘pageviews’ value in the ‘prop’ parameter in the previous version of the URL, we can construct the following request:

<https://en.wikipedia.org/w/api.php?action=query&titles=Mannheim&prop=coordinates&format=xml>

Pasting this query into the browser, results in a response from the server containing the

```

▼<api batchcomplete="">
  ▼<query>
    ▼<pages>
      ▼<page _idx="99627" pageid="99627" ns="0" title="Mannheim">
        ▼<pageviews>
          <pvip date="2018-11-03" xml:space="preserve">562</pvip>
          <pvip date="2018-11-04" xml:space="preserve">593</pvip>
          <pvip date="2018-11-05" xml:space="preserve">617</pvip>
          <pvip date="2018-11-06" xml:space="preserve">528</pvip>
          <pvip date="2018-11-07" xml:space="preserve">546</pvip>
          <pvip date="2018-11-08" xml:space="preserve">517</pvip>
          <pvip date="2018-11-09" xml:space="preserve">546</pvip>
        </pageviews>
      </page>
    </pages>
  </query>
</api>

```

**Figure 22.1 Response from the Wikipedia API (Mannheim page views between November 3 and November 9, 2018)**

```

▼<api batchcomplete="">
  ▼<query>
    ▼<pages>
      ▼<page _idx="99627" pageid="99627" ns="0" title="Mannheim">
        ▼<coordinates>
          <co lat="49.48888889" lon="8.46916667" primary="" globe="earth"/>
        </coordinates>
      </page>
    </pages>
  </query>
</api>

```

**Figure 22.2** Response from the Wikipedia API (Mannheim coordinates)

XML document in Figure 22.2. The document is quite similar to the document in Figure 22.1. The only changes are in the center of the document, which now contains the latitude and longitude values associated with the Mannheim Wikipedia page.<sup>5</sup> Having familiarized ourselves with the idea of manipulating URLs to gather all the information we need, let's return to the XML document in Figure 22.1, which can teach us even more about the basics of web data collection.

The XML format is quite similar to the HTML format, hence we can learn many of the properties of the latter document type from studying the former. Additionally, many of the same tools are used to extract information from XML and HTML documents. XML documents are nothing more than plain text documents that are structured in a particular way. The structuring elements are called tags. Such tags are enclosed by a '<' and a '>' sign. Typically, an opening tag is mirrored by a closing tag, which is indicated by a '/' character. Together, an opening and a closing tag form a node in a document. This sounds much more complicated than it is. Consider as an example the '<query>' tag in Figure 22.1, which is matched by the '</query>' tag further down in the document.

One important characteristic of XML-like documents is their hierarchical structure. In a well-formatted document, an outer node encapsulates one or more inner nodes. For instance, the '<query>' node encapsulates the '<pages>' node, which encapsulates further nodes still. Nodes in an XML document

and in an HTML document in particular can also encapsulate content. In the example in Figure 22.1, the daily page view figures are included in the document between the opening and the closing '<pvip>' tags. Compare this to the output in Figure 22.2, where all the sought-after information is embedded in the document structure. Lastly, the opening tags frequently have additional parameters and values associated with them, the so-called attributes and attribute values. For instance, the '<pvip>' tag in Figure 22.1 contains the attribute 'date' with the seven date values for the Mannheim page views.

As we have pointed out, an XML document is nothing more than a structured text document. Therefore, you could easily download the document to your hard drive and open and manipulate it with a basic text editor. The version of the document in Figure 22.1, i.e., the version with the indentation, is nothing more than an interpretation of the document by the browser that is intended to improve the readability of the document. The indentation reflects the hierarchical document structure where more indented nodes reside at lower levels. Incidentally, the interpretation of the document by the browser is why the document might look slightly different for you if you have been following along with the example, as each browser may render the document slightly differently.

We began the discussion in this section by suggesting that one of the advantages of data collection via APIs over web scraping is the



fact that the data comes in much more useful data formats that can easily be converted into conventional data matrices. While neither the document in Figure 22.1 nor that in Figure 22.2 look particularly simple at first glance, they are in fact quite bare bones data formats from which the relevant pieces of information can easily be extracted. We will not elaborate how to do this here, as the actual extraction would require some minor coding and we want to remain at a more principled level in this introductory section. Suffice it to say that CSS selectors and XPath expressions are two syntaxes which have been specifically designed for the purpose of extracting information from XML-type documents. Both make use of the hierarchical structure of XML documents and their various components (tags, attributes) to extract the relevant pieces of information. (See the final section for suggestions on how to implement these ideas in practice.)

Moving from the practical back to the principle, one additional benefit of APIs over web scraping is that the provider of the information clearly encourages the data collection and often explicitly spells out the terms of usage in the API documentation, putting our collection efforts on more secure ethical and legal footing. Finally, APIs may occasionally even enable us to access data that we could not access otherwise. In sum, the clear benefits of collecting data via APIs over web scraping suggests that whenever we embark on a new data collection project, we should ascertain whether the information can be acquired through an API.

These advantages notwithstanding, there are several downsides of gathering data from programming interfaces. On a technical level, APIs are somewhat idiosyncratic, as each provider gets to define a set of valid inputs and operations with which we must familiarize ourselves before interacting with the interface. Therefore, when a user wants to collect data from a programming interface, this typically means having to engage with the documentation of the API. The need

to adapt our tools to the specific case is of course equally true when scraping data from a website, but in this case the data provider cannot place any arbitrary restrictions on our arsenal. Related to this point is the possibility that providers change the functionality of their interfaces over time, potentially breaking the code you wrote to interact with an API. At least this point is no different for web scraping, where changes in the source code of a website can easily break your code.

In addition to these rather technical challenges, greater concerns typically stem from the restrictions that providers frequently place on their interfaces. From the point of view of the providers, restrictions are perfectly reasonable, as the simplicity of querying APIs is both a blessing and a curse for them. The ease of requesting huge amounts of data in very high frequency can cause an enormous burden on the servers and certainly invites disruptive behavior by malicious users. Providers employ a number of tactics to combat such behavior that all tend to be detrimental to our collection efforts. First, many APIs require users to register their application and to provide their authentication credentials as part of their query. This is more of a hassle than a limitation which allows providers to track how often users access their services, and who they are, in order to identify bad behavior.

More problematic are the limitations which are frequently placed on the data that can be queried from the services. Consider the Wikipedia API once more for a practical example. Say we were interested in the page views for a set of pages in the last year. We could change the 'pviptdays' parameter in the initial URL to the value '365' to gather the page views for a whole year:

```
https://en.wikipedia.org/w/api.php?action=query&
titles=Mannheim&prop=pageviews&pviptdays=365
&format=xml
```

At the time of writing, when we paste this URL into the browser, we receive data for the past 60 days along with a warning message that users may not access page views for a

longer time period. Another restriction that we typically face is a cap on the number of requests we can make in a given time period or the volume of data that can be queried. Yet another restriction could be set with regard to the data that can be queried in the first place. At the time of writing, the free-of-charge version of the Twitter API not only restricts backward searches to seven days, but also restricts users to searches in a database containing only a sample of the tweets which were published in this time frame. In light of these limitations, organizations – be they civil society actors or private organizations – have occasionally set up parallel services that collect the data as it becomes available, to redistribute the data with a different set of restrictions. Naturally, the ability of these services to provide the data is highly dependent on the terms of usage, and in many cases these supplementary services come with a price tag.

We want to close this discussion by pointing out that a great many APIs have been set up to provide easy access to data of all sorts. We encourage users to take a moment and check out some services. Familiarize yourselves with the documentation to see whether you can make sense of the API properties and identify the similarities to the components that we have discussed so far. One excellent starting point is the site *ProgrammableWeb*,<sup>6</sup> which has categorized thousands of application programming interfaces. Maybe you can even come up an interesting research project when you see what type of data is available.

One final benefit of using APIs for web data collection should not be left unmentioned. So far, we have accessed APIs directly with our browser. We could also access these sites using a scripting language such as Python or R and extract whatever piece of information we are interested in with few lines of code. For many of the most popular APIs, specific modules have been written in these languages that mimic the functionality of the programming interfaces by assigning the API operations to specific functions of these modules. This is to say that if a module has been programmed for

an API, we do not have to engage with the API documentation, but can simply use the functionality of the module to extract whatever data we are looking for, such that the module turns our request into a valid query and sends it to the API to retrieve the data.

## **Web Scraping**

In the previous section, we have accessed APIs to retrieve data in the XML format. In web scraping, we typically access documents in the HTML format, the most common document format on the web. The basics of HTML are not all that different from XML, so you will be able to recognize many of the features that we have outlined in the previous section. We should point out right away that modern HTML documents are characterized by a number of features that go well beyond this basic model. We will only introduce the basic ideas in this section, point to some complications in the following section and refer you to the further readings in the final section on how to deal with these complications in real-world applications. We want to emphasize, however, that despite these complications, the majority of use cases can still be accomplished with the tools that we outline in this section.

Like XML, HTML documents are hierarchically structured with the same basic features as XML, i.e., HTML documents are composed of nodes, tags and attributes. The main difference between XML and HTML is that HTML tags have a pre-defined meaning which is the same for every web page you will ever visit. By contrast, the tags and attributes in the XML documents we considered in Figures 22.1 and 22.2, e.g., ‘<query>’, ‘<pages>’, were defined by the curators of the Wikipedia API to best structure the data that is output by this specific API.

For HTML documents, there is only a limited set of tags with a predefined meaning and the tags have attributes associated with them that further specify the nodes. This limitation

enables the interpretation of an HTML document by the browser, since there is no ambiguity regarding the purpose of a particular tag in the overall document structure. Put differently, when you open an HTML document in your browser, you are, in fact, accessing a plain text document that looks quite similar to Figures 22.1 and 22.2 and the browser simply provides you with an interpreted version of the source code.

A lot follows from these considerations for the purpose of scraping data from the web. First, the great variability of the web is really only a superficial matter. Underlying this variety are always the same structuring elements, which allow browsers to make sense of the documents. Let's elaborate this with two examples. Only a handful of tags are used to embed a table in HTML source code. The entire table is always nested inside a '<table>' node, table rows are always indicated by a '<tr>' node, and so on. This means that in order to extract tabular information from a web page, it suffices to identify the relevant '<table>' node and to extract all the information encapsulated by the node. Indeed, due to the simplicity of identifying these types of structures within HTML code, numerous convenience functions have been written for the purpose of converting tabular information in HTML source code to

conventional data matrices. Another example is the '<li>' node, which inevitably encapsulates a list. Again, if the sought-after information comes in the form of a list, it is quite simple to identify and extract the relevant elements from the HTML source code using the querying syntaxes we mentioned in the previous section – CSS selectors and XPath expressions.

Let's consider a practical example. Many studies have relied on hyperlinks in websites for inferences about networks between issues or organizations (cf. Ackland and O'Neil, 2011; McNutt and Pal, 2011). The network datasets in these studies are often impressive but the underlying technical ideas are perfectly straightforward. For the purpose of a mock example, we consider the Mannheim Wikipedia entry. An excerpt from the site is provided in Figure 22.3. Our aim is to extract all hyperlinks to other Wikipedia articles from the source code of the site to build a network of pages with Mannheim at the center of the network. In a real-world case study, we might go on to extract all links from the resulting pages and repeat the process to build a comprehensive network or we might repeat the process for all German cities over a certain population threshold and build a network for each city to analyze whether and where these networks overlap.

## Mannheim

From Wikipedia, the free encyclopedia

Coordinates: 49°29′20″N 8°28′9″E﻿ / ﻿

*This article is about the city in Germany. For other uses, see Mannheim (disambiguation).*

**Mannheim** (German pronunciation: [ˈmanhɛm] ⓘ (ⓘ listen); Palatine German: *Monnem* or *Mannem*) is a city in the southwestern part of Germany, the third-largest in the German state of Baden-Württemberg after Stuttgart and Karlsruhe with a 2015 population of approximately 305,000 inhabitants. The city is at the centre of the larger densely populated Rhine-Neckar Metropolitan Region which has a population of 2,400,000<sup>[3]</sup> and is Germany's eighth-largest metropolitan region.

Mannheim is located at the confluence of the Rhine and the Neckar in the northwestern corner of Baden-Württemberg. The Rhine separates Mannheim from the city of Ludwigshafen, just to the west of it in Rhineland-Palatinate, and the border of Baden-Württemberg with Hesse is just to the north. Mannheim is downstream along the Neckar from the city of Heidelberg.

Mannheim is unusual among German cities in that its streets and avenues are laid out in a grid pattern, leading to its nickname "die Quadratestadt" ("The City of Squares"). The eighteenth century Mannheim Palace, former home of the Prince-elector of the Palatinate, now houses the University of Mannheim.



**Figure 22.3** Interpreted version of the Mannheim Wikipedia page (excerpt), November 13, 2018

```

<p>
"Mannheim is located at the confluence of the "
<a href="/wiki/Rhine" title="Rhine">Rhine</a>
" and the "
<a href="/wiki/Neckar" title="Neckar">Neckar</a>
" in the northwestern corner of Baden-Württemberg. The Rhine separates Mannheim from the city of "
<a href="/wiki/Ludwigshafen" title="Ludwigshafen">Ludwigshafen</a>
", just to the west of it in "
<a href="/wiki/Rhineland-Palatinate" title="Rhineland-Palatinate">Rhineland-Palatinate</a>
", and the border of Baden-Württemberg with "
<a href="/wiki/Hesse" title="Hesse">Hesse</a>
" is just to the north. Mannheim is downstream along the Neckar from the city of "
<a href="/wiki/Heidelberg" title="Heidelberg">Heidelberg</a>
"
"
</p>

```

**Figure 22.4** Source code of the Mannheim Wikipedia page (excerpt, November 13, 2018)

To implement this idea in a practical web scraping application, we would typically begin by inspecting the source code of the page. All common web browsers allow you to right-click on a site and select the option of displaying the source code. Alternatively, you could download the source code to your hard drive and open the document in a text editor to check the underlying document structure. If we were to actually implement this use case, we would download the source code to our hard drive and then operate on it with a scripting language of our choice. A small portion of the page's source code, containing the second paragraph of the page in Figure 22.3, is displayed in Figure 22.4.

Once again, the indentation of the document is not part of the source code but an interpretation of the document by the browser that is intended to help human readers make sense of the document. We can clearly see the similarities between this document and the XML document we examined in the previous section. The document is hierarchically structured with opening and closing tags that encapsulate additional nodes and pieces of content. Inspecting the document a little closer, we find that the whole paragraph is enclosed by a '<p>' node. '<p>' nodes always wrap around paragraphs in HTML documents. If we were to extract all the '<p>' nodes from the document, we would get rid of much of the unwanted information, such that we would be left with almost only the content we might be looking for if our goal was to run a text analysis on the page.

As our interest only lies in extracting the internal links on the Wikipedia domain, we find that the links are – always – embedded in an '<a>' node. Specifically, the opening and closing '<a>' tags encapsulate the text that is displayed as the link to users in the interpreted version of the HTML document. Consider the first link, which reads:

```
<a href="/wiki/Rhine" title="Rhine">Rhine</a>
```

Going back to the interpreted version of the document in Figure 22.3, you will find that the first link in the second paragraph displays the encapsulated content between the two '<a>' tags: 'Rhine'. The interesting piece of information for our mock example is not the name of the link, however, but the value of the 'href' attribute in the opening tag: '/wiki/Rhine'. This sequence of characters describes the path for the Wikipedia entry on the Rhine river.

At this point we have all the technical pieces assembled to construct a link network of arbitrary size starting with the Mannheim page at its center. In the first step, we would simply write a query to extract all the '<a>' nodes and specifically the value of the 'href' attribute associated with the '<a>' nodes. From this list, we would discard all the links to resources that do not reside on the Wikipedia domain. For instance, at the time of writing, the first link to an external domain is 'https://www.mannheim.de/'. Note that external links do not complicate this discussion in any way. These links are still embedded in an '<a>' node and the link target

is still associated with the ‘href’ attribute. In this particular case, the link looks as follows.

```
<a rel="nofollow" class="external text"
xhref="https://www.mannheim.de/">www.
mannheim.de</a>
```

We note that additional attributes are associated with the ‘<a>’ node that we do not want to discuss here, but we find the ‘href’ attribute easy enough. We would shorten the remaining list of internal links – either links that specify a path on the Wikipedia domain or links that provide a full URL on the target domain – by discarding links that do not link to HTML pages, but to images or the like.

The next step in our mock research project is just as simple. We would take the URL for the Mannheim page – ‘https://en.wikipedia.org/wiki/Mannheim’ – and discard the path component from the URL ‘/wiki/Mannheim’ to replace it one by one with the new links, such as the one we identified above to construct new URLs: ‘https://en.wikipedia.org/wiki/Rhine’; then we would download the source code of the new page and start over. At this point, it is just a matter of repeating these few steps as often as we care to, to build a hyperlink network of arbitrary size.

We hope that this example has given you some intuition of how even simple web scraping tools allow us to build sizeable datasets by automating the collection tasks. Automating the tasks is greatly simplified by the fact that websites are all composed of the same building blocks. Let’s close this section with some final remarks on web scraping in practice. First, and related to the mock example we just discussed, it is not uncommon that we build our web scraping applications on the basis of an index page of one kind or another to collect a set of links to additional sites. For example, when gathering data from an online news source, we might start the data collection on an index page which contains links to a set of articles. We would first extract the links and open the individual pages in a second step to extract the content of the news articles. To be sure, the links on

such index pages may not link to HTML documents at all, but rather to PDF files or similar, which we could download with the same tools in order to extract the information from them in a second step.

As a rule of thumb, it is useful to distinguish between the collection and the extraction steps in web data collection. In practice, we are often interested in extracting data from thousands of HTML documents. Since we are bound to make mistakes in the extraction step, it is helpful to first store all the target pages on your hard drive and then attempt to extract the relevant information from the target pages. This way, we create as little traffic on the server as possible, which we should always strive for in web scraping. The additional upside of distinguishing between the collection and the extraction step is that you have a local copy of the target material, which might be valuable to ensure the replicability of your work when the data on the server changes. Before moving on, we encourage you to take a look at the source code of a couple of websites. You should already be able to recognize bits and pieces and to make sense of the overall structure of the page.

## CURRENT CHALLENGES IN WEB DATA COLLECTION

While the basic tools outlined in the previous section are not difficult to put into practice, there are a number of challenges that we currently face in our efforts to automatically collect data from the web. This section serves to elaborate some of these challenges and, where possible, hint at some solutions. To structure the discussion, we distinguish between technical and conceptual challenges.

### *Technical Challenges*

We have outlined the basic building blocks of conventional websites and how they help us

in collecting data from the web. We now have to backpedal a little and concede that these elaborations have been truer in the online world of the early 2000s than in today's web. Modern web architectures increasingly move away from this basic model and continue to add elements that make the task of the web scraper a little more challenging. Foremost among these challenges is the incorporation of non-static elements into websites. Think for instance about websites that continuously load content as you scroll through them. For a practical example, do an image search on Google and continue to scroll to the bottom of the page. You will notice that the site grows longer as you continue to scroll down to offer you additional images associated with your search term. This architecture is designed to keep you from having to push a button for additional images, ensuring a smoother browsing experience. At the same time, the designers have wisely decided that it makes little sense to load all possible pictures from the start, as this would mean loading a bunch of images without any sense of how many the user actually wants to see.

What happens is that a script is running in the background that appends the source code and loads additional pictures from the server only when it becomes necessary. The problem for web scrapers is that when we download the source code of a page, we typically only get the initial version of the source code, since we do not run the embedded scripts. One solution when faced with dynamic web pages is to mimic the user behavior with tools that allow for automating the browser in order to extract the information you are looking for.

Continuing with the example, this might mean accessing the Google image search for a particular search term in any browser with a scripting language and then running a function that continuously scrolls to the bottom of the page in order to load additional images. While this might sound technically elaborate, the tools are fairly straightforward to use. Originally developed for testing web applications, we can make use of these tools

to automate our browsers with the aim of collecting data. Nevertheless, despite these somewhat unfortunate trends in modern web development, most websites are still sufficiently static that we can achieve most of our goals in web data collection with the traditional tools.

A second technical challenge is that a website-centric view of the web is increasingly outdated. Many of the most popular services on today's web are exclusively built for the mobile screen, creating additional hurdles for the data collection. In the worst of cases, if these services do not provide access to their data via an API, this might mean having to emulate a mobile screen on your computer to run the app and access the data from there. Again, this is more of a developing challenge and, at least for now, most services still provide an accompanying website.

One final challenge is the sheer size of the data that occasionally results from our collection efforts. Especially when we engage with social media platforms where our research questions sometimes require that we continuously collect data, we might have to learn more complex tools specifically designed to cope with the data stream. One possible solution to the problem of too extensive data could be to think about how and whether to sample the data. Monroe (2013) rightly points out that for many applications, sampling is not a viable option. Indeed, even in our mock example, we probably need the full set of articles to understand the network of Wikipedia pages. That being said, collecting the full set of documents should not be the default just because we can. Before embarking on a new data collection project, researchers should think about what data they need to answer their research question. Think, for example, about traditional analyses of political communication, where scholars have gained valuable insights just by looking at a small subset of articles written about a particular subject (e.g., Eberl et al., 2017; de Vreese et al., 2006). These same considerations should guide our collection efforts when we engage with data from the web.

## ***Conceptual Challenges***

Beyond these more technical challenges, there are also conceptual concerns that need to be kept in mind when collecting data from the web. The first among these are issues related to data quality and data generation mechanisms. By collecting data from the web, we can typically sidestep the problem of data sparsity quite easily. At the same time, we need to be particularly mindful about how the data came to be on the web in the first place. Such concerns are quite apparent when we try to infer public opinion from utterances on social media. Not only are social media users far from a representative sample of the general public (Mellon and Prosser, 2017), the propensity to voice an opinion in such a public forum likely distorts the sample even further, begging the question of what we can learn about the public from social media.

Although we have seen some elegant efforts in recent years to estimate public opinion from immensely unrepresentative samples (Buttice and Highton, 2013; Wang et al., 2015), such models require a baseline of sociodemographic information to which we typically do not have access when we engage with social media posts. Lacking a baseline of socio-demographic information for weighting the posts, we have to take them at face value and hope for the best when trying to make inferences about public opinion.

A particularly telling example in this regard is the work by Tumasjan and colleagues (2011) who forecast the 2009 German federal election using social media data. Their results have been criticized for disregarding the Pirate party as the then newcomer in the German party system (Jungherr et al., 2012). Jungherr and colleagues suggest that when incorporating the Pirate party into the dataset, the forecast would have seen the Pirate party win the election, which did not even manage to cross the 5% exclusion threshold. The results by Jungherr and colleagues suggest that the Pirate party had a particularly web-savvy and outspoken group of supporters,

creating substantial biases in the dataset. To be sure, this is not to say that research on social media is not worthwhile, but it is important to keep the data generation mechanisms in mind when engaging with such data. For instance, the countless studies on elite usage of social media are much less problematic, as we can describe the population of interest and how the sample differs from the population (e.g., Ernst et al., 2017; Jackson and Lilleker, 2011; Vergeer and Hermans, 2013).

One final challenge relates less to the individual collection efforts and more to the aggregate consequences of relying on data from the web for social research. There is a considerable gap between the social science research agenda and the availability of web data, where some areas are characterized by excessive data availability while data in other areas is no less sparse today than before the advent of the web. Along with this disparity comes the concern that if we let our research focus be guided by data availability, we might disregard important research questions.

There is both a narrow and a broad variant of this concern. In the narrow version, we frequently engage with data just because it is available – not because it provides the best information for the point we are trying to make. Possibly the most well-known version of this phenomenon is the research focus on Twitter over competing social media platforms, most notably Facebook. This is not due to a collective perception among scholars that Twitter is the most important social media platform – it simply reflects the fact that Twitter provides easier access to its data. Clearly, we cannot learn all we might want to learn about public usage of social media platforms from Twitter, both because the user base differs drastically between the platforms and because platform design governs user behavior.

In a broad version of this concern, there is good reason to wonder whether data availability governs the very scope of certain research programs, such that some topics might be a little excessively researched while our efforts could be better spent in areas where data is

less easily accessible. The importance of social media for modern societies notwithstanding, one might wonder, for example, whether the ease of gathering data from social media platforms has left researchers asking too often whether social media might be an interesting road to go down. To be sure, the standards for choosing research questions are no different when engaging with web data than in other areas, so we want to close this discussion with the simple suggestion that it is important to take particular care to make the case for the theoretical and societal relevance of our research questions (e.g., Lehnert et al., 2007) and to avoid the temptation to have our research efforts be guided by what data we can easily access.

## AN APPLIED RESEARCH EXAMPLE

Despite the technical and conceptual challenges briefly sketched above, scholars have, of course, put web data to good use in numerous exciting applications. In this section we want to highlight one example from the literature in greater detail. Our emphasis here will be on the question of how we might replicate the study and not so much on the results. For this discussion, we have picked the study by Lucas and colleagues (2015). In their work, the authors provide two applications of the Structural Topic Model (cf. Roberts et al., 2014), a model for the quantitative analysis of text. Quantitative text analysis frequently goes hand in hand with web data collection, as data from the web often takes the shape of more or less structured text (cf. Grimmer and Stewart, 2013; Wilkerson and Casas, 2017). The aim of the model is to assign topical categories to a collection of texts where the researcher does not pre-structure the results much beyond specifying the number of topics to be estimated. We will not discuss the specifics of the model any further, as we are only interested in how the researchers have collected and processed the data for their application.

Lucas and colleagues provide two applications. In the first application, they classify a set of teachings by Muslim clerics. We will focus on the second application, where the authors analyze a sample of Chinese and Arabic social media posts that deal with the revelations of the NSA surveillance programs by Edward Snowden. Lucas and colleagues aim to assess how users in China and in the Arab world have reacted to this story and whether the two discourses were systematically different.

The authors begin by assembling data from Twitter, for the Arab-speaking social media posts, and from Sina Weibo, as the Chinese equivalent, containing mentions of Snowden in the period between June 1 and June 30, 2013. Lucas et al. collect their data from a third-party data provider, Crimson Hexagon, not least since the APIs for both platforms are quite restrictive in terms of how far back users can collect data. Nevertheless, in principle the same data could be collected from the companies directly, as they provide access to their data via programming interfaces. We encourage you to check out the API documentations for Twitter<sup>7</sup> and Sina Weibo.<sup>8</sup> Both require the creation of an account to start collecting data, but otherwise it is fairly straightforward to gather data from these platforms and to replicate the analysis by Lucas et al. for a topic of choice.

To run the analysis, the authors convert the texts into one language in order to run a single, comprehensive model on all social media posts. They choose to convert the texts into English, for which they make use of the Google translation algorithm. They send their queries to the Google server using the `translateR` add-on in R, but it is simple enough to write out the request by hand without having to rely on this module. We have provided a valid query to the Google translation API and we can easily identify the URL components we have introduced above:

```
https://www.googleapis.com/language/translate/v2?key=userkey&q=web%20data%20collection%20is%20surprisingly%20simple&source=en&target=fr
```



The URL contains a scheme ('https'), a domain ('www.googleapis.com'), a path ('language/translate/v2') and the specific query with four elements: 'key', 'q', 'source' and 'target'. 'q' contains the specific sentence to be translated. The only new element here is the fact that the space characters in the query are replaced with the sequence '%20', as spaces are invalid characters in URLs. The query reads: 'web data collection is surprisingly simple'. The 'source' parameter tells the server what language the query is written in (English); 'target' is the language to be converted into (French). The only element that is not a true representation of the actual query is the value of the 'key' parameter. Similar to the two previous APIs, Google requires that users register with the service before using it. Registering is quite straightforward and we encourage readers to create an account to replicate the example.<sup>9</sup> When you replace the 'userkey' value in the 'key' parameter with your own private key and paste the URL into your browser, you will get Figure 22.5 as the response.

My French is not quite what it used to be, but as far as I can tell the sentence expresses the same sentiment as the input sentence. From here, it is simply a matter of running all messages through the translation service and running the quantitative text analysis in a second step. The example highlights once more that interacting with application programming interfaces is absolutely straightforward and provides quite simple access to data, or in this case to services that operate on data. The great simplicity and power of automated web data collection stems from the fact that once

you have figured out how to run an operation once, it is simple enough to generalize the operation to run it many times to build large datasets with few lines of code.

## WEB DATA COLLECTION IN PRACTICE

To conclude this chapter, we would like to provide readers with some guidance as to where they can find help for a practical web data collection application. While there are dedicated programs and platforms built around specific APIs, we discourage the use of such black box solutions for the purpose of web data collection. Instead, it is useful to learn a programming language to be able to adapt the tools to a variety of scenarios.

We recommend the use of the R programming language for several reasons. First, R has made tremendous strides in recent years in the area of web data collection and a number of packages have been published that allow you to accomplish a variety of tasks in web data collection. Additionally, a number of packages have been written in R to interact with APIs, such that you can access these services from within R. The benefit of R for social scientists in particular is that the language has become enormously popular in the area of statistical computing, such that a number of readers might already be familiar with the language, allowing you to collect the data without having to learn an entirely new language. Finally, the strength of R in the area of data analysis and data visualization

```
{
  "data": {
    "translations": [
      {
        "translatedText": "la collecte de données Web est étonnamment simple"
      }
    ]
  }
}
```

**Figure 22.5** Response from the *Google Translate* server to the query

enables running the whole analysis from start to finish in the same environment.

Even though the benefits of web data collection should be evident to most scholars with an interest in quantitative research and even though the underlying technology has not dramatically changed for many years, there are surprisingly few comprehensive treatments of the subject. One book-length discussion of the ideas introduced in this chapter is provided by Munzert and colleagues (2014). Using the R language, the volume provides an applied introduction to web data collection, while also addressing many of the more complex problems not addressed in this chapter, such as interacting with forms, providing authentication credentials, or dealing with dynamic websites.

A second, and similar, volume is the work by Nolan and Temple Lang (2014). The book differs from the work by Munzert and colleagues in that it is a little less applied and possibly less useful for beginners. At the same time, the book provides assistance even in edge cases that may not be covered by Munzert et al. (2014). As pointed out, however, the great majority of data collection scenarios can be accomplished with the most basic tools, making the interaction with the more fringe tools a somewhat excessive exercise for most ordinary users. It should be pointed out that both volumes have suffered a little from the more recent developments in the area of web data collection, as some of the current core packages in R were not available at the time of the books' writing and are therefore not addressed in either volume. Nevertheless, as the underlying principles for web data collection have not changed, it is still worthwhile to check out how the general ideas sketched out above are implemented in practice.

Finally, while we recommend the use of R for web data collection, R is far from the only language that can be used for the task of gathering data from the web. A second and frequently used scripting language is Python, which also provides a number

of well-developed packages in the area of quantitative text analysis, which is often of interest for analyzing data from the web. An excellent introduction on how to collect data from the web using Python is provided by Mitchell (2015).

## Notes

- 1 Application programming interfaces have many applications beyond providing access to data. We restrict our discussion to those cases that are most relevant for social scientists interested in web data collection.
- 2 Later in this section, we highlight a resource that provides an overview of existing APIs.
- 3 The second example searches tweets containing the phrase 'metoo': <https://twitter.com/search?q=metoo&src=typd>. The third example searches for entries on Google Scholar for 'anthony downs': <https://scholar.google.com/scholar?hl=en&q=anthony+downs>. Note that the empty space in the search query in the latter example is replaced by a '+' sign, as empty spaces are not valid in URLs. As we are only interested in highlighting the similarities between different URLs, we refrain from discussing the parameters of the additional examples. We encourage you to check out some examples of URLs in the real world to see whether you are able to identify the components.
- 4 <https://en.wikipedia.org/w/api.php?action=help&modules=query>
- 5 Incidentally, we note that the output contains the parameter 'globe' with the parameter value 'earth', so apparently Wikipedia is ready to go for the human colonization of Mars.
- 6 <https://www.programmableweb.com/apis>
- 7 <https://developer.twitter.com/en.html>
- 8 [http://open.weibo.com/wiki/API文档\\_V2/EN](http://open.weibo.com/wiki/API文档_V2/EN)
- 9 <https://console.cloud.google.com/>

## REFERENCES

- Ackland, Robert, and Mathieu O'Neil. 2011. 'Online Collective Identity: The Case of the Environmental Movement.' *Social Networks* 33(3): 177–90.
- Alvarez, R. Michael, ed. 2016. *Computational Social Science: Discovery and Prediction*. Cambridge: Cambridge University Press.

- Behnke, Joachim, Andreas Blätte, Kai-Uwe Schnapp and Claudius Wagemann, eds. 2018. *Computational Social Science: Die Analyse von Big Data*. Baden-Baden: Nomos.
- Buttice, Matthew K., and Benjamin Highton. 2013. 'How Does Multilevel Regression and Poststratification Perform with Conventional National Surveys?' *Political Analysis* 21(4): 449–67.
- Eberl, Jakob-Moritz, Hajo G. Boomgaarden and Markus Wagner. 2017. 'One Bias Fits All? Three Types of Media Bias and Their Effects on Party Preferences.' *Communication Research* 44(8): 1125–48.
- Ernst, Nicole, Sven Engesser, Florin Büchel, Sina Blassnig and Frank Esser. 2017. 'Extreme Parties and Populism: An Analysis of Facebook and Twitter across Six Countries.' *Information, Communication and Society* 20(9): 1347–64.
- Grimmer, Justin, and Brandon M. Stewart. 2013. 'Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts.' *Political Analysis* 21(3): 267–97.
- Jackson, Nigel, and Darren Lilleker. 2011. 'Microblogging, Constituency Service and Impression Management: UK MPs and the Use of Twitter.' *Journal of Legislative Studies* 17(1): 86–105.
- Jungherr, Andreas, Pascal Jürgens and Harald Schoen. 2012. 'Why the Pirate Party Won the German Election of 2009 or the Trouble with Predictions: A Response to Tumasjan, A., Sprenger, T. O., Sander, P. G., & Welpel, I. M. "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment".' *Social Science Computer Review* 30(2): 229–34.
- Kämpf, Mirko, Eric Tessenow, Dror Y. Kennett and Jan W. Kantelhardt. 2015. 'The Detection of Emerging Trends Using Wikipedia Traffic Data and Context Networks.' *PLoS One* 10(12): 1–19.
- Lehnert, Matthias, Bernhard Miller and Arndt Wonka. 2007. 'Increasing the Relevance of Research Questions: Considerations on Theoretical and Social Relevance in Political Science.' In *Research Design in Political Science: How to Practice What They Preach*, eds Thomas Gschwend and Frank Schimmelfennig. Houndmills: Palgrave Macmillan, 21–38.
- Lucas, Christopher, Richard Nielsen, Margaret E. Roberts, Brandon M. Stewart, Alex Storer and Dustin Tingley. 2015. 'Computer-Assisted Text Analysis for Comparative Politics.' *Political Analysis* 23(2): 254–77.
- McNutt, Kathleen, and Leslie A. Pal. 2011. '"Modernizing Government": Mapping Global Public Policy Networks.' *Governance* 24(3): 439–67.
- Mellon, Jonathan, and Christopher Prosser. 2017. 'Twitter and Facebook Are Not Representative of the General Population: Political Attitudes and Demographics of British Social Media Users.' *Research and Politics* 4(3). <https://doi.org/10.1177/2053168017720008>
- Mitchell, Ryan. 2015. *Web Scraping with Python: Collecting Data from the Modern Web*. Beijing: O'Reilly.
- Monroe, Burt L. 2013. 'The Five Vs of Big Data Political Science: Introduction to the Virtual Issue on Big Data in Political Science.' *Political Analysis* 21(V5): 1–9.
- Munzert, Simon, Christian Rubba, Peter Meißner and Dominic Nyhuis. 2014. *Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining*. Hoboken: Wiley.
- Nolan, Deborah, and Duncan Temple Lang. 2014. *XML and Web Technologies for Data Sciences with R*. New York: Springer.
- Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson and David G. Rand. 2014. 'Structural Topic Models for Open-Ended Survey Responses.' *American Journal of Political Science* 58(4): 1064–82.
- Roll, Uri, John C. Mittermeier, Gonzalo I. Diaz, Maria Novosolov, Anat Feldman, Yuval Itescu, Shai Meiri and Richard Grenyer. 2016. 'Using Wikipedia Page Views to Explore the Cultural Importance of Global Reptiles.' *Biological Conservation* 204: 42–50.
- Tumasjan, Andranik, Timm O. Sprenger, Philipp G. Sandner and Isabell M. Welpel. 2011. 'Election Forecasts with Twitter: How 140 Characters Reflect the Political Landscape.' *Social Science Computer Review* 29(4): 402–18.
- Vergeer, Maurice, and Liesbeth Hermans. 2013. 'Campaigning on Twitter: Microblogging and Online Social Networking as Campaign Tools in the 2010 General Elections in the Netherlands.' *Journal of Computer-Mediated Communication* 18(4): 399–419.

- de Vreese, Claes H., Susan Banducci, Holli A. Semetko and Hajo G. Boomgaarden. 2006. 'The News Coverage of the 2004 European Parliamentary Election Campaign in 25 Countries.' *European Union Politics* 7(4): 477–504.
- Wang, Wei, David Rothschild, Sharad Goel and Andrew Gelman. 2015. 'Forecasting Elections with Non-Representative Polls.' *International Journal of Forecasting* 31(3): 980–91.
- Wilkerson, John, and Andreu Casas. 2017. 'Large-Scale Computerized Text Analysis in Political Science: Opportunities and Challenges.' *Annual Review of Political Science* 20: 529–44.

# How to Use Social Media Data for Political Science Research

Pablo Barberá and Zachary C. Steinert-Threlkeld

Citizens across the globe spend an increasing proportion of their daily lives on social media websites, such as Twitter, Facebook or Instagram. Their activities on the sites generate granular, time-stamped footprints of human behavior and personal interactions, sometimes with longitude and latitude coordinates. A sizable proportion of these digital traces have to do with politics – social media is an increasingly popular source of political news, as well as a forum for political debates on which virtually every political candidate running for office is now present.

The data generated from these interactions is in many cases freely available for research purposes and provides a depth and breadth that was unimaginable even one decade ago. Its high degree of spatial and temporal granularity allows the study of behavior at low levels of aggregation but also at a more macro scale and from a comparative perspective. The fact that human behavior is observed

unobtrusively also facilitates collecting data at a larger scale and reduces certain types of biases. This set of advantages makes social media data a new and exciting source of information to study key questions about political and social behavior.

At the same time, the volume and heterogeneity of this new type of data present unprecedented methodological challenges. Several sources of bias can limit the generalizability of our findings. Often it is difficult to connect online interactions with offline behavior. Unlike data created by governments or collected by researchers, the data-generating process is not always known. For example, we don't know if the platform was running a randomized experiment at the time of data collection or whether content was being blocked by the internet provider. And, of course, many scholars have raised concerns about the ethics of collecting data from individuals without, in some cases, obtaining their informed consent.

In this chapter, we provide a fair assessment of these advantages and limitations in the use of social media as data generators for research in the social sciences, with a particular emphasis on the study of political behavior. We illustrate these strengths and weaknesses with examples from two types of studies: those where social media is being used merely as a source of data, and those where the focus is on how social media is transforming different political phenomena. The first group includes research that uses social media to measure public opinion, the ideology of citizens and elites, the structure of social networks, government censorship, conflict dynamics and elite rhetoric, and where social media sites are used as a new space to conduct affordable field experiments. The second set of studies deals with questions such as how social media platforms contribute to the success of collective action events, how they are transforming election campaigns and whether their usage is contributing to greater political polarization and the spread of misinformation.

Throughout the chapter, we consider ‘social media data’ as any type of information obtained from websites whose value primarily comes through user interactions. The best-known examples include Facebook, Instagram, Twitter and Sina Weibo, all of which provide infrastructure to facilitate users sharing, and commenting on, content. The content can be user generated, such as status updates or photos from personal devices, or created by a third party, such as a newspaper, and shared by users. Excluded from this definition are sites such as reddit and YouTube, whose primary purpose is to surface content from elsewhere and comment on it. Despite that stipulation, the issues described throughout this chapter apply almost equally to these social media-like websites.

To further explore the opportunities and limitations of social media data, we also provide an in-depth description of an applied example that uses Twitter data to study the dynamics of protest movements in Egypt and Bahrain in 2011.

## ADVANTAGES OF SOCIAL MEDIA DATA

Perhaps the most important advantage of social media compared to traditional sources of data in the social sciences, such as surveys or government records, is the ability to unobtrusively collect information about a large sample of individuals with only minimal costs. Being able to observe subjects in a real-world environment where they are engaging in ordinary interactions reduces the likelihood that individuals change their behavior in response to knowing they’re part of a social science research project, which minimizes Hawthorne and social desirability biases.<sup>1</sup> Although the set of APIs available for academic research has shrunk in recent years (Freelon, 2018), large amounts of such data are still freely available at a minimal cost, which facilitates comparative and longitudinal analysis of any phenomenon.

This new ability to collect information about human behavior at scale is revolutionizing the state of the art across many fields. One such example is the study of social networks. This literature had traditionally relied on either small-scale networks, such as those related to high-school classrooms (Moreno, 1934) or social groups (Zachary, 1977), or on partial views of a network reconstructed through survey responses to questions about social ties, such as in the study of communication networks (Huckfeldt et al., 2004; Mutz, 2002). Large-scale network datasets from social media sites have offered new evidence to answer some of the key standing questions within this field (see e.g. Bakshy et al., 2015; Ugander et al., 2011; Lerman et al., 2012).

A second notable advantage of social media data is that its homogeneity in format and content facilitates systematic comparisons for different types of actors, across multiple countries, and over time. This homogeneity applies to multiple dimensions: constraints on text length; the use of similar language and textual marks (e.g. hashtags); the ability

of different types of actors to interact directly (e.g. via shares or retweets), among other many platform affordances and features. This strength of social media data is already leading to important breakthroughs in the analysis of political mobilization (Anastasopoulos and Williams, 2017; Freelon et al., 2016) and agenda-setting dynamics (Barberá et al., 2019), as well as in the comparative study of political polarization (Bright, 2018) and censorship (Hobbs and Roberts, 2018).

Third, social media data offers unparalleled temporal and spatial granularity. It allows researchers to observe longitudinal trends to identify events and also to study geographic patterns. This advantage can be especially relevant in autocracies or conflict areas, since it would be difficult or impossible to gather real-time data in such situations. Social media data have proven particularly useful for understanding the dynamics of Syria's protests and subsequent civil war (O'Callaghan et al., 2014; Freelon et al., 2015; Kostyuk and Zhukov, 2017); the same is true for Ukraine (Gruzd and Tsyganova, 2015; Zhukov, 2015; Wilson, 2017; Driscoll and Steinert-Threlkeld, 2020). In the last section of this chapter we offer an additional example that focuses on protests in Egypt and Bahrain leading to the Arab Spring. For technical discussions of using temporal data to detect events, see the computer science literature on event detection (Sakaki et al., 2010; Cadena et al., 2015; Alsaedi et al., 2017) or work by Golder and Macy (2011) on the relationship between diurnal and seasonal mood patterns. For the possibility of social media data for the generation of political event data, see Zhang and Pan (2019) and Steinert-Threlkeld (2019).

Fourth, despite obvious concerns about the representativeness of samples obtained from social media (a point to which we return next), for some groups of actors virtually the entire population is present on social networking platforms. For example, more than 85% of world leaders have active Twitter or Facebook accounts (Barberá and Zeitzoff,

2017) and virtually all members of the US Congress also maintain profiles on these sites (Pew Research Center, 2017), making it possible to make externally valid inferences about their communication strategies and rhetoric with data obtained from these accounts.

As we consider samples of ordinary citizens, it is worth emphasizing that representativeness of social media users is only a concern if the behavior under study is thought to vary according to variables by which users on social media differ from the population from which they are drawn. For example, Malik et al. (2015) find that a 1% increase in the population size of a census block group correlates with 5.7% fewer geolocated Twitter users. As a result, using Twitter to understand behaviors correlated with living in rural areas, such as sentiments toward a trade war or support for political candidates, is likely to underestimate those behaviors in the population.<sup>2</sup> Similarly for studying older or poorer Americans.

On the other hand, many behaviors should be less sensitive to non-representative samples. For example, factors affecting information contagion – weak and strong ties, the influence of media and celebrities, etc. – probably do not vary by age, location or race. During the Arab Spring, users of social media were certainly not representative of their populations (Ghannam, 2011; Breuer, 2012); the vast majority of protesters in Tahrir Square, for example, learned about the protests from satellite television or face-to-face communication (Wilson et al., 2012). It is not clear, however, why patterns of information diffusion found on Twitter at this time (Brym et al., 2014; Borge-Holthoefer et al., 2015; Steinert-Threlkeld, 2017b) should be assumed to be distinct from information spreading outside of social media. Other outcomes, such as diurnal patterns of behavior or user to user interactions, for example, should also not vary on observables.<sup>3</sup>

The fifth advantage of social media data, that many well-known offline behaviors also

occur online, supports the claim – representativeness issues matter, but probably not as much as feared – of the previous paragraph. Social media users’ ideology is recoverable from the structure of their social network and maps onto offline estimates of ideology (Barberá, 2015; Bond and Messing, 2015). Dunbar’s Number (Dunbar et al., 2015), attitudinal homophily (Bliss et al., 2012), diurnal activity (Golder and Macy, 2011) and geographic constraints (Takhteyev et al., 2012) occur offline and online, and Facebook users’ scores on the Big Five personality traits (extraversion, neuroticism, conscientiousness, agreeableness and openness to experience) are recoverable from their behavior on the site (Gosling et al., 2011). Twitter users also exhibit homophily with respect to age, gender and political affiliation (Zamal et al., 2012), and homophily of likes on Facebook permits similar behavioral and personality inference (Kosinski et al., 2013a). The preponderance of evidence suggests it is reasonable to expect that many, perhaps most, behaviors studied using social media are analogues to what would be observed offline, if it were possible to observe these behaviors at scale offline. It is rarely possible, of course, to observe these behaviors at scale offline.

## LIMITATIONS OF SOCIAL MEDIA DATA

A fair assessment of how social media data can be used in social science research requires also a discussion of its key limitations, as well as potential ways to address them. Probably the most important challenge to overcome, as we discussed in the previous section, is that social media users are not a representative sample of the population in any given country. As a result, an analysis of social media data without any type of adjustment may reveal patterns that are not generalizable. Despite this valid concern, it is worth making two points. First, samples that are not

representative can still be scientifically relevant. For example, we may treat the sample of users posting about politics on Twitter as a set of ‘opinion leaders’ that can be more influential than other ordinary citizens. Second, there are actually different types of sampling bias, and it is important to quantify them. We could have sampling bias whenever sociodemographic characteristics are correlated with *both* our outcome variables and the propensity to be present on social media. But there may also be self-selection *within* samples of social media users, particularly when samples are collected at the tweet level, based on whether it mentions or not a set of keywords. For example, if our sample includes tweets mentioning names of parties or political candidates, it will oversample individuals with extreme political identities, because they tend to tweet about politics more frequently (Barberá and Rivero, 2015).<sup>4</sup>

One potential solution to this set of problems would be to apply similar weighting methods as in survey research, where low response rates can lead to similar concerns about sampling bias. However, we are limited by what Golder and Macy (2014: 141) call the ‘privacy paradox’: ‘[social media] data are at once too revealing in terms of privacy protection, yet also not revealing enough in terms of providing the demographic background information needed by social scientists.’

A different type of sampling issues that is also worth mentioning is those due to black-box proprietary sampling algorithms used by social media companies. For example, Morstatter et al. (2013) showed that the Twitter API does not return a truly random sample if we are collecting from the ‘spritzer’ and not sampling using keywords.<sup>5</sup> Why this difference occurs, however, is unclear from existing work, as Twitter populates the spritzer based on a tweet’s millisecond timestamp (Pfeffer and Mayer, 2018). More work is needed, however, to determine how sensitive this result is to studying particular



topics or times, as well as to whether the observed differences are relevant enough to be significant.

A different type of challenge lies in the ability to connect online and offline behavior. The fact that we observe a pattern in an online setting does not mean that it would necessarily replicate in an offline setting. This could be due to affordances of the platforms – e.g. the fact that anonymity is easier in online settings makes vitriol and incivility more likely to occur in interpersonal communications – or to the types of ties that people develop on social media. However, as discussed in the previous section, there is increasing evidence that this is not the case. For example, two recent studies demonstrate that the structure of networks and the nature of social ties are similar in social media and offline networks (Bisbee and Larson, 2017; Jones et al., 2013).

An additional difficulty of conducting network analysis using social media data, especially Twitter, is defining what an *edge* is. Connections form more easily on social media than offline, so it is common to infer edge strength via other behavior. On Facebook, appearing together in a photograph or tagging someone in a status is commonly used to define ‘true’ friends. On Twitter, researchers will infer an edge if a user retweets or mentions another user; for topic modeling, edges usually come from the co-occurrence of hashtags. For Twitter, retweet and user mentions are preferred because they can be easily extracted from the streaming API, whereas finding an account’s followers requires more work and is severely hindered by rate limits which Twitter imposes on how frequently a researcher can request data. A shortcoming of this approach is that it usually creates cross-section data, removing any temporal information about a network. Steinert-Threlkeld (2017a) introduces a method that works within the Twitter rate limits to measure an account’s influence as it changes every day. This approach requires more engineering than relying on retweets

and user mentions delivered via the REST API, but it permits longitudinal observation.

A third limitation of social media data is replicability, for three reasons. First, sharing raw data is difficult. Twitter, the most common social media platform studied, only allows an individual to share 50,000 tweets per day ‘via non-automated means’; this quantity is not large.<sup>6</sup> For studies with millions of tweets, sharing tweets therefore requires either moderate amounts of programming to build a front-end to the dataset to ensure replicators do not receive more than 50,000 tweets per day or a human in the loop for however many days are required to share the entire dataset. Neither option is easy, and both detract from research productivity.<sup>7</sup> The best method for sharing a full dataset is therefore to share the identification number of each tweet, as Twitter has no cap on that quantity *when used for academic research*.<sup>8</sup> For Python and R scripts to download tweets using their ID numbers, see Steinert-Threlkeld (2018).

Sharing tweet IDs invokes the second reason, which is programmer skill. Because downloading tweets requires connecting to Twitter’s application programming interface (API), and the returned data is made available in a format (JSON) with which many researchers may not be familiar, replication requires programming skills that many replicators may not possess. A new R package, *rehydratoR*, simplifies this process. It takes a list of tweet IDs and saves the results to .csv files, in addition to accounting for Twitter rate limits (Coakley et al., 2019).

The third reason is ‘post rot’. If a user deletes a tweet or account, that tweet (and all tweets from that account) is not later recoverable. A tweet included in a study may therefore not be available to a later replicator. While the rate of tweet or user rot is not known, as far as we are aware, anecdotal reports suggest it is about  $\frac{1}{10}$  of tweets after one year. Whether or not this rot changes a replicator’s inference is not known, although there is some evidence that rot does not occur randomly (Timoneda, 2018). Even if the

overall findings may not change, point estimates probably will.

If replicators start with the aggregated data that authors use for their analysis, then these three reasons are obviated. Restrictions from institutional review boards also often require that authors do not share tweet IDs because it would allow other researchers to identify the subjects in the study. In that case, sharing aggregated data is the only approach to replication. This outcome is especially the case in experiments on social media (Munger, 2017; Siegel and Badaan, 2018).

Ethical concerns are, of course, important beyond the discussion about replicability. A controversial recent study that experimentally manipulated the content of Facebook users' news feed to explore whether emotions are contagious (Kramer et al., 2014) opened a debate about what the notion of informed concern means in an environment in which companies constantly manipulate features of their product using experimental methods. Even observational studies with publicly available data can raise ethical concerns because individuals may not have the expectation that their data is being analyzed for research. Even if data is de-anonymized, often it is possible to re-identify personal data (Zimmer, 2010). It is important to note, however, that this problem is present in any other study that uses individual-level data, including those that use survey data. Some recent work in the area of differential privacy (Dwork, 2008) offers exciting new possibilities to strike a balance between data access and privacy.

The fifth limitation is norms of users and companies. On user norms, the primary difference between Facebook and Twitter, and one that is not often appreciated, is that users on Twitter maintain a norm of public production while most users on Facebook opt for privacy. A small minority of Twitter users maintain protected profiles, making their data as inaccessible as the vast majority of Facebook users' data. Companies also maintain their own norms around how much

data to expose via their APIs, so the kind of research possible is at the whim of companies' API policies. For example, Facebook and Instagram used to provide much more user and graph information via their APIs, but they have become much more restrictive in response to recent controversies about data protection. Twitter has made it more difficult to obtain a developer account, the first step to collecting data, though changes to its API have otherwise been positive.

## USING SOCIAL MEDIA TO MEASURE POLITICAL BEHAVIOR

To illustrate the strengths and limitations of the use of social media data in the social sciences, we now offer an overview of recent research within this subfield. Some of the earliest work that took advantage of the opportunities offered by this new source of data was focused on measuring public opinion. Initial claims that lauded the potential of tweets to predict election results (Tumasjan et al., 2010) were soon rebuked by more systematic analysis demonstrating that those predictions were in many instances no better than random guesses (Gayo Avello et al., 2011). With some notable exceptions, such as the work of Ceron et al. (2014), the probability that we will be replacing public opinion surveys with metrics based on social media still seems unlikely (Klašnja et al., 2016). However, there is clear value in social media as a *complement* to survey data, both as an early indicator of changes in public opinion and as a possible signal on unpolled topics or areas (Beauchamp, 2017).

Even if aggregate-level opinion is hard to measure, there is plenty of evidence that social media can reveal a lot of information about citizens' characteristics and behavior. Kosinski et al. (2013b) and Youyou et al. (2015) found that Facebook likes are highly predictive of private traits such as party preference, age, gender, sexual orientation,

alcohol use, psychological traits, etc. It was this type of analysis that Cambridge Analytica allegedly used to target voters with personalized ads through social media, although there is no evidence regarding its effectiveness. Regarding the political domain more specifically, Barberá (2015) and Bond and Messing (2015) demonstrated that the political accounts that each citizen likes on Facebook or follows on Twitter offer enough data to estimate political ideology on a continuous scale (liberal to conservative) with high accuracy. Other recent work (Radford and Sinclair, 2016) suggests that the text of social media messages could also be used to estimate political preferences.

Since most autocratic governments around the world try to exercise some type of control over social media sites, we can also use data from these sources to better understand digital repression strategies. For example, King et al. (2013, 2014) found that it is possible to 'reverse engineer' online censorship using a sophisticated system that scrapes social media in real-time and then checks what has been deleted and why. Their analysis revealed that the Chinese government prioritizes deleting content that could lead to collective action but that criticism of the government is allowed. Another way in which governments can exercise digital repression is through trolls or bots that flood a platform with pro-regime information or with topics to distract from a politically sensitive issue. By muddying the waters, this tactic allows the government to silence dissidents' attempts at political coordination while maintaining plausible deniability that censorship has occurred (Little, 2015; King et al., 2017; Keller et al., 2017). These tactics also make it difficult for researchers to separate true beliefs from pro-government noise. One potential solution is to identify bots using bot detection methods (Ferrara et al., 2016), although this becomes more difficult if we consider trolls, which are accounts controlled by humans.

In the same way that press releases or campaign speeches are often used to examine

elite rhetoric (see e.g. Grimmer et al., 2012), tweets and Facebook posts by politicians can also be a useful source of data to understand political communication. The series of reports on congressional rhetoric released by the Pew Research Center are a good example. There is also a growing body of comparative work on topics such as campaign strategies or populism (Theocharis et al., 2016; Nulty et al., 2016; Stier et al., 2017), including for instance politicians' attempts to overwhelm social media with false messages (a technique also known as *astroturfing*) so that challengers' support appears smaller than it may actually be (Munger et al., 2018). It is also possible to recover politicians' ideology from the images they share on Facebook, which may be especially useful for estimating policy positions of challengers who do not have a voting record or sizable donations Xi et al., 2020.

At a more aggregate level, social media can reveal conflict dynamics with detail unavailable via other approaches. For example, students of repression and dissent generally code repression as a binary or categorical variable. Advances in computer vision techniques can be applied to thousands of protest photos to generate a continuous measure of state and protester violence Steinert-Threlkeld et al., 2020. These techniques can also generate estimates of the race and gender of participants, as well as the number of participants (Sobolev et al., 2020). Though social media tend to have some sort of location bias, it is reasonable to expect that this bias is less than in newspapers because of the lower barriers to entry and unlimited publication space (Steinert-Threlkeld, 2019). Social media data can also reveal daily changes in interstate conflict dynamics (Zeitsoff, 2011; Zeitsoff et al., 2015; Zeitsoff, 2018) and civil wars (Zhukov, 2015; Driscoll and Steinert-Threlkeld, 2020) as well as document more protests in China than newspapers have, by an order of magnitude (Zhang and Pan, 2019). It also has been used to study Black Lives Matter protests in the United States (Anastasopoulos and Williams, 2017; Chen et al., 2017).

Finally, social media can also be a space to conduct affordable field experiments. In a pioneering article, Munger (2017) developed a way to deploy experiments on Twitter using bots to convey treatments. His study discovered that, when users who use racist slurs are exposed to a bot that criticizes their actions, they change their behavior, but only when the ‘punishment’ is coming from someone similar to them. Compared to field experiments deployed in an offline environment, administering treatments has a much lower cost, although it requires an effort in terms of hardware and programming skills. Building a bot to administer a treatment is more challenging than passively collecting data from an API endpoint. Despite this limitation, this approach represents a promising blueprint for future experimental studies with high external validity (Siegel and Badaan, 2018; Coppock et al., 2016).

### **UNDERSTANDING HOW SOCIAL MEDIA AFFECTS POLITICAL BEHAVIOR**

An important part of the ongoing research that uses social media data focuses not necessarily on using these sites as a source of information about behavior, but instead on their potential as a transformative political force whose effects we are only starting to understand. There’s perhaps no better example of this type of work than the efforts to understand how digital technologies were a catalyst in the recent wave of protests around the world, starting with the Arab Spring. Some of the earlier optimism about the democratizing power of social media (see e.g. Valenzuela, 2013) was followed by a wave of skepticism. Authors such as Gladwell (2010) and Morozov (2012) warned against the rise of ‘slacktivism’ fueled by social media sites, which facilitate engagement in protests but also disincentivize commitment and the type of dedicated and trained

activism that can turn revolutionary fever into action. However, recent empirical research in political science has demonstrated that it is precisely the ability of social media to bring in peripheral individuals without resources or deep engagement that can lead to the success of collective action event (Barberá et al., 2015a; Steinert-Threlkeld, 2017b).

This evolving narrative about the impact of technology on protest mirrors the public discussion about how social media has been used in political campaigns. Barack Obama’s election campaigns are widely acknowledged to have brought the data revolution to politics (Kreiss, 2012). An important part of his success lay in understanding the native digital audience and deploying an extensive data collection and analysis platform. In contrast, Mitt Romney’s campaign famously required 22 people to approve every single tweet from the candidate’s account (Kreiss, 2016). But the optimism about how digital technologies can empower grassroots movements was put to the test after Donald Trump’s election in 2016, and the alleged use by his campaign of micro-targeted ads that were tailored to voters’ preferences. Although there is still a significant lack of research about whether targeted advertising on social media can be effective at mobilizing or persuading voters, the existing work generally finds null or very small effects, giving us reasons to be skeptical (Broockman and Green, 2014; Eckles et al., 2018; Kalla and Broockman, 2018; Bond et al., 2012).

Another key research question in this field has been whether news consumption and political conversations through social media may be one of the factors explaining the recent rise in political polarization and extremism around the world. A common argument in the literature is that social networking sites make it easier for citizens to isolate themselves into communities of like-minded individuals, where political agreement is the norm and individuals can avoid being exposed to any opinion that may

challenge their ideological views (Sunstein, 2018). This process could be exacerbated by ranking algorithms that filter out any content that users may dislike (Pariser, 2011). These concerns are not new – Putnam (2000) already expressed concerns about cyberbalkanization nearly two decades ago – but they have re-emerged and are often identified as an additional factor explaining the recent rise in political polarization. Despite these concerns, empirical evidence that this process is happening is scarce: individuals are exposed to more diverse views on social media than in offline settings (Bakshy et al., 2015; Barnidge, 2017; Fletcher and Nielsen, 2018), cross-ideological interactions are frequent even in relation to highly contentious topics (Barberá et al., 2015b), and the increase in polarization in the US has been largest among those who are least likely to be active on social media (Boxell et al., 2017).

Finally, perhaps the most timely topic that has received the broadest media attention is the extent to which social media contributes to the spread of misinformation. Contrary to the conventional wisdom that exposure to false news during the 2016 presidential election was nearly universal, a growing consensus is emerging in the literature that points to important asymmetries in the extent to which this type of information is shared and consumed. Conservative, older citizens are more likely to consume misinformation on social media (Guess et al., 2019) and exposure appears to be concentrated on a small minority of users. Whether or not being exposed to false news can actually affect attitudes and behavior is unclear, and evidence is mixed (Jamieson, 2018; Allcott and Gentzkow, 2017).

### **CASE STUDY: ANALYZING THE DYNAMICS OF PROTEST MOVEMENTS USING TWITTER DATA**

In this last section, we demonstrate the ability of social media as a data generator through a

case study of activists and government accounts from Egypt and Bahrain in early 2011, during the Arab Spring. An analysis of data collected from Twitter explores the question of whether activists can activate offline protest through their online activity on social media.

The phrase ‘Arab Spring’ refers to the large-scale protests that occurred throughout the Middle East and North Africa from December 2010 through the end of 2011. On December 17, 2010, Mohamed Bouazizi self-immolated to protest the seizure of his fruit cart. His action inspired local protests that became national, and the national protests spread to other countries once President Ben Ali fled to Saudi Arabia on January 14, 2011. Hosni Mubarak would abdicate on February 14, and almost every country in the region experienced some form of protest.

Though there is much evidence that the protests were spontaneous (Wilson et al., 2012; Steinert-Threlkeld, 2017b), activists played important roles at various points. Before the protests, they held workshops about peaceful resistance, showed movies about protest or that were critical of a regime, and taught participants how to create political graffiti. These activities demonstrated to others that there exists discontent with a regime, and did so in such a way that others knew that others knew there is discontent. Generating political coordination has the same effect as advertising: the more people know about a product (the protest), the more likely are people to buy it (join the protest) (Chwe, 1998). For example, Egyptian activists deliberately discussed (advertised) the January 25 protests in taxi cabs because they knew the drivers would spread the information to their passengers (Lim, 2012). However, there is still little systematic research on the extent to which activists contributed to the spread of protest – that is the research question we try to address here.

The Arab Spring presents an ideal case to demonstrate the strengths of social media

data. Because the involved countries were – and largely still are – repressive, identifying networks of activists, gaining their trust and administering surveys was (is) a resource-intensive process. For example, even after Hosni Mubarak abdicated, individuals in Tahrir Square were afraid to respond to Western researchers' surveys (Tufekci and Wilson, 2012). Any work coming from that process would be difficult to compare to other situations because the researcher would have invested so much time in gathering data for a particular set of actors in one country. Even within a particular country, it would be difficult to study many movements at once.

Once a researcher has established him or herself in a repressive country, administering surveys presents additional challenges. Respondents are often suspicious of foreigners conducting research or do not want to be seen associating with them. Local enumerators may be hired, adding time and expense to the data gathering process. If an unforeseen event occurs, such as a protest, the researcher is unlikely to be able to pivot to study it. Even if the researcher can study this unexpected event, she or he will be hard pressed to gather many observations per day, not to mention many observations across many locales per day. Survey work is hard, especially in places where it is infrequently conducted.

For the reasons provided earlier in this chapter, social media data ameliorates many of these concerns. To demonstrate how they do so, we analyze 19 activists across four social movements in Egypt and Bahrain, and examine whether their social media behavior led to an increase in offline protest.

In Egypt, we focus on the April 6 movement, which started in early 2008 as a Facebook page rallying support for striking textile workers at a government enterprise in Mahalla al-Kubra, a city of 535,000 inhabitants located 70 miles north of Cairo. Large-scale strikes and protests focused on working conditions and pay had occurred since 2006, sparking a periodic series of worker actions throughout Egypt over the next two years

(Beinin, 2009). The most important event was a large strike that was called for April 6, 2008, which the government reacted to by preemptively arresting activists and closing off public spaces nationwide (Gunning and Baron, 2013: 59–61). The movement persisted at a subdued level of activity – not for lack of trying – for the next three years and would become a central actor in the 2011 mobilization.

The second social movement in Egypt is the anti-sexual harassment movement. Egyptian public spaces have long been dangerous for women (Amar, 2011). As protests increased in Egypt throughout the first decade of the new millennium, so did reports of sexual assault at these events; in many cases, these assaults were linked to civilians hired by the Interior Ministry for that purpose (Langohr, 2013). In response to these events, an assortment of civil society organizations emerged, most notably the Nadeem Center, the Egyptian Center for Women's Rights and the Nazra for Feminist Studies.

We also study the two main organizations in support of human rights in Bahrain, the Bahrain Center for Human Rights (BCHR) and Bahrain Human Rights Society (BHRS), which were founded in 2002 and were still active at the time of the 2011 protests.

To study these movements, we rely on data that contains a complete history of the 19 most important activists within these movements on Twitter. The data was purchased from Sifter, a third-party reseller, and spans the period from January 11, 2011 through April 5, 2011. This time period was selected because it encompasses the time leading to each country's main protest period, the time during the main protest period (January 25–February 11 in Egypt; February 14–March 17 in Bahrain) and the time after the protests.<sup>9</sup> Sifter returned 58,376 tweets; each includes metadata on the number of followers of the account, the number of people the account follows and a character string describing the device from which the tweet was created. Table 23.1 details these data.<sup>10</sup>

**Table 23.1 Descriptive statistics from Sifter data**

Account	Followers	Friends	Tweets	Twitter.com	HTTps	iPhone	Android	BlackBerry	Windows	Nokia	Retweets	Mention	Hashtag	Group
Shabab6april	4229	78	833	0.67	0.00	0.00	0.00	0.00	0.00	0.00	0.09	0.16	0.87	April 6
mrmeit	2377	244	16262	0.84	0.00	0.00	0.00	0.00	0.00	0.00	0.16	0.59	0.32	April 6
AsmaaMahfouz	1418	88	201	0.98	0.00	0.00	0.00	0.00	0.00	0.00	0.07	0.57	0.16	April 6
waleedrashed	694	329	3265	0.77	0.00	0.00	0.00	0.22	0.00	0.00	0.36	0.76	0.24	April 6
Seldemerdash	1616	277	11832	0.14	0.00	0.67	0.00	0.00	0.00	0.00	0.19	0.54	0.22	Anti-SH
Anti-SHmap	726	154	234	0.51	0.00	0.11	0.00	0.00	0.00	0.00	0.32	0.43	0.38	Anti-SH
SorayaBahgat	315	267	344	0.11	0.00	0.00	0.00	0.89	0.00	0.00	0.15	0.66	0.42	Anti-SH
MariamKirollos	127	37	564	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.52	0.66	0.60	Anti-SH
Ribeska	15	53	23	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.42	0.33	0.83	Anti-SH
ZeinabSabet	2	7	5	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	Anti-SH
alaa	13109	371	38755	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.28	0.92	0.29	NoMilTrials
Monasosh	8595	251	9479	0.47	0.00	0.00	0.00	0.53	0.00	0.00	0.18	0.68	0.40	NoMilTrials
NABEELRAJAB	7285	433	2985	0.80	0.00	0.16	0.00	0.00	0.00	0.00	0.10	0.22	0.84	Hum. Rights
BahrainRights	6687	574	4641	0.00	0.00	0.63	0.00	0.00	0.00	0.00	0.28	0.68	0.92	Hum. Rights
MARYAMALKHAWAJA	4790	187	1112	0.45	0.00	0.30	0.00	0.25	0.00	0.00	0.43	0.54	0.76	Hum. Rights
angrarakbiya	3274	144	2575	0.80	0.00	0.00	0.00	0.13	0.00	0.00	0.10	0.62	0.28	Hum. Rights
SAIDYOUSIF	765	56	1574	0.49	0.00	0.51	0.00	0.00	0.00	0.00	0.34	0.47	0.84	Hum. Rights

Note: Accounts are sorted alphabetically within each movement. Columns from Twitter.com through Nokia refer to the percent of tweets from each source; for example, 84% of @mrmeit's tweets are through Twitter's website. Columns Retweets through Hashtag refer to the percentage of each account's tweets which are a retweet, contain a user mention or contain at least one hashtag.

Did the activities on Twitter by these key leaders have any effect on subsequent protest events in these two countries? Contrary to the conventional wisdom that highly visible activists were a determinant in the spread of protest across these two countries, here we show that there is little evidence supporting that assertion. Table 23.2 presents regression where we test that hypothesis through a multivariate regression on the number of protests with a range of variables related to the social media activity of leaders before those protests took place. The data on protests was extracted from the Integrated Conflict Early

Warning System (ICEWS) (Boschee et al., 2015).<sup>11</sup> To measure activists' social media behavior, we use a set of different variables that include the percentage of tweets from activists, the percentage of tweets that are activist retweets, the percentage of tweets that are activists mentioning others, the percentage of tweets with hashtags that are from activists and the percentage of tweets with links from activists.<sup>12</sup>

Our first model in Table 23.2 shows that none of the variables related with activists' behavior on Twitter predicts subsequent protest. And in fact, *Activist Coordination Tweet*  $\%_{i,t-1}$ , the

**Table 23.2 Main results**

	DV: $Protest_{i,t}$	
	(1)	(2)
Activist Tweet $\%_{i,t-1}$	2.58 (6.98)	1.72 (7.47)
Activist Retweet $\%_{i,t-1}$	5.65 (5.35)	6.26 (5.54)
Activist Mention $\%_{i,t-1}$	-2.23 (7.32)	-1.15 (8.08)
Activist Hashtag $\%_{i,t-1}$	-1.22 (1.13)	-1.27 (1.14)
Activist Link $\%_{i,t-1}$	9.75 (11.80)	11.19 (12.18)
Protest $_{i,t-1}$	0.01 (0.01)	0.01 (0.01)
Repression $_{i,t-1}$	0.01 (0.02)	0.01 (0.02)
Non-Activist Coordination $_{i,t-1}$	5.31*** (0.85)	5.33*** (0.85)
Activist Coordination Tweet $\%_{i,t-1}$	-253.02** (116.86)	-218.72* (132.53)
Activist Coordination $_{i,t-1}$		-181.11 (393.02)
Intercept	-2.06*** (0.49)	-2.08*** (0.50)
Country FE	Yes	Yes
N	168	168
Log Likelihood	-473.42	-473.33

\*p < .1; \*\*p < .05; \*\*\*p < .01

Note: All activist variables except for Activist Hashtag  $\%_{i,t-1}$  use the total tweets from country, as the denominator. Activist Hashtag  $\%_{i,t-1}$  uses the total tweets with hashtags from country.



percentage of tweets coordinating protests that are from activists, negatively correlates with protest.<sup>13</sup> The second model in Table 23.2 uses a different measure of activist coordination but is otherwise identical to the first. Here, *Activist Coordination*<sub>*i,t-1*</sub> is the interaction of *Non-Activist Coordination*<sub>*i,t-1*</sub> and *Activist Hashtag* %<sub>*i,t-1*</sub>. *Non-Activist Coordination*<sub>*i,t-1*</sub> is the Gini coefficient for hashtags per country-day; if a large percentage of those hashtags are produced by activists, then hashtag coordination comes from activists. This measure is not statistically significant, while *Activist Coordination Tweet*

%<sub>*i,t-1*</sub> remains so. Note that in both models *Non-Activist Coordination*<sub>*i,t-1*</sub> is significant with a p-value less than .01.

Activists in both countries frequently use mobile devices for communication, but they also used desktop computers. One may argue that our results in the previous set of regression models are null because they do not focus specifically on calls for actions from whenever the activists are actually in the streets engaging in political protest. To address this concern, we take advantage of a key piece of metadata that comes with tweets, the ‘tweet\_source’ field. This field is

**Table 23.3 Robustness checks – tweets from phones**

	Protest	
	(1)	(2)
Activist Tweet % <sub><i>i,t-1</i></sub>	-6.28 (8.06)	-6.14 (8.05)
Activist Retweet % <sub><i>i,t-1</i></sub>	-1.36 (6.85)	-1.91 (6.90)
Activist Mention % <sub><i>i,t-1</i></sub>	8.40 (8.99)	8.02 (9.00)
Activist Hashtag % <sub><i>i,t-1</i></sub>	-0.84 (1.11)	-0.50 (1.10)
Activist Link % <sub><i>i,t-1</i></sub>	10.90 (11.71)	7.93 (11.73)
Protest <sub><i>i,t-1</i></sub>	0.01* (0.01)	0.01 (0.01)
Repression <sub><i>i,t-1</i></sub>	0.01 (0.02)	0.01 (0.02)
Non-Activist Coordination <sub><i>i,t-1</i></sub>	5.44*** (0.84)	5.60*** (0.85)
Activist Coordination Tweet % <sub><i>i,t-1</i></sub>	-271.44** (116.12)	-437.98*** (168.53)
Mobile Phone % <sub><i>i,t-1</i></sub>	6.04* (3.52)	5.44 (3.52)
Activist Coordination Tweet % <sub><i>i,t-1</i></sub> * Mobile Phone % <sub><i>i,t-1</i></sub>		4,767.43 (4,194.30)
Intercept	-2.09*** (0.49)	-2.17*** (0.49)
Country FE	Yes	Yes
N	168	168
Log Likelihood	-472.11	-471.26

\*p <.1; \*\*p <.05; \*\*\*p <.01

a string created by Twitter to reflect the provenance of a tweet. We find that our results do not change when we analyze specifically the proportion of tweets originated on a mobile device. Although the percentage of an activist's tweets that are from a mobile phone does positively correlate with subsequent protest and has a p-value between .05 and .10, the result does not hold when that value is interacted with the percentage of tweets that are about coordination. This result makes sense, as a tweet does not say whether or not it comes from a mobile device, so a tweet from a mobile device does not provide a signal to others that the author has mobilized.

## CONCLUSION

The aim of this chapter was to offer a broad overview of how social media data is currently being used in social science research, as well as a detailed account of the main strengths and weaknesses of this source of information about political and social behavior. Overall, our assessment shows the immense and still largely untapped potential of social media as data generators. This means we are still at an early stage in the development of standards and best practices in the different literatures that increasingly rely on this type of data, but also that we should expect to see the results of much exciting new research being published over the next few years.

One particular challenge that we didn't discuss at length is the extent to which research that uses social media data can yield findings that can be generalizable across domains and over time, as opposed to idiosyncratic to the particular site whose data is being used. It is certainly the case that some of the websites we study did not even exist 10 years ago, which begs the question of whether they will still exist 10 years from now. We have plenty of examples of successful sites that eventually disappeared, such as MySpace or

Friendster. However, our view is that even if Facebook or Twitter eventually disappear, findings derived from research on these sites will survive and remain valid. Contrary to the common view that characterizes social media interactions as not occurring in the 'real world', behavior on these sites indeed mirrors offline behavior, and thus it can help reveal the mechanisms that drive human behavior, not only on these platforms, but in people's lives more generally.

## Notes

- 1 Of course, the fact that behavior on social media takes place in public may introduce other types of social desirability bias.
- 2 Observing a group over time obviates the numerator problem.
- 3 'On observables' is, of course, a major qualification.
- 4 Keyword self-selection can be mitigated by connecting to the streaming API and downloading a 1% sample of tweets in real time. A downside of this approach is that events or keywords that are not very popular are less likely to appear in the stream since it is a sample.
- 5 Morstatter et al. (2013) show divergence in trends in (1) the number of tweets, (2) topics, and (3) some network features in tweets related to Syria from 12.14.2011 to 01.10.2012. Overall, it is clear that a random sample collected from streaming API may not be a perfectly representative sample of all of Twitter.
- 6 Our discussion here focuses on Twitter, as the main social media platform offering easy access to public data for research purposes. Although similar data used to be available for Instagram and Facebook, recent changes to their platform policies mean that their APIs are essentially no longer available for research purposes (Freelon, 2018).
- 7 An alternative would be to develop a framework that can produce synthetic data with similar properties as the full dataset while ensuring privacy (Raab et al., 2016), but to our knowledge this approach has not been developed for social media datasets yet.
- 8 If the ID numbers are distributed for non-academic purposes, no more than 1.5 million per 30 day period can be shared.
- 9 The end of the main protest period in Egypt is defined as Mubarak's resignation. In Bahrain, protests ended following a major assault on the Pearl Roundabout, the main protest site in Manama;

this assault occurred three days after Gulf Cooperation Council forces, led by Saudi Arabia, marched into Bahrain, and the Pearl Roundabout was dismantled on March 18. Protesters would not again succeed in occupying the circle.

- 10 For additional information about data formats and code to collect and analyze Twitter data, see the materials available in Steinert-Threlkeld (2018).
- 11 ICEWS is a Department of Defense project, led by Lockheed Martin and Michael Ward, that reads newspapers and extracts events. It represents a substantial modification and extension of Philip Schrodt's Kansas Events Data System (KEDS) and Textual Analysis by Augmented Replacement Instructions (TABARI) (Schrodt et al., 1994; Gerner et al., 2002). ICEWS reads thousands of news sources, including non-English ones, and applies a heavily modified version of TABARI, leading to much lower rates of false positives than other machine-coded events data.
- 12 Note that *Activist Hashtag*  $\%_{i,t-1}$  is calculated slightly differently from *Activist Tweet*  $\%_{i,t-1}$ , *Activist Retweet*  $\%_{i,t-1}$ , *Activist Mention*  $\%_{i,t-1}$ , and *Activist Link*  $\%_{i,t-1}$ . *Activist Hashtag*  $\%_{i,t-1}$  is calculated as the percentage of all tweets with hashtags that are tweets from activists, but the other four take the total number of tweets from the activists' country on that day as the denominator. The variables are modeled differently to reflect the information consumption process on Twitter. When one sees a tweet on Twitter, it is presented as part of a sequence of reverse chronological tweets. If one views tweets containing a hashtag, however, all tweets in the subsequent reverse chronological sequence contain that hashtag. The determinant of the length of the latter sequence is therefore all tweets containing that hashtag while the length of all tweets one sees is better approximated by all tweets on that day.
- 13 These coordination tweets are those determined by a supervised topic model. For more details, see Steinert-Threlkeld (2017b).

## REFERENCES

- Allcott, Hunt and Matthew Gentzkow. 2017. 'Social media and fake news in the 2016 election.' *Journal of Economic Perspectives* 31(2):211–236.
- Alsaedi, Nasser, Pete Burnap and Omer Rana. 2017. 'Can we predict a riot? Disruptive event detection using Twitter.' *ACM Transactions on Internet Technology* 17(2):1–26.
- Amar, Paul. 2011. 'Turning the gendered politics of the security state inside out? Charging the police with sexual harassment in Egypt.' *International Feminist Journal of Politics* 13(3):299–328.
- Anastasopoulos, Lefteris and Jake Williams. 2017. 'Understanding collective action with Bayesian social action identification algorithms.' Unpublished manuscript, URL: <https://scholar.harvard.edu/files/janastas/files/computation-social-action-advances.pdf>
- Bakshy, Eytan, Solomon Messing and Lada A. Adamic. 2015. 'Exposure to ideologically diverse news and opinion on Facebook.' *Science* 348(6239):1130–1132.
- Barbera, Pablo. 2015. 'Birds of the same feather Tweet together: Bayesian ideal point estimation using Twitter data.' *Political Analysis* 23(1): 76–91.
- Barberá, Pablo and Gonzalo Rivero. 2015. 'Understanding the political representativeness of Twitter users.' *Social Science Computer Review* 33(6):712–729.
- Barberá, Pablo and Thomas Zeitzoff. 2017. 'The new public address system: why do world leaders adopt social media?' *International Studies Quarterly* 62(1):121–130.
- Barberá, Pablo, John T. Jost, Jonathan Nagler, Joshua A. Tucker and Richard Bonneau. 2015b. 'Tweeting from left to right: is online political communication more than an echo chamber?' *Psychological Science* 26(10):1531–1542.
- Barberá, Pablo, Andreu Casas, Jonathan Nagler, Patrick Egan, Richard Bonneau, John T. Jost, and Joshua Tucker. 2019. 'Who leads? Who follows? Measuring issue attention and agenda setting by legislators and the mass public using social media.' *American Political Science Review* 113(4): 883–901.
- Barberá, Pablo, Ning Wang, Richard Bonneau, John T. Jost, Jonathan Nagler, Joshua Tucker and Sandra González-Bailón. 2015a. 'The critical periphery in the growth of social protests.' *PLOS One* 10(11):e0143611.
- Barnidge, Matthew. 2017. 'Exposure to political disagreement in social media versus face-to-face and anonymous online settings.' *Political Communication* 34(2):302–321.
- Beauchamp, Nicholas. 2017. 'Predicting and interpolating state-level polls using Twitter textual data.' *American Journal of Political Science* 61(2):490–503.

- Beinin, Joel. 2009. Workers' struggles under 'socialism' and neoliberalism. In *Egypt: The Moment of Change*, ed. Philip Marfleet and Rabab El-Mahdi. London: Zed Books pp. 68–86.
- Bisbee, James and Jennifer M. Larson. 2017. 'Testing social science network theories with online network data: an evaluation of external validity.' *American Political Science Review* 111(3):502–521.
- Bliss, Catherine A., Isabel M. Kloumann, Kameron Decker Harris, Christopher M. Danforth and Peter Sheridan Dodds. 2012. 'Twitter reciprocal reply networks exhibit assortativity with respect to happiness.' *Journal of Computational Science* 3(5):388–397.
- Bond, Robert and Solomon Messing. 2015. 'Quantifying social media's political space: estimating ideology from publicly revealed preferences on Facebook.' *American Political Science Review* 109(1):62–78.
- Bond, Robert M., Christopher J. Fariss, Jason J. Jones, Adam D. I. Kramer, Cameron Marlow, Jaime E. Settle and James H. Fowler. 2012. 'A 61-million-person experiment in social influence and political mobilization.' *Nature* 489(7415):295–298.
- Borge-Holthoefer, Javier, Walid Magdy, Kareem Darwish and Ingmar Weber. 2015. Content and network dynamics behind Egyptian political polarization on Twitter. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work and Social Computing*, Vancouver, BC, Canada. pp. 700–711.
- Boschee, Elizabeth, Jennifer Lautenschlager, Sean O'Brien, Steve Shellman, James Starz and Michael Ward. 2015. 'ICEWS coded event data.' URL: <http://dx.doi.org/10.7910/DVN/28075>
- Boxell, Levi, Matthew Gentzkow and Jesse M. Shapiro. 2017. 'Greater internet use is not associated with faster growth in political polarization among US demographic groups.' *Proceedings of the National Academy of Sciences* p. 201706588.
- Breuer, Anita. 2012. The role of social media in mobilizing political protest: evidence from the Tunisian Revolution. Technical report, German Development Institute Bonn.
- Bright, Jonathan. 2018. 'Explaining the emergence of political fragmentation on social media: the role of ideology and extremism.' *Journal of Computer-Mediated Communication* 23(1):17–33.
- Broockman, David E. and Donald P. Green. 2014. 'Do online advertisements increase political candidates name recognition or favorability? Evidence from randomized field experiments.' *Political Behavior* 36(2):263–289.
- Brym, Robert, Melissa Godbout, Andreas Hoffbauer, Gabe Menard and Tony Huiquan Zhang. 2014. 'Social media in the 2011 Egyptian uprising.' *The British Journal of Sociology* 65(2):266–292.
- Cadena, Jose, Gizem Korkmaz, Chris J. Kuhlman, Achla Marathe, Naren Ramakrishnan and Anil Vullikanti. 2015. 'Forecasting social unrest using activity cascades.' *PLOS One* 10(6):e0128879.
- Ceron, Andrea, Luigi Curini, Stefano M. Iacus and Giuseppe Porro. 2014. 'Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens political preferences with an application to Italy and France.' *New Media & Society* 16(2):340–358.
- Chen, Ted Hsuan Yun, Chen, Paul Zachary and Christopher J. Fariss. 2017. 'Who protests? Using social media data to estimate how social context affects political behavior.' Unpublished manuscript. URL: [https://ssl.uh.edu/hobby/\\_docs/events/FarissWhoProtestSocialMediaData.pdf](https://ssl.uh.edu/hobby/_docs/events/FarissWhoProtestSocialMediaData.pdf)
- Chwe, Michael. 1998. 'Believe the hype: solving coordination problems with television advertising.' URL: <http://www.chwe.net/michael/h.pdf>
- Coakley, Kevin, Christine Kirkpatrick and Zachary C. Steinert-Threlkeld. 2019. 'rehydratoR.' URL: <https://cran.r-project.org/web/packages/rehydratoR/index.html>
- Coppock, Alexander, Andrew Guess and John Ternovski. 2016. 'When treatments are tweets: a network mobilization experiment over Twitter.' *Political Behavior* 38(1):105–128.
- Driscoll, Jesse and Zachary C. Steinert-Threlkeld. 2020. 'Social media and Russian territorial irredentism: some facts and a conjecture?' *Post-Soviet Affairs*. Forthcoming.
- Dunbar, R. I. M., Valerio Arnaboldi, Marco Conti and Andrea Passarella. 2015. 'The structure of online social networks mirrors those in the offline world.' *Social Networks* 43:39–47.

- Dwork, Cynthia. 2008. Differential privacy: a survey of results. In *International Conference on Theory and Applications of Models of Computation*. Springer pp. 1–19.
- Eckles, Dean, Brett R. Gordon and Garrett A. Johnson. 2018. 'Field studies of psychologically targeted ads face threats to internal validity.' *Proceedings of the National Academy of Sciences* p. 201805363.
- Ferrara, Emilio, Onur Varol, Clayton Davis, Filippo Menczer and Alessandro Flammini. 2016. BotOrNot: a system to evaluate social bots. In *Proceedings of the 25th International Conference Companion on World Wide Web, Montréal, Québec, Canada*, pp. 273–274.
- Fletcher, Richard and Rasmus Kleis Nielsen. 2018. 'Are people incidentally exposed to news on social media? A comparative analysis.' *New Media & Society* 20(7):2450–2468.
- Freelon, Deen. 2018. 'Computational research in the post-API age.' *Political Communication* 35(4): 665–668.
- Freelon, Deen, Marc Lynch and Sean Aday. 2015. 'Online fragmentation in wartime: a longitudinal analysis of Tweets about Syria, 2011–2013.' *The ANNALS of the American Academy of Political and Social Science* 659(1):166–179.
- Freelon, Deen, Charlton Mcllwain and Meredith Clark. 2016. *Beyond the hashtags: #Ferguson, #Blacklivesmatter, and the online struggle for offline justice*. Center for Media and Social Impact, School of Communication, American University.
- Gayo Avello, Daniel, Panagiotis T. Metaxas and Eni Mustafaraj. 2011. Limits of electoral predictions using twitter. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, Barcelona, Spain, pp. 490–493.
- Gerner, Deborah J., Philip A. Schrodt, Rajaa Abu-Jabr and Omur Yilmaz. 2002. Conflict and Mediation Event Observations (CAMEO): a new event data framework for the analysis of foreign policy interactions. In *Annual Meeting of the International Studies Association*.
- Ghannam, Jeffrey. 2011. Social media in the Arab world: leading up to the uprisings of 2011. Technical report, Center for International Media Assistance Washington D.C.
- Gladwell, Malcolm. 2010. 'Small change: why the revolution will not be tweeted.' *The New Yorker*. September 27, 2010.
- Golder, Scott A. and Michael W. Macy. 2011. 'Diurnal and seasonal mood vary with work, sleep, and daylight across diverse cultures.' *Science* 333(6051):1878–1881.
- Golder, Scott A. and Michael W. Macy. 2014. 'Digital footprints: opportunities and challenges for online social research.' *Annual Review of Sociology* 40:129–152.
- Gosling, Samuel D., Adam A. Augustine, Simine Vazire, Nicholas Holtzman and Sam Gaddis. 2011. 'Manifestations of personality in online social networks: self-reported Facebook-related behaviors and observable profile information.' *Cyberpsychology, Behavior and Social Networking* 14(9):483–488.
- Grimmer, Justin, Solomon Messing and Sean J. Westwood. 2012. 'How words and money cultivate a personal vote: the effect of legislator credit claiming on constituent credit allocation.' *American Political Science Review* 106(4):703–719.
- Gruzd, Anatoliy and Ksenia Tsyganova. 2015. 'Information wars and online activism during the 2013/2014 crisis in Ukraine: examining the social structures of pro- and anti-Maidan groups.' *Policy & Internet* 7(2):121–158. URL: <http://doi.wiley.com/10.1002/poi3.91>
- Guess, Andrew, Jonathan Nagler and Joshua Tucker. 2019. 'Less than you think: prevalence and predictors of fake news dissemination on Facebook.' *Science Advances* 5(1):eaau4586.
- Gunning, Jeroen and Ilan Zvi Baron. 2013. *Why Occupy a Square: People, Protests and Movements in the Egyptian Revolution*. C. Hurst & Co. Publishers.
- Hobbs, William R. and Margaret E. Roberts. 2018. 'How sudden censorship can increase access to information.' *American Political Science Review* 112(3): 621–636.
- Huckfeldt, Robert, Paul E. Johnson and John Sprague. 2004. *Political Disagreement: The Survival of Diverse Opinions within Communication Networks*. Cambridge University Press.
- Jamieson, Kathleen Hall. 2018. *Cyberwar: How Russian Hackers and Trolls Helped Elect a President: What We Don't, Can't, and Do Know*. Oxford University Press.

- Jones, Jason J., Jaime E. Settle, Robert M. Bond, Christopher J. Fariss, Cameron Marlow and James H. Fowler. 2013. 'Inferring tie strength from online directed behavior.' *PLOS One* 8(1):e52168.
- Kalla, Joshua L. and David E. Broockman. 2018. 'The minimal persuasive effects of campaign contact in general elections: evidence from 49 field experiments.' *American Political Science Review* 112(1):148–166.
- Keller, Franziska B., David Schoch, Sebastian Stier and JungHwan Yang. 2017. How to manipulate social media: analyzing political astroturfing using ground truth data from South Korea. In *ICWSM*. pp. 564–567.
- King, Gary, Jennifer Pan and Margaret E. Roberts. 2013. 'How censorship in China allows government criticism but silences collective expression.' *American Political Science Review* 107(2):326–343.
- King, Gary, Jennifer Pan and Margaret E. Roberts. 2014. 'Reverse-engineering censorship in China: randomized experimentation and participant observation.' *Science* 345 (6199):1251722.
- King, Gary, Jennifer Pan and Margaret E. Roberts. 2017. 'How the Chinese government fabricates social media posts for strategic distraction, not engaged argument.' *American Political Science Review* 111(3):484–501.
- Klašnja, Marko, Pablo Barberá, Nick Beauchamp, Jonathan Nagler and Joshua Tucker. 2016. Measuring public opinion with social media data. In *The Oxford Handbook of Polling and Survey Methods*.
- Kosinski, Michal, David Stillwell and Thore Graepel. 2013a. 'Private traits and attributes are predictable from digital records of human behavior.' *Proceedings of the National Academy of Sciences of the United States of America* 110(15):5802–5805.
- Kosinski, Michal, David Stillwell and Thore Graepel. 2013b. 'Private traits and attributes are predictable from digital records of human behavior.' *Proceedings of the National Academy of Sciences* p. 201218772.
- Kostyuk, Nadiya and Yuri M. Zhukov. 2017. 'Invisible digital front: can cyber attacks shape battlefield events?' *Journal of Conflict Resolution*: 1–31. doi:10.1177/0022002717737138
- Kramer, Adam D. I., Jamie E. Guillory and Jeffrey T. Hancock. 2014. 'Experimental evidence of massive-scale emotional contagion through social networks.' *Proceedings of the National Academy of Sciences* p. 201320040.
- Kreiss, Daniel. 2012. *Taking Our Country Back: The Crafting of Networked Politics from Howard Dean to Barack Obama*. Oxford University Press.
- Kreiss, Daniel. 2016. 'Seizing the moment: the presidential campaign's use of Twitter during the 2012 electoral cycle.' *New Media & Society* 18(8):1473–1490.
- Langohr, Vickie. 2013. "'This is our square": fighting sexual assault at Cairo protests.' *Middle East Research and Information Project* 43(Fall). URL: <http://www.merip.org/mer/mer268/our-square>
- Lerman, Kristina, Rumi Ghosh and Tawan Surachawala. 2012. 'Social contagion: an empirical study of information spread on Digg and Twitter follower graphs.' *arXiv preprint arXiv:1202.3162*.
- Lim, Merlyna. 2012. 'Clicks, cabs, and coffee houses: social media and oppositional movements in Egypt, 2004–2011.' *Journal of Communication* 62(2):231–248.
- Little, Andrew T. 2015. 'Communication technology and protest.' *Journal of Politics* 78(1):152–166.
- Malik, Momin M., Hemank Lamba, Constantine Nakos and Jurgen Pfeffer. 2015. Population bias in geotagged Tweets. In *9th International AAAI Conference on Weblogs and Social Media*. pp. 18–27. AAAI Press.
- Moreno, Jacob Levy. 1934. *Who Shall Survive? A New Approach to the Problem of Human Interrelations*. Nervous and Mental Disease Publishing Co., Washington DC.
- Morozov, Evgeny. 2012. *The Net Delusion: The Dark Side of Internet Freedom*. PublicAffairs.
- Morstatter, Fred, Jürgen Pfeffer, Huan Liu and Kathleen M. Carley. 2013. Is the sample good enough? Comparing data from Twitter's streaming API with Twitter's Firehose. In *7th ICWSM, Cambridge, MA*. AAAI Press.
- Munger, Kevin. 2017. 'Tweetment effects on the Tweeted: experimentally reducing racist harassment.' *Political Behavior* (39)3: 629–649.
- Munger, Kevin, Richard Bonneau, Jonathan Nagler and Joshua A. Tucker. 2018. 'Elites Tweet to get feet off the streets: measuring

- regime response to protest using social media.' *Political Science Research and Methods* First View: 1–20. doi:10.1017/psrm.2018.3
- Mutz, Diana C. 2002. 'Cross-cutting social networks: testing democratic theory in practice.' *American Political Science Review* 96(1):111–126.
- Nulty, Paul, Yannis Theocharis, Sebastian Adrian Popa, Olivier Parnet and Kenneth Benoit. 2016. 'Social media and political communication in the 2014 elections to the European Parliament.' *Electoral Studies* 44:429–444.
- O'Callaghan, Derek, Nico Prucha, Derek Greene, Maura Conway, Joe Carthy and Pdraig Cunningham. 2014. 'Online social media in the Syria conflict: encompassing the extremes and the in-betweens.'
- Pariser, Eli. 2011. *The Filter Bubble: What the Internet Is Hiding from You*. Penguin UK.
- Pew Research Center. 2017. Partisan conflict and congressional outreach. *Pew Research Center Report*.
- Pfeffer, Jürgen and Katja Mayer. 2018. 'Tampering with Twitter's sample API.' *EPJ Data Science* 7, Art. 50:1–21.
- Putnam, Robert D. 2000. Bowling alone: America's declining social capital. In *Culture and Politics*. Springer pp. 223–234.
- Raab, Gillian M., Beata Nowok and Chris Dibben. 2016. 'Practical data synthesis for large samples.' *Journal of Privacy and Confidentiality* 7(3):67–97.
- Radford, Jason and Betsy Sinclair. 2016. 'Electronic homestyle: Tweeting ideology.'
- Sakaki, Takeshi, Makoto Okazaki and Yutaka Matsuo. 2010. Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of the 19th International World Wide Web Conference*. ACM: 851–860.
- Schrodt, Philip A., Shannon G. Davis and Judith L. Weddle. 1994. 'Political science: KEDS – a program for the machine coding of event data.' *Social Science Computer Review* 12(4):561–587.
- Siegel, Alexandra A. and Vivienne Badaan. 2018. '#No2Sectarianism: experimental approaches to reducing sectarian hate speech online.'
- Sobolev, Anton, Keith Chen, Jungseock Joo and Zachary C. Steinert-Threlkeld 2019. 'News and geolocated social media accurately measure protest size.' Working paper available at [https://www.anderson.ucla.edu/faculty\\_pages/keith.chen/papers/WP\\_MeasuringProtestSize.pdf](https://www.anderson.ucla.edu/faculty_pages/keith.chen/papers/WP_MeasuringProtestSize.pdf) (Accessed on 19 February 2020).
- Steinert-Threlkeld, Zachary C. 2017a. 'Longitudinal network centrality using incomplete data.' *Political Analysis* 25(3): 308–328.
- Steinert-Threlkeld, Zachary C. 2017b. 'Spontaneous collective action: peripheral mobilization during the Arab Spring.' *American Political Science Review* 111(2):379–403.
- Steinert-Threlkeld, Zachary C. 2018. *Twitter as Data*. Cambridge University Press.
- Steinert-Threlkeld, Zachary C. 2019. 'Comment: the future of event data is images.' *Sociological Methodology* 49(1): 68–75.
- Steinert-Threlkeld, Zachary C., Jungseock Joo, and Alexander Chan. 2020. 'How violence affects protests.' *APSA Preprints*. doi: 10.33774/apsa-2019-bv6zd-v2.
- Stier, Sebastian, Lisa Posch, Arnim Bleier and Markus Strohmaier. 2017. 'When populists become popular: comparing Facebook use by the right-wing movement Pegida and German political parties.' *Information, Communication & Society* 20(9):1365–1388.
- Sunstein, Cass R. 2018. *#Republic: Divided Democracy in the Age of Social Media*. Princeton University Press.
- Takhteyev, Yuri, Anatoliy Gruzd and Barry Wellman. 2012. 'Geography of Twitter networks.' *Social Networks* 34(1):73–81.
- Theocharis, Yannis, Pablo Barberá, Zoltán Fazekas, Sebastian Adrian Popa and Olivier Parnet. 2016. 'A bad workman blames his tweets: the consequences of citizens' uncivil Twitter use when interacting with party candidates.' *Journal of Communication* 66(6):1007–1031.
- Timoneda, Joan. 2018. 'Where in the world is my tweet: detecting irregular removal patterns on Twitter.' *PLOS One* 13(9):e0203104.
- Tufekci, Zeynep and Christopher Wilson. 2012. 'Social media and the decision to participate in political protest: observations from Tahrir Square.' *Journal of Communication* 62(2):363–379.
- Tumasjan, Andranik, Timm Oliver Sprenger, Philipp G. Sandner and Isabell M. Welpe. 2010. Predicting elections with Twitter: what 140 characters reveal about political sentiment. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*. pp. 178–185.

- Ugander, Johan, Brian Karrer, Lars Backstrom and Cameron Marlow. 2011. 'The anatomy of the Facebook social graph.' *arXiv preprint arXiv:1111.4503*.
- Valenzuela, Sebastián. 2013. 'Unpacking the use of social media for protest behavior: the roles of information, opinion expression, and activism.' *American Behavioral Scientist* 57(7):920–942.
- Wilson, Robert E., Samuel D. Gosling and Lindsay T. Graham. 2012. 'A review of Facebook research in the social sciences.' *Perspectives on Psychological Science* 7(3):203–220.
- Wilson, Steven Lloyd. 2017. 'Detecting mass protest through social media.' *Journal of Social Media in Society* 6(2):5–25.
- Xi, Nan, Di Ma, Marcus Liou, Zachary C. Steinert-Threlkeld, Jason Anastasopoulos, and Jungseock Joo. 2020. Understanding the Political Ideology of Legislators from Social Media Images. In *Proceedings of the International AAAI Conference on Web and Social Media*. Forthcoming.
- Youyou, Wu, Michal Kosinski and David Stillwell. 2015. 'Computer-based personality judgments are more accurate than those made by humans.' *Proceedings of the National Academy of Sciences* 112(4):1036–1040.
- Zachary, Wayne W. 1977. 'An information flow model for conflict and fission in small groups.' *Journal of Anthropological Research* 33(4):452–473.
- Zamal, Faiyaz Al, Wendy Liu and Derek Ruths. 2012. Homophily and latent attribute inference: inferring latent attributes of Twitter users from neighbors. In *Sixth International AAAI Conference on Weblogs and Social Media*. pp. 387–390.
- Zeitsoff, Thomas. 2011. 'Using social media to measure conflict dynamics: an application to the 2008–2009 Gaza conflict.' *Journal of Conflict Resolution* 55(6):938–969.
- Zeitsoff, Thomas. 2018. 'Does social media influence conflict? Evidence from the 2012 Gaza conflict.' *Journal of Conflict Resolution* 62(1): 29–63.
- Zeitsoff, Thomas, John Kelly and Gilad Lotan. 2015. 'Using social media to measure foreign policy dynamics: an empirical analysis of the Iranian–Israeli confrontation (2012–13).' *Journal of Peace Research* 52(3):368–383.
- Zhang, Han and Jennifer Pan. 2019. 'CASM: a deep-learning approach for identifying collective action events with text and image data from social media.' *Sociological Methodology* 49:1–48.
- Zhukov, Yuri M. 2015. 'Trading hard hats for combat helmets: the economics of rebellion in eastern Ukraine.' *Journal of Comparative Economics* 44(1):1–15.
- Zimmer, Michael. 2010. "But the data is already public": on the ethics of research in Facebook.' *Ethics and Information Technology* 12(4):313–325.





# Spatial Data

David Darmofal and Christopher Eddy

Political scientists analyze spatial data. We may think of our units of analysis primarily as block groups, zip codes, states, countries or dyads, but all of these data are also spatial data. Indeed, all political science data are spatial data (Darmofal, 2015) since all political behaviors, processes and events take place at spatial locations. But because we think of our units primarily in non-spatial terms, we often miss the implications that decades of geographic research provides for our analyses.

Spatial data present both opportunities and challenges for political scientists. Areal units such as those above are often geocoded so that we know the precise spatial locations of their boundaries and their centroids. Such geocoding allows us to define their neighbors and conduct spatial analyses. This is important because our spatial units often exhibit spatial dependence – similarities or dissimilarities in measures of interest that are related to our units' spatial locations. As a result, we typically need to model our areal units using

spatial models such as spatial lag or spatial error models. Such spatial modeling can shed new insights that are not possible when we conceive of our data in non-spatial terms.

The spatial nature of our data should also lead us to think about our units themselves and what they are capturing. It should also lead us to think about how our choice of units affects our substantive results. Many of the units of analysis that we employ in political science are arbitrary areal units created for governmental administrative purposes rather than because they capture contexts of substantive interest for social and political phenomena. Block groups, census tracts and zip codes are examples of such administrative units. Even when our units do carry a substantive meaning, such as countries, they might not be the appropriate level of analysis – either substate or regional units might make more sense for a particular analysis. Either way, we have a levels of analysis question, a problem well known to political scientists (see, e.g., Singer, 1961; Ray, 2001).

But because political scientists don't primarily think of our data as spatial data, we don't necessarily think of the levels of analysis problem in spatial terms and, as a consequence, we miss the spatial insights that geographers have been considering for decades.

In short, we need to be sensitive to how our results depend upon our choice of spatial units and what these units are actually capturing as contextual units. These issues, reflected in two related concerns – the modifiable areal unit problem (MAUP) and the uncertain geographic context problem (UGCoP) – are the focus of our chapter. In highlighting these issues, we hope to encourage political scientists and geographers toward greater dialogue with each other.

The modifiable areal unit problem – the fact that our substantive results are dependent upon how we divide a spatial plane into spatial units – is well known outside of political science. Openshaw and Taylor's (1979) classic analysis of the subject in 'A Million or So Correlation Coefficients: Three Experiments on the Modifiable Areal Unit Problem' has been cited more than 1,400 times across a variety of disciplines, ranging from geography to demography to sociology.<sup>1</sup> Its influence in political science, however, has been much more limited. A Google Scholar analysis of publications in three of the leading general journals in the discipline (the *American Political Science Review*, the *American Journal of Political Science* and *The Journal of Politics*) finds very few references to this subject that is so central to research in other disciplines. In fact, this Google Scholar search found that *no articles* in the discipline's flagship journal, the *American Political Science Review*, have included the term 'modifiable areal unit problem'. The story is not much better in the other two journals. Only three *American Journal of Political Science* articles have included the term, all since 2007.<sup>2</sup> Only three *Journal of Politics* articles have used the term, all since 2012.<sup>3</sup>

Is the story any more positive when we move away from general journals and look

at the leading subfield journals? Three fields in the discipline that frequently employ areal units are public administration, comparative politics and international relations. A Google Scholar search finds zero uses of the term 'modifiable areal unit problem' in three leading public administration journals (*Public Administration Review*, the *Journal of Public Administration Research and Theory* and the *Journal of Policy Analysis and Management*). In comparative politics, there is only one article in *World Politics*<sup>4</sup> that uses the term, two articles in *Comparative Political Studies*<sup>5</sup> and zero articles in *Comparative Politics*. In international relations, there are zero uses of the term in articles in *International Organization* and *International Studies Quarterly*, and only two articles that use the term in the *Journal of Conflict Resolution*.<sup>6</sup>

The uncertain geographic context problem (UGCoP) is a more recent concern in the geographic literature, identified by Mei-Po Kwan in a highly influential 2012 article in the *Annals of the Association of American Geographers*, 'The Uncertain Geographic Context Problem'. Kwan defines UGCoP as

the problem that findings about the effects of area-based attributes could be affected by how contextual units or neighborhoods are geographically delineated and the extent to which these areal units deviate from the 'true causally relevant' geographic context (the precise spatial configuration of which is unknown in most studies to date; Diez-Roux and Mair, 2010: 134). (p. 959)

In other words, where MAUP recognizes that results can differ depending upon how the spatial plane is partitioned, UGCoP emphasizes the importance of identifying the theoretically appropriate context for understanding particular behaviors, processes and events. It is the substantive complement to the methodological MAUP concern.

Kwan's UGCoP article has been cited more than 440 times in the six years since its publication. However, UGCoP is absent from the political science journals discussed above. None of our leading general journals has published an article that includes the

term ‘uncertain geographic context problem’. None of the top subfield journals discussed above has published such an article either.

In short, two prominent concerns in geography and other disciplines – ones highly relevant for and related to longstanding concerns in political science – have not been utilized much within the discipline of political science. We believe that political scientists can benefit significantly from a greater familiarity with and use of these concepts. After highlighting how the research on contextual influences and racial attitudes can be strengthened by a more careful consideration of the underlying spatial and geographical context, we turn to discuss both UGCoP and MAUP in greater detail.

## CONTEXT IN POLITICAL RESEARCH

Scholars of political science have long been curious about the influence of contextual factors on political behaviors and how such factors may affect the formation of political, policy and racial attitudes. *Context* has been used to explain voting behavior, electoral participation and a growing variety of political and policy positions. According to Books and Prysby (1988: 215), the goals of contextual analysis are at least two-fold: ‘to uncover the extent of contextual effects and to unravel the mechanisms by which these effects occur’. While early research in this vein was typically narrow, focusing primarily on one specific relationship, researchers began focusing more explicitly on the role of context over time. Despite a steady supply of research seeking to understand context in relation to political behavior and public opinion, contextually based inquiries continue to face a variety of analytical and methodological challenges.

A prominent debate within the contextual theory literature apropos the role of context on racial attitudes underscores well the importance of recognizing both the theoretical concerns of UGCoP and the methodological

challenges highlighted by MAUP. By reviewing the evolution of literature surrounding a well-known contextually based theory, the racial threat hypothesis (Key, 1949), the following section illustrates how our ability to understand and accurately capture measures of context can be enhanced by a careful consideration of the geographic literature on context. Today, nearly all studies of racial context and individual behavior include at least a passing reference to Key’s *Southern Politics in State and Nation* (1949). In his book, Key offers a contextual hypothesis of perceived racial threat, suggesting that as the perceived size of a local minority population increases, so too will perceived feelings of racial threat among the predominant population. Ultimately, this is argued to increase levels of racial animosity and boost support for racially hostile policies among members of a predominant group. Indeed, studies lending support to the racial threat hypothesis are in no short supply.<sup>7</sup>

While Key focuses solely on the South, subsequent studies sought to extend the applicability of the racial threat hypothesis to northern urban centers (Katznelson, 1982), urban electoral politics (Browning and Marshall, 1986; Browning et al., 1984) and states more generally (Huckfeldt and Kohfeld, 1989). To be sure, the literature on the racial threat hypothesis has been far from conclusive. Some studies have concluded with null findings, while others have uncovered relationships between minority population size and majority attitudes that are the *opposite* of what Key’s theory would predict (Voss and Miller, 2001).

The ‘predominant explanation’ to these contradictory results, according to Newman (2013: 376), ‘is the countervailing predictions of intergroup threat and contact theory’. In contrast to the racial threat hypothesis, contact theory (Allport, 1954) assumes that intergroup interactions have the potential to positively influence political attitudes and behavior.<sup>8</sup> Whereas the racial threat hypothesis predicts a negative relationship between

rising minority populations and the racial attitudes of a predominant group in a given geographic unit, contact theory predicts larger out-group populations to be positively related to racial attitudes. The underlying theoretical assumption is that intergroup social interaction reduces or corrects negative stereotypes, such as those often portrayed in the mass media. Like the racial threat hypothesis, there is a healthy supply of research supporting the idea that racial contact, under certain conditions, can produce positive attitudes toward out-group members.<sup>9</sup>

Others have suggested that inconsistent conclusions from racial threat theory are more the result of analytical or methodological shortcomings. In a study extending the premise of racial threat theory to attitudes regarding immigration, Newman (2013: 376–7) notes: ‘research on power threat has produced one of the most central puzzles in the contextual research – the notoriously inconsistent findings for group-size based measures of ethnic context on citizens’ attitudes and policy preferences.’ ‘In response to the poor empirical performance of the power-threat hypothesis,’ Newman (2013: 377) writes, ‘the literature has seen the emergence of a new wave of contextual research possessing a sharper focus on stipulating the conditions under which group-size-based measures of racial and ethnic context lead to amity or enmity between groups (Branton and Jones, 2005; Oliver and Mendelberg, 2000; Oliver and Wong, 2003)’.

While earlier scholars such as Przeworski (1974) recognize measurement challenges in the use of aggregate contexts to examine contextual effects, little attention was given to the selection of theoretically meaningful geographic units. Because context can be measured at varying geographic levels, additional sources of variation have been commonly overlooked.<sup>10</sup> Indeed, many studies now find that substantive results depend on the level of measurement related to the contextual unit of analysis. Carsey (1995), for example, reports that the effect of black population densities

on white voter behavior is different at the county/state levels and the precinct/borough levels. Baybeck (2006) also finds different effects at the block-group and city levels. In other words, the commonly reported negative association between minority populations and white racial attitudes may diminish when examined from smaller geo-units. Further, Newman et al. (2015: 172) find a similar pattern when comparing zip codes and county measures, reporting that

citizens’ perceptions of their context are more responsive to their more immediate versus distal residential context. This result also reinforces the concern among the contextual research community that scholars should strive to use smaller geo-units to capture contextual effects, at least when such units correspond to theoretical process presumably operative at the neighborhood level, such as intergroup contact.

Oliver and Wong (2003) urge that ‘when comparing the impact of contact and conflict across racial environments, the racial composition of both the micro and macro contextual units need to be considered’ (p. 570). Again, despite the most consistent evidence supporting racial threat being captured at larger environmental units (counties and metro-regions), racial contact is more likely to be affected by smaller geographic units. Rocha and Espino (2009: 423–424) incorporate levels of residential segregation into the racial environment, ultimately uniting the racial threat and contact hypotheses. They find that segregation complicates the racial context, making ‘the conditions for either racial threat or social contract more likely’, and echo Oliver and Wong (2003) by warning that ‘future research on Anglo racial attitudes cannot rely on simple definitions of racial context, such as the size of the minority population, but also must take account of such influences as the spatial dispersion of minorities’. Adding to this line of thought, Baybeck (2006: 386) argues that ‘racial context needs to be considered as a complex system of overlapping spatial units’. Moreover, he examines how multiple

contexts interact, finding that 'there is no "one" context', and thus advises scholars to 'consider the overlapping political and social boundaries that surround individuals and that these contexts may interact in surprising yet systematic ways' (p. 395).

### MODIFIABLE AREAL UNIT PROBLEM

As stated earlier, many of the areal units that political scientists employ, such as block groups, census tracts and zip codes, are arbitrary and modifiable units developed for purposes of governmental administration and not for purposes of social science (Openshaw and Taylor, 1981: 60).<sup>11</sup> The arbitrary and modifiable nature of many areal objects is problematic, for as Openshaw and Taylor (1979) note, different areal objects can produce fundamentally different relationships between aggregate units or, as they colorfully note, 'a million or so correlation coefficients' (p. 127). This dependence of estimates on the areal objects of analysis is known as the modifiable areal unit problem (MAUP).

The modifiable areal unit problem comprises two problems. The *scale problem* refers to the dependence of findings on the number of areal units into which a spatial plane is divided. A given plane may be divided into any of an arbitrary number of polygons, with results differing fundamentally depending upon the  $n$  that is chosen. The *aggregation problem* refers to the dependence of results on the way that the spatial plane is divided into a particular set number of polygons. Thus, for example, there is a multitude of ways that a spatial plane can be divided into 50 polygons, and estimates of relationships between units will be dependent upon the researcher's choice of how this division is accomplished.

The arbitrary and modifiable nature of many areal objects, such as those discussed above, cannot be remedied; it is an inherent part of their nature. The fact that these units are not unique and fixed, however, would

not be problematic for applied researchers if neither the scale problem nor the aggregation problem obtained. That is, if spatial results did not vary as 500 small polygons were aggregated into 5 large polygons, or depending upon which 5 large polygons were chosen, then areal units would remain modifiable but this would not present a problem for estimation. Unfortunately, there is ample empirical evidence that the scale and aggregation problems are common and can be quite severe (see, e.g., Openshaw, 1983). As but one example, Openshaw and Taylor (1981) show that correlation coefficients for the 99 counties in Iowa varied from  $-0.97$  to  $+0.99$  depending upon how these counties were aggregated. Moreover, equally problematic, as Openshaw (1983: 5) notes, is that there is no general, systematic pattern to the effects on correlation coefficients that can aid with correction of these effects.

As Openshaw and Taylor (1981) show, a variety of solutions have been proposed for the modifiable areal unit problem, but many of these are problematic, as they impose additional, arbitrary criteria at the discretion of the researcher or ignore the fundamentally geographic nature of spatial data. For example, a spatial filtering approach due to Tobler (1969, 1975) may be employed to produce smoothed maps of underlying patterns which remove the noise resulting from aggregation effects. As Openshaw and Taylor note, however, this 'solution' involves an infinite regress since any filtering analysis based upon aggregate data will, by definition, be dependent upon the aggregate areal units employed for the analysis. Even more problematic from a model-based social science perspective is an inductive 'optimal zoning' approach in which the researcher specifies the spatial dependence she expects to observe based upon her understanding of the demographic, political, sociological or other phenomena she seeks to explain. The researcher then compares results using different definitions of the areal objects. The researcher then employs the polygons that are most consistent

with the type of spatial dependence between units (e.g., strongly positive, weakly negative) that she expects given the subject of her analysis (Openshaw and Taylor, 1981: 66–7). Clearly, such an inductive approach runs counter to a model-based social science perspective in which hypotheses are developed independent of the data used to test them.

Much preferable is a theoretically based choice of areal objects. However, as the UCGoP recognizes, such theory is often lacking. As Kwan notes, we can never be certain that we have employed the appropriate contextual units for our question.

The plausibility of this approach will depend upon the researcher's substantive question and the availability of appropriate areal objects. For some analyses, however, there is clear theoretical guidance on the proper polygons to employ in a spatial analysis. A critical question in international relations is the role that geography plays in conditioning the probability that countries will engage in militarized conflict with each other (see, e.g., Starr and Most, 1976, 1983; Most and Starr, 1980, 1982). All else equal, spatially proximate countries should be more likely than spatially distant countries to engage in militarized conflict with each other because proximity affords opportunities for conflict. Here, the choice of polygons – countries – is clear, and it would make little sense either to aggregate or disaggregate these polygons since it is countries that engage in international conflicts. Similarly, if our research question focuses on municipalities' tax policies, we would want to collect data at the municipality level. We might well expect negative spatial autocorrelation in municipality-level tax policies as suburban municipalities pursue low tax strategies to attract economic development away from center cities with higher tax rates. We would not want to collect data within municipalities as a substitute, as these data would exhibit positive spatial autocorrelation because units nested within the same municipality share tax policies decided at the municipal level of government.

For other questions, however, the match between the theoretical areal objects of interest and available areal objects will be less clear. The example of block groups, tracts and zip codes as proxies for neighborhoods is a prime example. Research on appropriate geographic definitions of neighborhoods remains imprecise. One recent area of research has focused on 'bespoke neighborhoods' in which increasing neighborhood domains are built outward from each household unit by including the  $n$  nearest persons to that household, where  $n$  is an arbitrary threshold value such as 50, 500 or 1,000 (see, e.g., MacAllister et al., 2001; Johnston et al., 2005). Although this bespoke neighborhood approach affords some increased verisimilitude by allowing neighborhood definitions to vary by household locations, the use of arbitrary thresholds is unlikely to distinguish precisely between actual neighborhoods.

In the case of neighborhood data, the modifiable areal unit problem will remain a concern until researchers make advances in the definition of neighborhoods *and* data begins to be collected for these areal units. In cases where areal objects remain defined for purposes of government administration rather than social scientific concerns, the researcher is best advised to be guided by theoretical considerations and to employ areal objects that most closely approximate her theoretical conceptions. In all spatial analyses for which the theoretical areal objects of interest are not readily available, scholars should employ multiple areal definitions and report the robustness of their results to these alternative polygon definitions in much the same way as they currently report the robustness of their results to alternative model specifications.

## **THE UNCERTAIN GEOGRAPHIC CONTEXT PROBLEM (UGCoP)**

It's worth repeating here that Kwan defines UGCoP as

the problem that findings about the effects of area-based attributes could be affected by how contextual units or neighborhoods are geographically delineated and the extent to which these areal units deviate from the 'true causally relevant' geographic context (the precise spatial configuration of which is unknown in most studies to date; Diez-Roux and Mair, 2010, 134). (p. 959)

We can, in other words, think of the two components of the modifiable areal unit problem as inevitable consequences of utilizing areal data, while the uncertain geographic context problem refers to the substantive and theoretical identification of the proper areal unit and its use.

Because many areal units were drawn for governmental administrative purposes, they are unlikely to reflect the true contexts that are relevant for behaviors. The actual neighborhood or community contexts that influence behavior are extremely unlikely to line up exactly with administrative boundaries drawn for a different purpose. Moreover, relevant contexts may differ for differing behaviors. Economic competition may occur across a metropolitan area while political competition may occur within a city or a town.

Conceptually distinct contexts such as metropolitan areas, cities and towns lend themselves naturally to theoretical arguments about contextual effects. The UGCoP may have a solution when scholars are working with such contextual units. Not so, however, with many of the areal units that are designed for administrative purposes, such as block groups, census tracts or zip codes. When scholars use these latter areal units it beggars belief to argue that any of these contextual units are the geographic context in which the researcher is actually interested. Neighborhoods and other contexts are often simply not measurable with such administrative units. And as a consequence, scholars employing such units will often wish to employ an alternative approach to measuring context. Doing so can help scholars deal with both UGCoP and MAUP.

## APPROACHES TO ADDRESSING MAUP AND UGCoP

Although both MAUP and UGCoP are presented as 'problems' in the literature, this does not imply that they have solutions. MAUP is an inevitable methodological consequence of employing areal data from a spatial plane that can be partitioned in an infinite number of ways. UGCoP is a substantive problem that we often lack the theoretical knowledge of contextual influences on political behaviors to address. In short, neither MAUP nor UGCoP presents an easy solution.

Scholars have, however, derived various approaches for mitigating the problems posed by both MAUP and UGCoP. Three particularly promising approaches have received considerable scholarly attention. These approaches are bespoke neighborhoods, self-drawn maps and individualized GPS tracking data of political phenomena.

As Johnston et al. (2004: 351) note, the idea of bespoke neighborhoods was developed independently by two sets of researchers (Buck, 2001; Johnston et al., 2000; MacAllister et al., 2001) in the UK in the late 1990s. The basic idea of bespoke neighborhoods is to identify a different contextual environment for every spatial location in the plane. This is done by building out from a survey respondent's location to include a set number of nearest persons to the respondent's home or a set distance from this location. Thus, for example, using the smallest available areal units (say census blocks for US Census data), one then creates progressively larger 'neighborhoods' from respondent *i*'s location utilizing areal units capturing the nearest 50, 100, 500, 5,000, or any other arbitrary increment of people from that location. Alternatively, if one employs a distance metric, one could build out the neighborhoods by utilizing all units whose centroids are within 1 mile, 10 miles, 100 miles, and so on from respondent *i*'s location.

Two advantages of the bespoke neighborhoods approach are that it allows one to

examine neighborhoods of different scales and does so from an individual origin location. The former moves us beyond having a single arbitrary context and allows us to examine the effects of different contextual definitions. However, the contexts can still remain somewhat arbitrary in that there's often no substantive reason why the 50 nearest neighbors would have a greater effect than the 100 nearest neighbors on *i*'s behavior, nor why these are the appropriate cutpoints to examine (the same limitation holds with arbitrary distance cutoffs). Substantive spatial theory remains too imprecise to provide justifications for arbitrary cutpoints such as this (and indeed such cutpoints may be inappropriate in that a set number of nearest neighbors or a specific distance may not be the appropriate metrics for gauging contextual influences). Still, the bespoke neighborhoods approach provides an advantage in considering multiple alternative contextual definitions that can then be chosen based on measures such as RMSE or information criteria.

The second advantage of bespoke neighborhoods is its location-specific origin for the creation of contexts. This is helpful in moving us away from static administrative boundaries to recognize that substantively important contexts for individuals differ within such boundaries and often transcend them. Location-specific origins can provide greater verisimilitude in identifying the contexts that shape individuals' behavior and modeling their effects.

Perhaps ideally, though, one would retain the individual origin component of the bespoke neighborhoods approach while also allowing for greater flexibility in the building of neighborhoods from this origin. Wong et al. (2012) present such an approach via self-drawn maps. In their analysis respondents were given a map and asked to draw their community. They find that respondents' self-defined communities differ from areal units created for governmental administrative purposes such as block groups, that these communities differ across respondents and

that they vary considerably in size across respondents.

Importantly, however, Wong et al. find that respondents' self-defined communities also come with misperceptions about these communities. Specifically, both black and white respondents overestimate the percentage of African Americans in their communities and underestimate the percentage of whites (Wong et al., 2012: 1163). Of course, such misperceptions might be endemic to all contextual units, including administratively defined ones. Velez and Wong (2017) explore this and find that objective administrative boundaries (zip code tabulation areas, ZCTAs) actually predict respondents' demographic perceptions of their contexts better than user-defined boundaries do. This argues that an alternative, more objective approach that is not dependent upon self-reports of context may be needed to properly measure individuals' contexts.

Individualized global positioning system (GPS) tracking of individuals provides just such an approach to measuring context. The idea of this approach is to provide an individualized measure of context, but one based upon objective data drawn from GPS tracking of where subjects were located throughout the day instead of self-drawn (and potentially invalid) maps of subjects' contexts. Beyond employing objective data on individuals' contexts, this approach has the added advantage of providing a dynamic measure of context that recognizes that individuals' contexts differ over the course of the day (Moore and Reeves, 2016). As a result, the behaviors that individuals may be exposed to, e.g., in their neighborhood in the morning or evening may differ from those they encounter during the workday. A GPS tracking measure provides the ability to measure the amount of time spent – and when it was spent – in these various contexts.

Utilizing data on more than 400 individuals, Moore and Reeves (2016: 13) find that GPS tracking data provide a nuanced and dynamic measure of individuals' actual contexts that



more static measures miss. Specifically, they find that static administrative measures of contexts such as census blocks overstate the demographic homogeneity of individuals' contexts. Individuals encounter more members of groups underrepresented in static homogeneous census blocks. Static measures are also found to understate the contextual diversity that individuals encounter in their daily lives. And problems of aggregation are not dependent upon high levels of aggregation, as the same mismeasurement of context occurs at the extremely low level of the block group.<sup>12</sup>

## CONCLUSION

Political scientists have long been interested in contextual influences on political behaviors. Too often, however, we haven't consulted the geographic literature on context. Two central concerns in geography – the modifiable areal unit problem (MAUP) and the uncertain geographic context problem (UGCoP) – are directly applicable to political scientists' study of contextual effects.

Happily, political scientists have been active in developing advances in measuring context. While bespoken neighborhoods have seen only limited use within the discipline, political scientists such as Wong et al. (2012) and Moore and Reeves (2016) have been among those in the vanguard of using self-drawn maps and GPS tracking data to better measure context. For these approaches to gain more widespread use within political science, however, we need to move outside of disciplinary silos and more seriously engage with the geographic literature on the two problems these approaches seek to address – MAUP and UGCoP. In short, we need to take more seriously the insights of geographers on these two problems in our contextual studies in order to more fully utilize the new advances in defining and measuring contexts that are being developed by political scientists and other scholars.

For instance, studies of social, economic and political contexts should more frequently examine how the scale and aggregation problems are affecting their findings. How does the use of different contextual units affect the substantive results that scholars draw? Is there a more theoretically defensible contextual unit that the researcher could employ, one that helps to overcome the UGCoP? How do the results utilizing this contextual definition differ from those in the preceding MAUP-sensitive analysis? Finally, how do the effects of this theoretically based context differ from those based on self-drawn maps and GPS tracking data?

In short, context is central to our understanding of political phenomena. Gone are the days when political scientists felt comfortable treating units – whether survey respondents or countries – as atomistic entities divorced from their surroundings. But as a discipline we still do not take the insights of geographers on context, particularly on MAUP and UGCoP, seriously enough. Until we do, and examine the sensitivity of our results to different contextual definitions, we will not be able to make full use of the very real advances that political scientists and other scholars are developing in measuring context.

## Notes

- 1 Google Scholar data in this section were accessed December 20, 2018.
- 2 These articles are Cho and Gimpel (2007); Enos (2016); Hersh and Clayton Nall (2016).
- 3 The articles are Wong et al. (2012); Fraga (2016); Velez and Wong (2017).
- 4 The article is Kelemen and Pavone (2018).
- 5 The articles are Kirby and Ward (1987) and Nathan (2016).
- 6 The articles are Fjelde et al. (2014) and Schutte (2017).
- 7 See, e.g., Bobo (1988); Fossett and Kiecolt (1989); Giles and Buckner (1993); Giles and Evans (1986); Quillian (1996); Taylor (1998).
- 8 See Putnam (1966); Huckfeldt and Sprague (1987).
- 9 Bledsoe et al. (1995); Sigelman and Welch (1993).

- 10 Oliver and Mendelberg (2000).
- 11 Portions of this discussion of the modifiable areal unit problem were previously published in Darmofal (2015). Reprinted with permission.
- 12 The individualized GPS tracking approach does not, however, come without costs. As Kwan (2012: 966) notes, GPS tracking of subjects does present significant privacy concerns regarding subjects.

## REFERENCES

- Allport, Gordon W. 1954. *The Nature of Prejudice*. Cambridge: Addison-Wesley Publishing Company.
- Baybeck, Brady. 2006. 'Sorting Out the Competing Effects of Racial Context.' *Journal of Politics* 68(2): 386–396.
- Bledsoe, Timothy, Susan Welch, Lee Sigelman and Michael Combs. 1995. 'Residential Context and Racial Solidarity among African Americans.' *American Journal of Political Science* 39(2): 434–458.
- Bobo, Lawrence. 1988. 'Group Conflict, Prejudice, and the Paradox of Contemporary Racial Attitudes.' In *Eliminating Racism*, eds P. Katz and D. Taylor, pp. 85–114.
- Books, John, and Charles Prysby. 1988. 'Studying Contextual Effects on Political Behavior.' *American Politics Quarterly* 16(2): 211–238.
- Branton, Regina P., and Bradford S. Jones. 2005. 'Reexamining Racial Attitudes: The Conditional Relationship Between Diversity and Socioeconomic Environment.' *American Journal of Political Science* 49(2): 359–372.
- Browning, Rufus P., Dale Rogers Marshall, and David H. Tabb. 1984. *Protest Is Not Enough*. Berkeley: University of California Press.
- Browning, Rufus P., and Dale Rogers Marshall, eds. 1986. 'Black and Hispanic Power in City Politics: A Forum.' *PS* 19: 573–640.
- Buck, N. 2001. 'Identifying Neighbourhood Effects on Social Exclusion.' *Urban Studies* 38(12), 2251–2275.
- Carsey, Thomas M. 1995. 'The Contextual Effects of Race on White Vote Behavior: The 1989 New York City Mayoral Election.' *Journal of Politics* 57(1): 221–228.
- Cho, Wendy K. Tam, and James G. Gimpel. 2007. 'Prospecting for (Campaign) Gold.' *American Journal of Political Science* 51(2): 255–268.
- Darmofal, David. 2015. *Spatial Analysis for the Social Sciences*. New York: Cambridge University Press.
- Diez-Roux, Ana V., and Christina Mair. 2010. 'Neighborhoods and Health.' *Annals of the New York Academy of Sciences* 1186: 125–145.
- Enos, Ryan D. 2016. 'What the Demolition of Public Housing Teaches Us about the Impact of Racial Threat on Political Behavior.' *American Journal of Political Science* 60(1): 123–142.
- Fjelde, Hanne, and Lisa Hultman. 2014. 'Weakening the Enemy: A Disaggregated Study of Violence against Civilians in Africa.' *Journal of Conflict Resolution* 58(7): 1230–1257.
- Fossett, Mark A., and K. Jill Kiecolt. 1989. 'The Relative Size of Minority Populations and White Racial Attitudes.' *Social Science Quarterly* 70(4): 820–835.
- Fraga, Bernard L. 2016. 'Redistricting and the Causal Impact of Race on Voter Turnout.' *The Journal of Politics* 78(1): 19–34.
- Giles, Micheal W., and Arthur Evans. 1986. 'The Power Approach to Intergroup Hostility.' *Journal of Conflict Resolution* 30(3): 469–486.
- Giles, Micheal W., and Melanie A. Buckner. 1993. 'David Duke and Black Threat: An Old Hypothesis Revisited.' *The Journal of Politics* 55(3): 702–713.
- Hersh, Eitan D., and Clayton Nall. 2016. 'The Primacy of Race in the Geography of Income-Based Voting: New Evidence from Public Voting Records.' *American Journal of Political Science* 60(2): 289–303.
- Huckfeldt, Robert, and Carol Weitzel Kohfeld. 1989. *Race and the Decline of Class in American Politics*. Urbana: University of Illinois Press.
- Huckfeldt, Robert, and John Sprague. 1987. 'Networks in Context: The Social Flow of Political Information.' *American Political Science Review* 81(4): 1197–1216.
- Johnston, Ron, Kelvyn Jones, Simon Burgess, Carol Propper, Rebecca Sarker and Anne Bolster. 2004. 'Scale, Factor Analyses, and Neighborhood Effects.' *Geographical Analysis* 36(4): 350–368.
- Johnston, R. J., C. J. Pattie, D. F. L. Dorling, I. MacAllister, H. Tunstall and D. J. Rossiter.

2000. 'The Neighbourhood Effect and Voting in England and Wales: Real or Imagined?' In *British Elections and Parties Review*, Volume 10, eds P. J. Cowley, D. T. Denver, A. T. Russel and L. Harrison, 47–63. London: Frank Cass.
- Johnston, Ron, Carol Propper, Rebecca Sarker, Kelynn Jones, Anne Bolster, and Simon Burgess. 2005. 'Neighbourhood Social Capital and Neighbourhood Effects.' *Environment and Planning A* 37(8): 1443–1459.
- Katznelson, Ira. 1982. *City Trenches: Urban Politics and the Patterning of Class in the United States*. Chicago: University of Chicago Press.
- Kelemen, R. Daniel, and Tommaso Pavone. 2018. 'The Political Geography of Legal Integration: Visualizing Institutional Change in the European Union.' *World Politics* 70(3): 358–397.
- Key, Valdimer Orlando. 1949. *Southern Politics in State and Nation*. New York: Knopf.
- Kirby, Andrew M., and Michael D. Ward. 1987. 'The Spatial Analysis of Peace and War.' *Comparative Political Studies* 20(3): 293–313.
- Kwan, Mei-Po. 2012. 'The Uncertain Geographic Context Problem.' *Annals of the Association of American Geographers* 102(5): 958–968.
- MacAllister, I., R. J. Johnston, C. J. Pattie, H. Tunstall, D. F. L. Dorling and D. J. Rossiter. 2001. 'Class Dealignment and the Neighbourhood Effect: Miller Revisited.' *British Journal of Political Science* 31(1): 41–59.
- Moore, Ryan T., and Andrew Reeves. 2016. 'Defining Racial and Ethnic Context with Geolocation Data.' Manuscript.
- Most, Benjamin A., and Harvey Starr. 1980. 'Diffusion, Reinforcement, Geopolitics, and the Spread of War.' *American Political Science Review* 74(4): 932–946.
- Most, Benjamin A., and Harvey Starr. 1982. 'Case Selection, Conceptualizations and Basic Logic in the Study of War.' *American Journal of Political Science* 26(4): 834–856.
- Nathan, Noah L. 2016. 'Local Ethnic Geography, Expectations of Favoritism, and Voting in Urban Ghana.' *Comparative Political Studies* 49(14): 1896–1929.
- Newman, Benjamin J. 2013. 'Acculturating Contexts and Anglo Opposition to Immigration in the United States.' *American Journal of Political Science* 57(2):374–390.
- Newman, Benjamin J., Yamil Velez, Todd K. Hartman and Alexa Bankert. 2015. 'Are Citizens "Receiving the Treatment"? Assessing a Key Link in Contextual Theories of Public Opinion and Political Behavior.' *Political Psychology* 36(1): 123–131.
- Oliver, J. Eric, and Janelle Wong. 2003. 'Inter-group Prejudice in Multiethnic Settings.' *American Journal of Political Science* 47(4): 567–582.
- Oliver, J. Eric, and Tali Mendelberg. 2000. 'Reconsidering the Environmental Determinants of White Racial Attitudes.' *American Journal of Political Science* 44(3): 574–589.
- Openshaw, Stan. 1983. *The Modifiable Areal Unit Problem*. Norwich: Geo Books.
- Openshaw, Stan and Peter J. Taylor. 1979. 'A Million or So Correlation Coefficients: Three Experiments on the Modifiable Areal Unit Problem.' In *Statistical Applications in the Spatial Sciences*, ed. N. Wrigley, pp 127–144. London: Pion.
- Openshaw, Stan and Peter J. Taylor. 1981. 'The Modifiable Areal Unit Problem.' In *Quantitative Geography: A British View*, eds N. Wrigley and R. J. Bennett, pp. 60–69. London: Routledge.
- Przeworski, Adam. 1974. 'Contextual Models of Political Behavior.' *Political Methodology* 1(1): 27–61.
- Putnam, Robert D. 1966. 'Political Attitudes and the Local Community.' *American Political Science Review* 60(3): 640–654.
- Quillian, Lincoln. 1996. 'Group Threat and Regional Change in Attitudes toward African-Americans.' *American Journal of Sociology* 102(3): 816–860.
- Ray, James Lee. 2001. 'Integrating Levels of Analysis in World Politics.' *Journal of Theoretical Politics* 13(4): 355–388.
- Rocha, Rene R., and Rodolfo Espino. 2009. 'Racial Threat, Residential Segregation, and the Policy Attitudes of Anglos.' *Political Research Quarterly* 62(2): 415–426.
- Schutte, Sebastian. 2017. 'Violence and Civilian Loyalties: Evidence from Afghanistan.' *Journal of Conflict Resolution* 61(8): 1595–1625.
- Sigelman, Lee, and Susan Welch. 1993. 'The Contact Hypothesis Revisited: Black-White

- Interaction and Positive Racial Attitudes.' *Social Forces* 71(3): 781–795.
- Singer, J. David. 1961. 'The Level-of-Analysis Problem in International Relations.' *World Politics* 14(1): 77–92.
- Starr, Harvey, and Benjamin A. Most. 1976. 'The Substance and Study of Borders in International Relations Research.' *International Studies Quarterly* 20(4): 581–620.
- Starr, Harvey, and Benjamin A. Most. 1983. 'Contagion and Border Effects on Contemporary African Conflict.' *Comparative Political Studies* 16(February): 92–117.
- Taylor, Marylee C. 1998. 'How White Attitudes Vary with the Racial Composition of Local Populations: Numbers Count.' *American Sociological Review* 63(4): 512–535.
- Tobler, Waldo R. 1969. 'Geographical Filters and Their Inverses.' *Geographical Analysis* 1(3): 234–253.
- Tobler, Waldo R. 1975. 'Linear Operators Applied to Areal Data.' In *Display and Analysis of Spatial Data*, eds J. C. Davis and M. J. McCullagh, pp. 14–37. London: John Wiley & Sons.
- Velez, Yamil Ricardo, and Grace Wong. 2017. 'Assessing Contextual Measurement Strategies.' *The Journal of Politics* 79(3): 1084–1089.
- Voss, D. Stephen, and Penny Miller. 2001. 'Following a False Trail: The Hunt for White Backlash in Kentucky's 1996 Desegregation Vote.' *State Politics & Policy Quarterly* 1(1): 62–80.
- Wong, Cara, Jake Bowers, Tarah Williams and Katherine Drake Simmons. 2012. 'Bringing the Person Back In: Boundaries, Perceptions, and the Measurement of Racial Context.' *Journal of Politics* 74(4): 1153–1170.



# Visualizing Data in Political Science

Richard Traunmüller

## INTRODUCTION

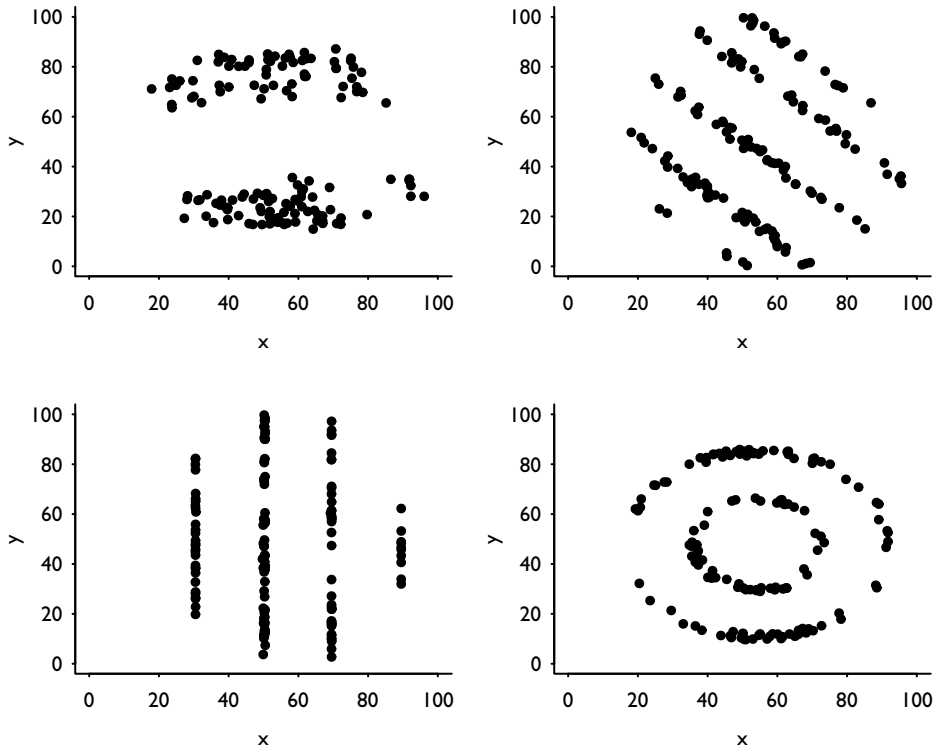
Much of modern political science is concerned with the analysis of data. Both the sheer mass of data and the variety of different data sources that are used to understand political processes have increased dramatically over the past years. Therefore, it should come as no surprise that political science has also developed a growing interest in data visualization. Indeed, few methods can match the utility of data visualization to explore, describe and communicate patterns in quantitative information.

The power of data visualization is easy to demonstrate. Figure 25.1 displays the relation between two variables in four different data sets. We can easily and almost instantly detect four very distinct data patterns: two well separated groups or clusters, different striped patterns that could indicate serial correlation and discrete variables, and a rather peculiar circular pattern.

Interestingly, these striking patterns would have been ignored had we summarized the

data using descriptive statistics that reduce the data points to fewer and more manageable numbers. For instance, we could have calculated the means and find that they are  $X = 54.3$  and  $Y = 47.8$  in all four data sets. Looking at the standard deviations for each variable would have yielded exactly the same in all four data sets:  $\sigma_X = 16.7$  and  $\sigma_Y = 26.8$ , respectively. Looking at the correlation between  $X$  and  $Y$  gives  $\rho_{YX} = -0.1$  in all cases and regressing  $Y$  on  $X$  in a simple linear model would have given the same intercept  $\alpha = 53.8$ , the same coefficient  $\beta = -0.1$  and the same measure of fit  $R^2 = .005$ . These data sets were generated by Matejka and Fitzmaurice (2017) and are a modern version of Anscombe's (1973) quartet that for decades has served scholars as a lesson to look at their data.

Data visualization is concerned with the visual representation of abstract variables and their relations. In this regard it differs from scientific visualization, which is used to visualize concrete physical objects or phenomena



**Figure 25.1** Four scatter plots of four data sets that show wildly different patterns—although summary statistics are identical

such as skeletons, planets or geographical topographies. Data visualization can be understood as a translation tool that assigns abstract numerical values to physical properties such as spatial position along a scale, the length of a bar or the geometric shape and color of a plotting symbol. However, data visualization involves more than simply mapping numbers to visual stimuli. Ideally, data visualization is a method that helps us and our audience understand the political world by assisting analytical thinking. It is common to distinguish two overarching goals of data visualization (e.g. Gelman and Unwin, 2013). *Data visualization for analysis* is mainly used to explore a data set, to diagnose potential problems – such as missing or implausible values – or to uncover unknown patterns and relations which suggest scientific hypotheses and modeling strategies. *Data visualization*

*for presentation*, on the other hand, serves as an efficient means to communicate the results of a data analysis and possibly has the goal of attracting attention and influencing human decisions.

Compared to tabular displays or numerical summaries, data visualization has several advantages (Anscombe, 1973; Cleveland, 1994; Jacoby, 1997; Jacoby and Schneider, 2010; Tufte, 2001; Keim and Ward, 2003; Ware, 2013). First, visualization easily handles large and even huge amounts of quantitative information. The reason is that visualization abstracts from single data points and instead turns them to an emergent new whole – a certain distribution or pattern. In this way virtually millions of data points can be easily perceived and processed. Second, visualization (usually) retains full information and does not rely on assumptions

concerning the distributional nature of the data.<sup>1</sup> Any parametric summary of data leads to a reduction and thus to a loss of potentially interesting information. They may also rely on explicit or implicit assumptions that are incompatible with the data or overly restrictive. Visualization allows for the discovery of unexpected patterns that are either interesting in their own right and thus constitute the end point of an analysis, or, alternatively and crucially, motivate follow-up questions and new directions for exploration. The third advantage of visualization is exactly this: it encourages the search for the sources of observed patterns and the processes that generated them. In this sense visualization can be viewed as an exploratory hypothesis generating device.

In what follows I will provide a selective overview of the state of the art of modern data visualization from a political science perspective. I will begin with a brief empirical analysis of graph use in current political science and then turn to data visualization as an exploratory tool for political science data. Next to table lens plots, I will introduce visual methods that were genuinely designed to display high-dimensional data structures: parallel coordinate plots and small multiple designs. I then turn to recent advances in data visualization that greatly expand the utility of visual methods: the visual exploratory model analysis and visual inference to protect against over-interpretation of random patterns.

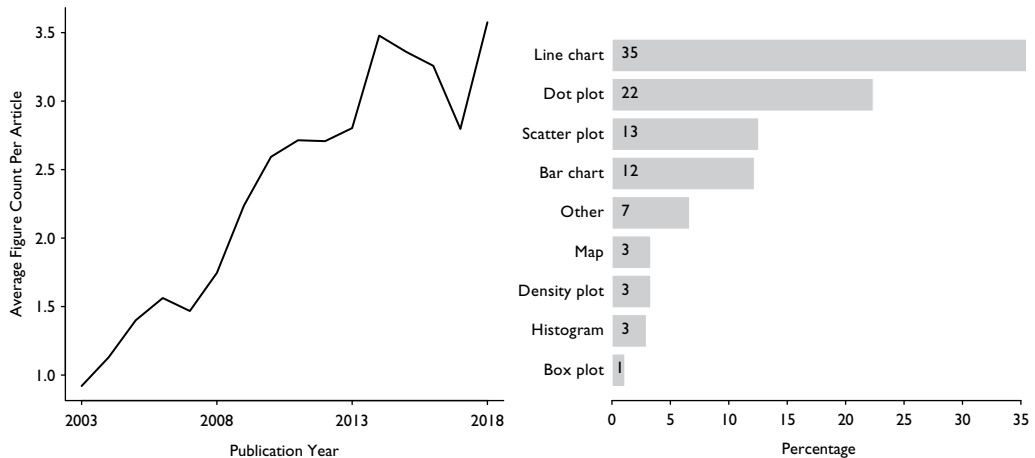
For a related up-to-date review of data visualization from a statistical perspective, see Cook et al. (2016); for a sociological perspective, see Healy and Moody (2014). Classic contributions, design advice and sources of inspiration are Bertin (1983), Tufte (2001, 2006) and Cleveland (1994). The important distinction between statistical graphics and infovis is discussed in Gelman and Unwin (2013) and in the ensuing debate. Good introductions to data visualization as an analytical tool are provided by Few (2012) and Unwin (2015). For data visualization as a presentational tool for communication, see Few (2009)

and Kirk (2015). A seminal experiment on the perception of statistical graphs is Cleveland and McGill (1984). Heer and Bostock (2010) replicate and Talbot et al. (2014) extend the results based on crowd-sourced experiments. The best reference for the cognitive psychological foundations of data visualization is Ware (2013). Wilkinson (2005) provides a formalization of data visualization based on the grammar of graphics, which was implemented in the ggplot package by Wickham (2010). Arguably the number one tool for data visualization in political science is the statistical programming language R, for which Murrell (2018), Chang (2012) and Healy (2018) are excellent references. Those interested in producing graphs using STATA should take a look at Mitchell (2012).

## GRAPH USE IN POLITICAL SCIENCE

Counter to many other methodological developments, data visualization is an invention of the social sciences. Historically, graphical methods have co-evolved with the advent of new social, political and economic data collections (Friendly, 2009). Many of the most widely used and successful graphical formats – such as the bar chart, the line chart and the pie chart (but not the scatterplot) – were developed by the Scottish political economist William Playfair to visualize political, economic and social data (Playfair, 1786, 1801). Several important milestones of data visualization – for instance, Minard's map of Napoleon's March to Russia or Nightingale's visualization of the causes of death of British soldiers during the Crimean War – are not only based on social scientific data, but also served very specific political purposes such as influencing policy makers. Finally, the greatest icon of modern data visualization, Edward Tufte, began his career as a political science professor at Yale University.

The reliance on data visualization fell somewhat out of fashion in the 20th century,



**Figure 25.2** Left panel: average number of figures in all articles published in the *AJPS*, 2003–18. Right panel: relative frequency of graphical formats

when ‘serious’ data analysis became associated with significance testing (e.g. Best et al., 2001). Nowadays, and ironically, this dominant mode of conducting statistical inference is itself under attack (Gill, 1999) and data visualization is experiencing a true renaissance. Based on an analysis of all articles published in the *American Journal of Political Science* between February 2003 and March 2018 and on the assumption that these are exemplary for the current state of the art in political science, Figure 25.2 demonstrates that graph use has dramatically increased over the past 15 years.<sup>2</sup> Whereas the average political science article in the discipline’s flagship journal contained roughly one (.92) figure in 2003, graph use has grown to an average of three and a half (3.58) figures per article in 2018. While I rely on figure count as a proxy for graph count, not every figure is necessarily a data visualization display of empirical data; there are also visual representations of mathematical functions, game theoretic decision trees or even just flow charts of theoretical arguments.

Among actual data visualizations, line charts are by far the most popular graphical format in political science: 35% of all graphs published in *AJPS* articles fall into this

category. The second most widely used format is some version of the dot plot (22%), potentially hinting at the influence of Cleveland (1993, 1994) on our discipline. Other common formats are scatter plots (13%) and bar charts (12%). The classical tools for describing continuous distributions – histograms, density plots and Tukey’s box plot – make up 7% of all data visualizations. Although political science has a strong reference to space and geography, maps make up only about 3% of all data visualizations. Finally, 7% fall into a residual category of other visualizations, including, for instance, visual representations of social networks, 3D wireframes and, yes, pie charts. Interestingly, 19% of all visualizations published in the area of political science now make use of color instead of remaining in black and white.

A common visualization technique in political science is to combine several plots into an overall visualization. In fact, only 54% of all visualizations consist of a single plot, whereas 40% combine multiple plots of the same format (in so-called *small multiple designs*, see further below) and 6% multiple plots of different formats (so-called *plot ensembles*, see further below). The average number of plots in these combined visualizations is 4 plots,



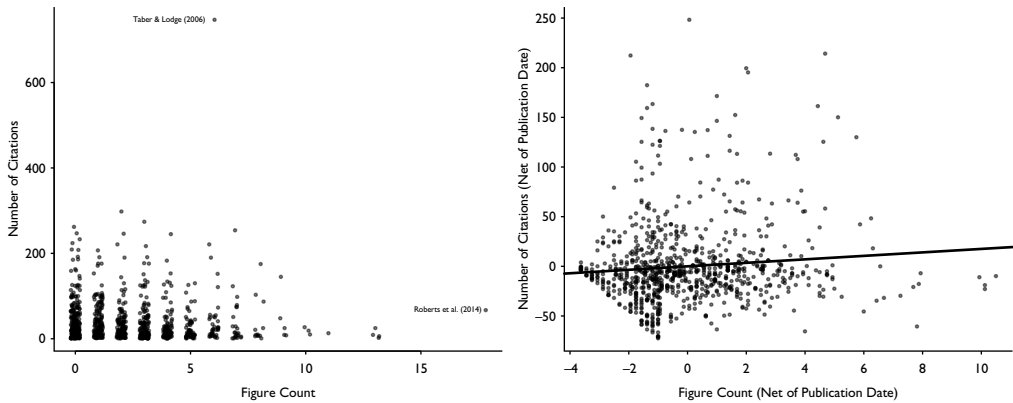
but in several instances many more plots are arranged together, with the maximum being no less than 36 tiny plots in one single visualization. Of all dot plots, 53% are used in the context of small multiples or plot ensembles, as are 50% of all scatter plots, 40% of all line charts and 35% of all bar charts.

To better understand the motivation behind graph use in political science, Kastlelec and Leoni (2007) went through every article from five issues of three leading political science journals – the February and May 2006 issues of *American Political Science Review*, the July 2006 issue of *American Journal of Political Science* and the winter and spring 2006 issues of *Political Analysis* – and found that political scientists never use graphs to present regression results. This has certainly changed over the past ten years. In our own sample, no less than 90% of all dot plots are displayed with error bars or confidence intervals and therefore almost 20% of all visualizations in political science are used to display coefficient estimates or experimental group means along with their associated inferential uncertainties. In addition, 32% of all line charts and therefore 11% of all graphs are marginal effect plots and predicted value plots. In this sense, political science uses graphical methods to visualize not only raw data, but also ‘cooked’ data. This is also an indication that data visualization and statistical modeling and inference are best seen as complementary instead of opposites.

Overall, only a limited number of graphical formats are currently used in political science and many of them are well-known standard types. As a result, these are also the graphical formats that, at least a priori, are likely to work best to convey your results to professional peers, conference audiences or reviewers and editors. This is so because they are part of a commonly shared ‘vocabulary’ and therefore incur low cognitive costs for the audience. According to cognitive psychologist and visualization expert Colin Ware (1998: 178), ‘making radically new designs is more interesting for the designer and leads

to kudos from other designers. But radical designs, being novel, take more effort on the part of the consumer’. Therefore, one should not underestimate the power of those simple and common graphical formats. At the same, and speaking from experience, one cannot overestimate editors’, reviewers’, and even co-authors’ reluctance to engage with innovative yet unusual data visualizations.

An interesting question concerns the impact of using graphs in one’s political scientific work. Echoing the late Stephen Hawking (1988), Barabasi (2010) quipped that there ‘is a theorem in publishing that each graph halves a book’s audience’. If this were true, urging political scientists to make heavy use of data visualization would be a rather difficult point to defend. The left panel of Figure 25.3 is a scatter plot of the (jittered) figure count versus the number of citations of the *AJPS* articles in our sample (citations counts are missing for the two most recent volumes). If anything, it seems to suggest that citations decrease with the number of figures included in an article. It also identifies two extreme cases: Taber and Lodge (2006) in terms of citations (747) and Roberts et al. (2014) in terms of number of figures. But clearly publication date confounds this relationship. The right panel therefore shows the same relation net of the fact that more recent publications need a while to be cited (and in addition excludes the two extreme observations). In other words, it shows the residuals of a simple regression of citations on volume number versus the residuals of a simple regression of figure count on volume number using an adjusted variable plot. The line thus corresponds to the regression line of the relation between citations and graph use, controlling for publication date. Each figure included in an *AJPS* article increases the citation count by 1.75 citations, and, with a p-value of .006, ‘significantly’ so (if one cared about this). Clearly, this is not a particularly strong relation, nor is it necessarily causal. Maybe more successful scholars just produce more graphs.



**Figure 25.3** The relationship between graph use and citation counts of *AJPS* articles, 2003–17. Left panel: simple scatter plot with jitter along the x-axis. Right panel: adjusted variable plot relating residuals of figure count to residuals of citations eliminating the effect of publication date

## VISUALLY EXPLORING AND DESCRIBING STRUCTURE IN POLITICAL SCIENCE DATA

Next to the compelling *presentation* of data summaries, statistical results and quantities of interest (Jacoby and Schneider, 2010; Kestellec and Leoni, 2007; King et al., 2001) statistical graphics can be used as *analytic* tools for various purposes and at various stages of the data analysis (Jacoby, 1997, 2000; Bowers, 2004; Bowers and Drake, 2005; Gelman, 2003; Gelman and Hill, 2007; Kerman et al., 2008). Data analysis in political science usually proceeds in several steps: (1) checking, cleaning and pre-processing the data, (2) exploring and describing structure in the data and (3) statistical modeling of the data and making inferences. Data visualization can help in each step of this analytic process: (a) to detect problems and anomalies in the data, (b) to explore and familiarize oneself with the data and generate hypotheses, and (c) to understand, check and diagnose models. Most of the visualizations produced in the process of data checking, initial data exploration and model evaluation won't find their way into journal articles or book chapters. However, because it is increasingly common to provide lengthy appendices and supplemental information online, at least

some of these visualizations could be documented for an interested audience.

A particular challenge in the visual exploration is that data sets in political science grow increasingly larger and more complex. Here largeness and complexity refer to both the number of observations  $N$  and the number of variables  $K$ . Both dimensions bring their challenges to data visualization (Unwin et al., 2006), but dealing with multidimensionality is particularly tricky.

### Table Lens Plots

Table lens plots are a visualization technique that allows the researcher to view a whole and possibly large data set at one glance (Tennekes et al., 2013). As such it gives a good initial overview and is useful for both checking the raw data for anomalies and exploring data to uncover structure. The basic idea of table lens plots is to first divide all  $N$  observations into  $H$  equally sized classes or *bins*. The distributions of the  $k$  dimensions or variables are then shown within each of these bins. For continuous variables, the distribution is shown using a bar for the means (along with the standard deviation). For categorical variables, the distribution of variable values is shown using stacked bar charts. Missing values are treated as their own category.

The example in Figure 25.4 visualizes data taken from the Swiss census 2010. The  $N = 371,221$  observations are first divided into  $H = 100$  row bins and then sorted by the continuous age variable (`AGE_HARM`). The young are at the bottom and the old at the top and we get an impression of the age distribution (panel A). We also immediately see several relations in the data. For instance, women (`SEX_HARM`) are slightly over-represented in the older age cohorts and the share of foreigners without citizenship (`NATIONALITYCAT_HARM`) is higher in younger cohorts. In addition, there is a clear age-specific pattern in employment status (`CURRACTIVITYSTATUSII`), where the young tend to be in education, the middle aged mostly in full-time employment and the elderly retired. Last, missing values seem to only occur in the religious affiliation variable (this information is voluntary in the Swiss census).

Sorting the observations in the table lens plot along the values of a categorical variable – in this case religious affiliation (`RELIGIOUSCOMMAGGII_HARM`) – reveals new patterns and relations (panel B). For instance, we now see that religious groups differ in their mean age. With a mean age of below 40, Muslims are the youngest religious group in Switzerland. Muslims also have the highest shares of non-citizens and the highest unemployment rates. Importantly, we are now also able to see how the missing values in the religious variable are related to missing values in other variables (which we missed in the previous visualization): respondents who did not disclose their religious affiliation (the bright red segment at the bottom of the table plot for religious belonging) are also less likely to indicate their current employment status.

Decreasing the number of bins to  $H = 10$  gives a smoothed and possibly better, because simpler, impression, albeit at the cost of losing detail (panel C). More detail emerges when we increase the number of bins to  $H = 300$  and zoom in on the data to between 54% and 62%, focusing on an interesting part (panel D). This nicely separates the religious groups, in the sense of forming homogenous

bins, and we find that the Jewish community is regionally highly concentrated (`RES_CANTON_HARM`). Filtering out every religious group (i.e. removing them from the visualization) to focus on Jews and ordering by employment status reveals that in this religious group it is overwhelmingly women who stay at home.

Table lens plots illustrate an important point in visual data exploration which they share with more conventional graphical formats, such as histograms or density plots. Exploratory graphs are essentially a model in that the ‘objective is to construct an abstraction that highlights the salient aspects of the data without distorting any features or imposing undue assumptions’ (Jacoby, 1997: 13). By definition, a model is a simplified representation of the world. As such it is always an abstraction that ignores some details. But simplification should not result in misrepresentation. Because table lens plots – just as histograms – divide the data into a discrete number of bins, they are reducing and potentially distorting the information in the raw data. In particular, the choice of the number and widths of the bins,  $d$ , impacts the appearance of the visualization and thus the patterns that become visible.

Thus these data visualizations come with a typical *variance–bias tradeoff* (cf. Jacoby, 1997). Narrow bins or bandwidths produce high-variance, low-bias graphs. That is, the graph closely follows the data (low bias) and shows much detailed variation (high variance). Broader bins or bandwidths produce low-variance, high-bias graphs. They give a smoother picture of the data that eliminates some of the details (low variance), but at the cost of deviating from the actual observed data (high bias). Since it is easy to change bin size and bandwidths it is advisable to always experiment with them to see how the visual pattern changes and what details emerge. A good general strategy is to start with narrow bins and bandwidths to see what the data have to say and then steadily increase them until a good representation which captures the most important features is found.

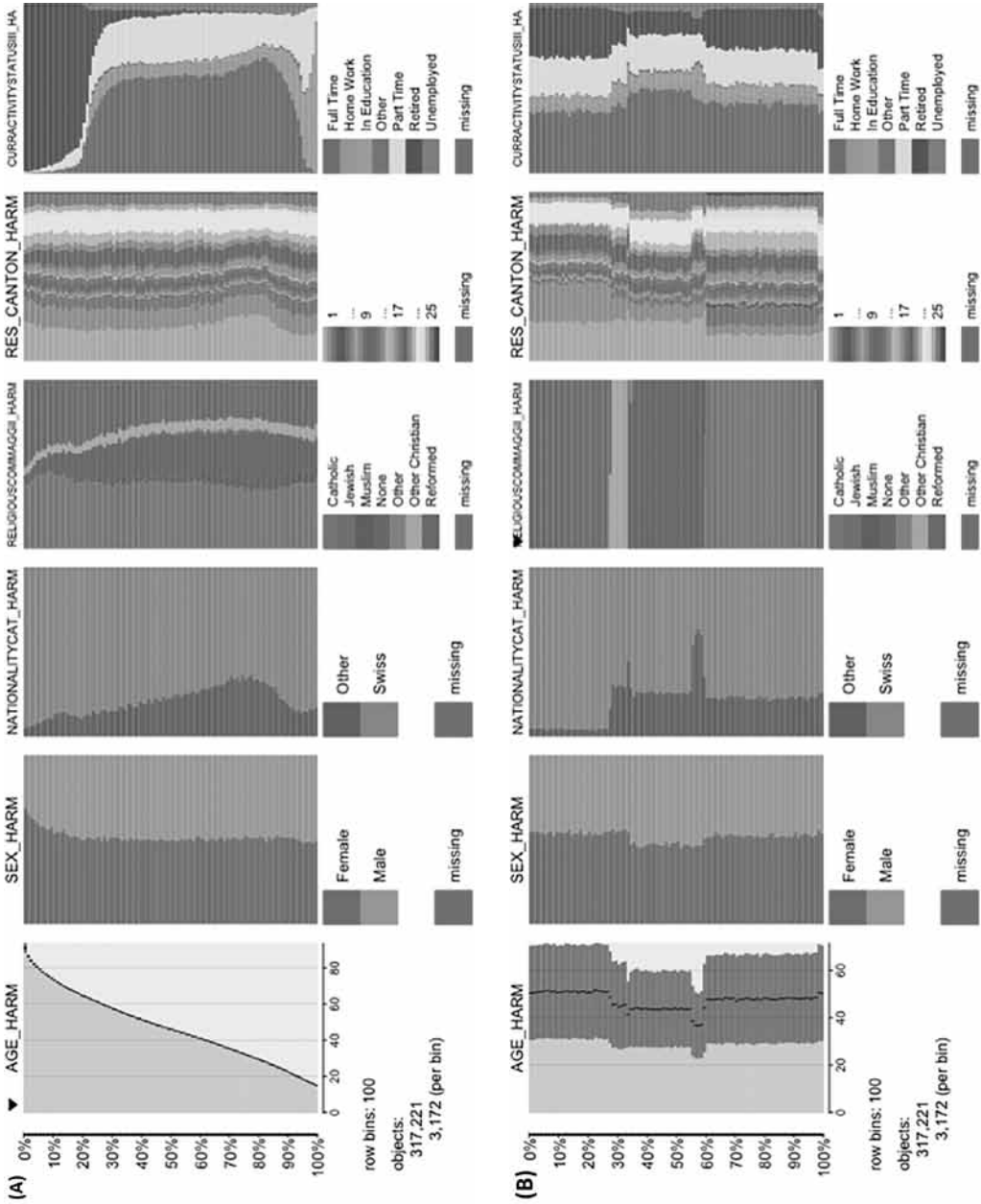


Figure 25.4 Table lens plots of Swiss census data 2010. See main text for description

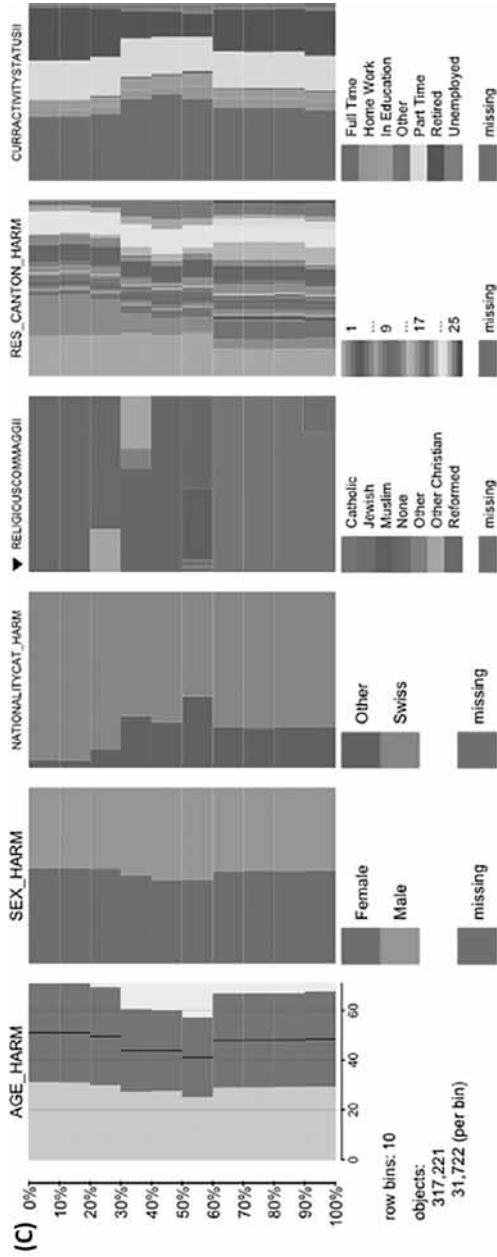


Figure 25.4 Continued

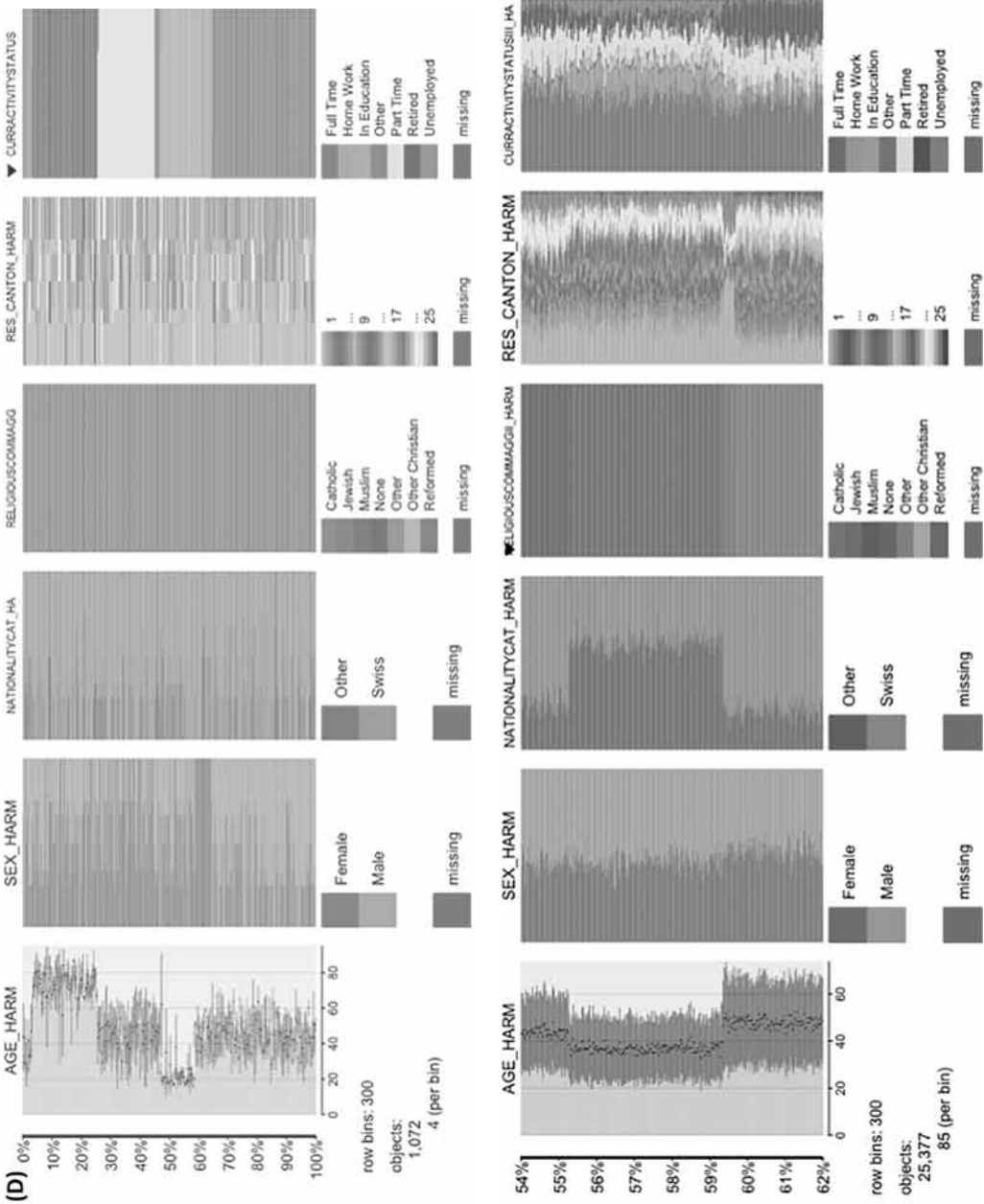


Figure 25.4 Continued

While table lens plots easily accommodate large data sets with many observations,  $N$ , they are limited in terms of the number of variables,  $K$ , they are able to reasonably visualize. Visualization techniques to deal with this ‘curse of dimensionality’ fall broadly into two categories: (a) complex visualization techniques that are explicitly designed to show a high number of dimensions or variables (e.g. *parallel coordinate plots*), and (b) visualization techniques that disaggregate high-dimensional data into a series of simpler, one- or two-dimensional graphics and arrange them in an effective way (e.g. *small multiple designs*).

### **Parallel Coordinate Plots**

A visualization method that is well suited for the analysis of high-dimensional data structures but rarely encountered in political science is the *parallel coordinate plot* (Inselberg, 2008; Wegman, 1990). A parallel coordinate plot solves the problem of the ‘curse of dimensionality’ by doing what its name implies: it maps  $K$  variables along  $K$  coordinates which are aligned in a parallel fashion, instead of orthogonally to each other. A single observation corresponds to a profile line that connects the variable values of the  $K$  variables. Next to overall variable distributions, parallel coordinate plots help find correlations between many variables at the same time and identify high-dimensional clusters in a data set. Negative correlations are visible as crossing, positive correlations as parallel lines. Clusters are visible as separations along an axis and how they are propagated across the plot.

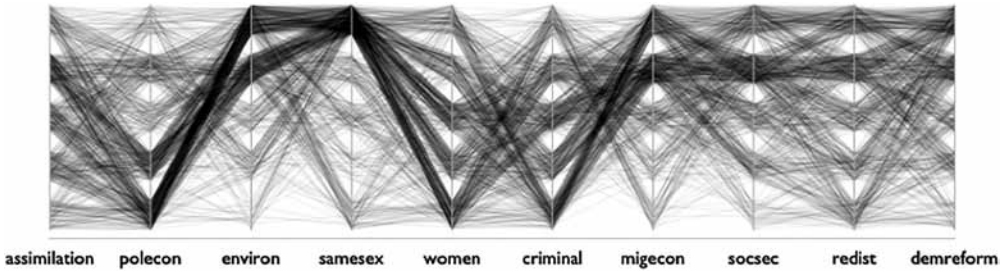
Using a parallel coordinates plot, the example in Figure 25.5 visualizes the policy preferences of  $N=1,069$  candidates for the state elections in Bavaria (2013), Hesse (2013) and Saxony (2014), comparing them across  $K=10$  policy areas.<sup>3</sup> Since the policy items were measured on a five-point Likert scale (higher values indicate higher support

for a policy), this parallel coordinate relies on two additional visualization techniques to deal with the common problem of overplotting – *jittering*, i.e. adding random noise to the variables’ values, and *alpha blending*, i.e. rendering the lines transparent.

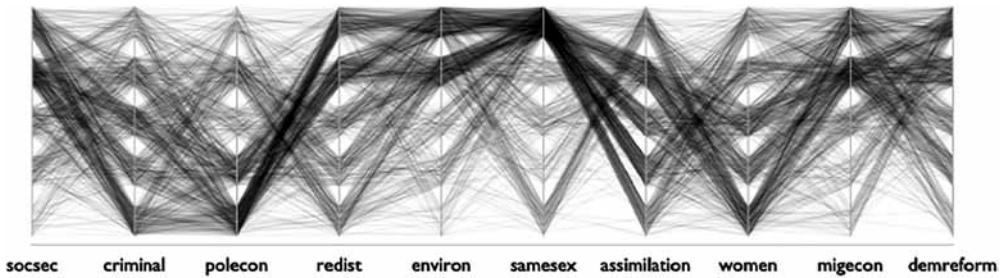
At first glance this plot may look intimidating, but several patterns emerge rather quickly. For instance, there seems to be strong consensus among candidates concerning environmental protection (environ) and the rights of same-sex couples (samesex). The bulk of the lines is concentrated on the upper ends of the two axes. More contested policy areas are affirmative action for women (women) and the punishment of criminals (criminal). Here the lines are distributed more or less equally across the five item values. In addition, the crossing of the lines indicates a strong negative correlation. Candidates that favor affirmative action for women tend to be against tighter laws for criminals, and vice versa.

Since the identification of patterns in parallel coordinate plots depends heavily on the sorting of the parallel axes, it is advisable to vary their order. For instance, a random re-sorting in Figure 25.6 shows that preferences for affirmative action for women is also negatively correlated to an immigration policy that stresses migrants’ assimilation (assimilation). Perhaps unsurprisingly, political candidates in favor of regulatory intervention in the economy (polecon, the item is reversed) also show strong support for social security (socsec).

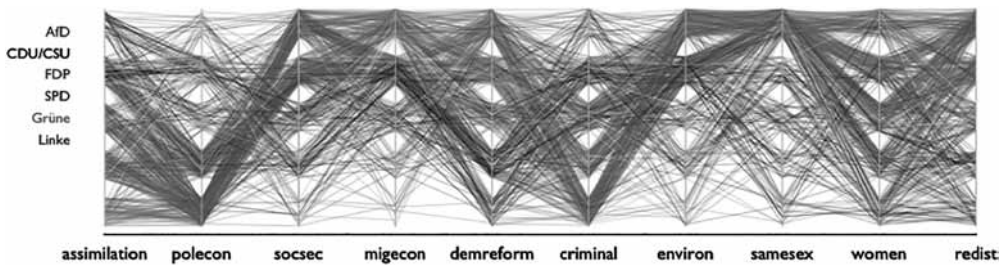
Making use of color, we are able to identify high-dimensional clusters in the parallel coordinate plot. Figure 25.7 colors all observations according to their party affiliation, where the colors correspond to the familiar signature colors of the six most important parties in the German party system (CDU/CSU, SPD, Greens, AfD, FDP, Lefts). As one would expect, policy preferences tend to cluster along party lines. Quite visible are the differences along the classic economic left–right dimension: regulatory intervention



**Figure 25.5** Parallel coordinate plot of the policy preferences of over a thousand political candidates



**Figure 25.6** Parallel coordinate plot of the policy preferences of over a thousand political candidates. Axes have been re-sorted



**Figure 25.7** Parallel coordinate plot of the policy preferences of over a thousand political candidates. Lines are colored according to party affiliation

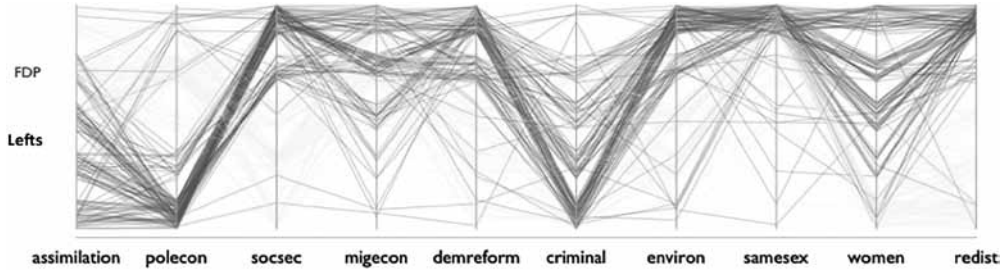
in the economy (polecon), social security (socsec), and redistribution (redistrib). Ideological party differences are better visible if we reduce the complexity somewhat and concentrate on a comparison of two parties while filtering out the rest (Figures 25.8–25.10). This idea of re-using the same graphic format on different subsets of the

data leads us directly to the next visualization technique.

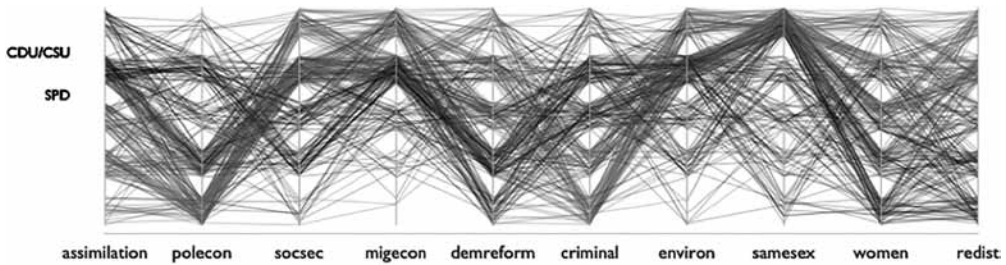
### ***Small Multiple Designs***

A particularly powerful visual strategy for multidimensional data is to repeatedly

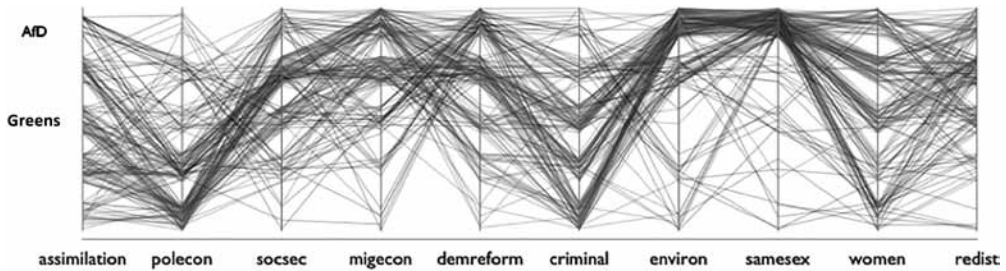




**Figure 25.8** Parallel coordinate plot of the policy preferences of political candidates of the FDP and the Left



**Figure 25.9** Parallel coordinate plot of the policy preferences of political candidates of the CDU/CSU and SPD



**Figure 25.10** Parallel coordinate plot of the policy preferences of political candidates of the AfD and the Greens

apply simple lower-dimensional graphical formats to  $G$  different subsets of the data and to arrange these  $G$  subplots in a meta-visualization. The subsets are themselves defined in terms of variable values or combinations of variable values. This technique has different names, such as *small multiples* (Tufte, 2001), *trellis displays* (Becker et al.,

1996) or *collections* (Bertin, 1983). A special case of this method is a scatter plot matrix that shows all  $G=K(K-1)$  bivariate scatter plots of  $K$  variables in one matrix. The key design feature of small multiples is that the single plots are shrunken in size, have the same appearance and size and also have constant axis scales. In other words,

single displays differ only in the subsets of the data they present. In this way it is possible to make very efficient subgroup comparisons as well as to identify conditional patterns and relations. The efficiency of small multiples is further increased by thoughtful ordering and arrangement.

A well-known political science example of a small multiple design is the visualization of (modeled) survey data on a controversial policy: the support of school vouchers (Figure 25.11, Gelman, 2009). The chosen graphical format is a choropleth map where color shade is used to encode the average support across the US and to give a sense of the geographic variation. Somewhat un-intuitively, green regions show lower support and orange regions higher support for school vouchers. We should also note that this color scheme is unfortunate given the possibility that there will be color-blind individuals in the audience. Be that as it may, the key design feature is the repeated application of the same map format to different subgroups of the data. In this case the subgroups are five different income groups, ranging from low-income on the left to high-income on the right. In this way it becomes clear that support for school vouchers increases with income in all states, with the exception of Wyoming and the Southwest. The small multiple design thus reveals that the relationship between income and policy preference depends on regional characteristics of the states.

We can expand this conditional analysis by bringing in a further variable, ethnic-religious group identity, and by arranging the single graphics in a table or matrix cross-classified by seven identities times five

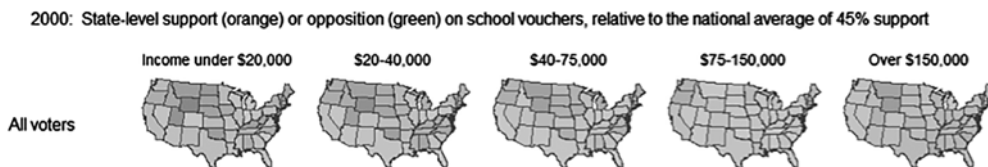
income groups (Figure 25.12). This shows that the idea of school vouchers is supported by high-income white Catholics and Evangelicals and by low-income Hispanics. Generally speaking, for whites the preference for school vouchers increases with income, whereas for blacks it decreases with income – a classic two-way interaction effect.

As in other visual displays, patterns are easier to detect if we sort the small multiples in a meaningful way (Figure 25.13). Since income groups already exhibit a natural order, whereas this is not the case for ethnic-religious identities, we can sort the rows of the small multiple design (roughly) according to their average political support for school vouchers. The sorting is done informally by eye (Bertin, 1983), but of course more advanced visualization methods could rely on a sorting algorithm to achieve the most efficient plot arrangement (e.g. Hurley, 2004). After sorting, a regional pattern for the policy preferences of the black population pops out. Blacks oppose school vouchers in the South and support them in other regions of the United States.

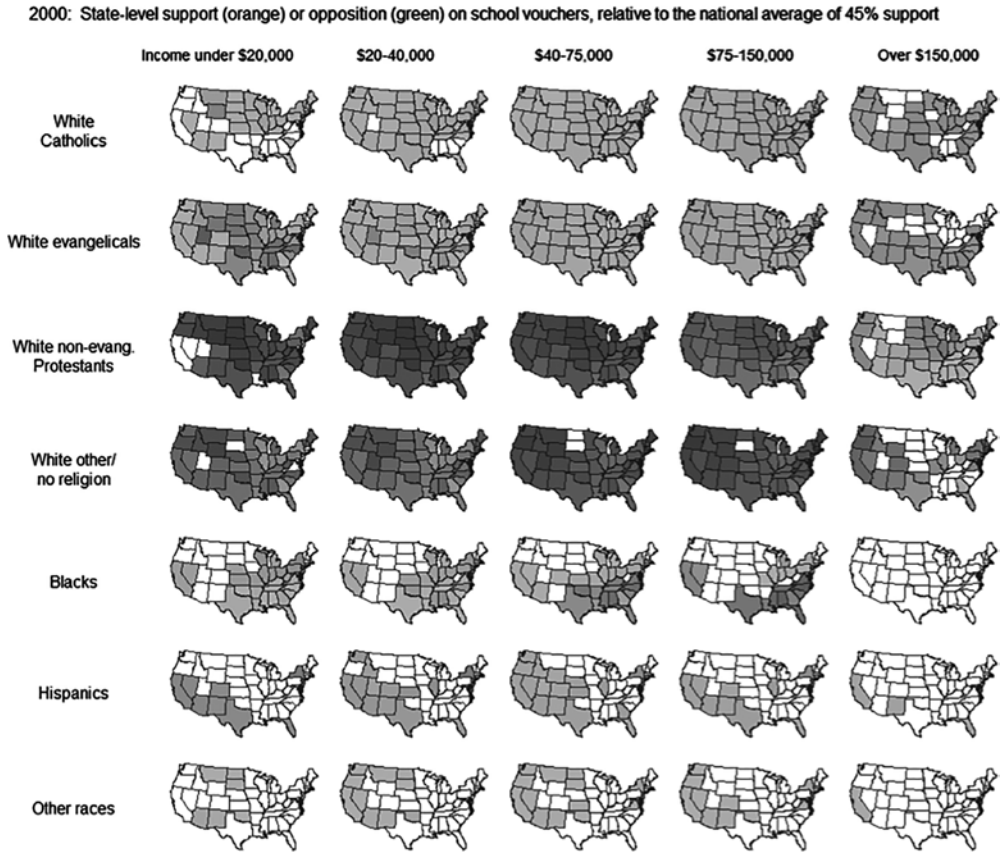
## RECENT ADVANCES IN DATA VISUALIZATION

### *Exploratory Model Analysis*

Exploratory data analysis (EDA) relies primarily on the visual display of data with the goal of discovering unknown structures and unexpected patterns in the data (Tukey, 1977). In a similar vein, visualization can be used to explore the unknown structures and unexpected implications of statistical models.



**Figure 25.11** Small multiples of support for school vouchers by geography and income (Gelman, 2009)



**Figure 25.12** Small multiples of support for school vouchers by geography, income and ethno-religious group (Gelman, 2009)

Exploratory model analysis (EMA) refers to the application of the methods and ideas of exploratory data analysis to statistical models (Gelman, 2003, 2004; Kerman et al., 2008; Unwin et al., 2003; Wickham, 2007; Wickham et al., 2015). Wickham et al. (2015) distinguish a total of five levels at which statistical models can be visualized, namely the model level ( $M$  measures of model fit), the model-estimate level ( $M \times K$  estimated coefficients, standard errors and t-values), the estimate level (summary of  $K$  estimates over many models), the model-observation level ( $M \times N$  residuals and influence measures) and finally the observation level ( $N$  original data and summaries of residual behavior).

The following example draws its motivation for using EMA from the requirements of robustness analysis, where we are interested in learning about parameter stability in many and potentially a huge number of different model specifications (Neumeier and Plümper, 2017). To illustrate, we re-analyze an influential sensitivity analysis conducted by Hegre and Sambanis (2006) on the determinants of civil war onset. In particular, we will look at all model specifications resulting from all possible combinations of 18 potential explanatory variables. The complete model space consists of just over a quarter million ( $2^{18}-1 = 262,143$ ) model specifications, excluding the empty intercept only model.

2000: State-level support (orange) or opposition (green) on school vouchers, relative to the national average of 45% support



**Figure 25.13** Small multiples of support for school vouchers by geography, income and ethno-religious group (Gelman, 2009). Rows have been re-sorted by support

Figure 25.14 (upper left panel) shows a *plot ensemble* that combines different graphs of different aspects of the model specifications (cf. Unwin, 2015). This way, more information is revealed than in any single plot alone. The plot ensemble combines two plots: (a) a scatter plot of the standardized coefficient estimates of the variable ‘military personnel’ (milper) across all models versus the models’ fit measured in terms of decrease in deviance, and (b) a parallel coordinate plot showing the standardized coefficient estimates for all 18

explanatory variables across all models. As we can see, there seem to be three broad clusters of model specifications that produce quite different coefficient estimates of the relationship between army size and the propensity for civil war.

To get a better understanding of which model specifications produce these three distinct clusters, we can rely on methods of interactive data visualization such as *brushing and linking* (Becker and Cleveland, 1987; Cook and Swayne, 2007; Theus and Urbanek,

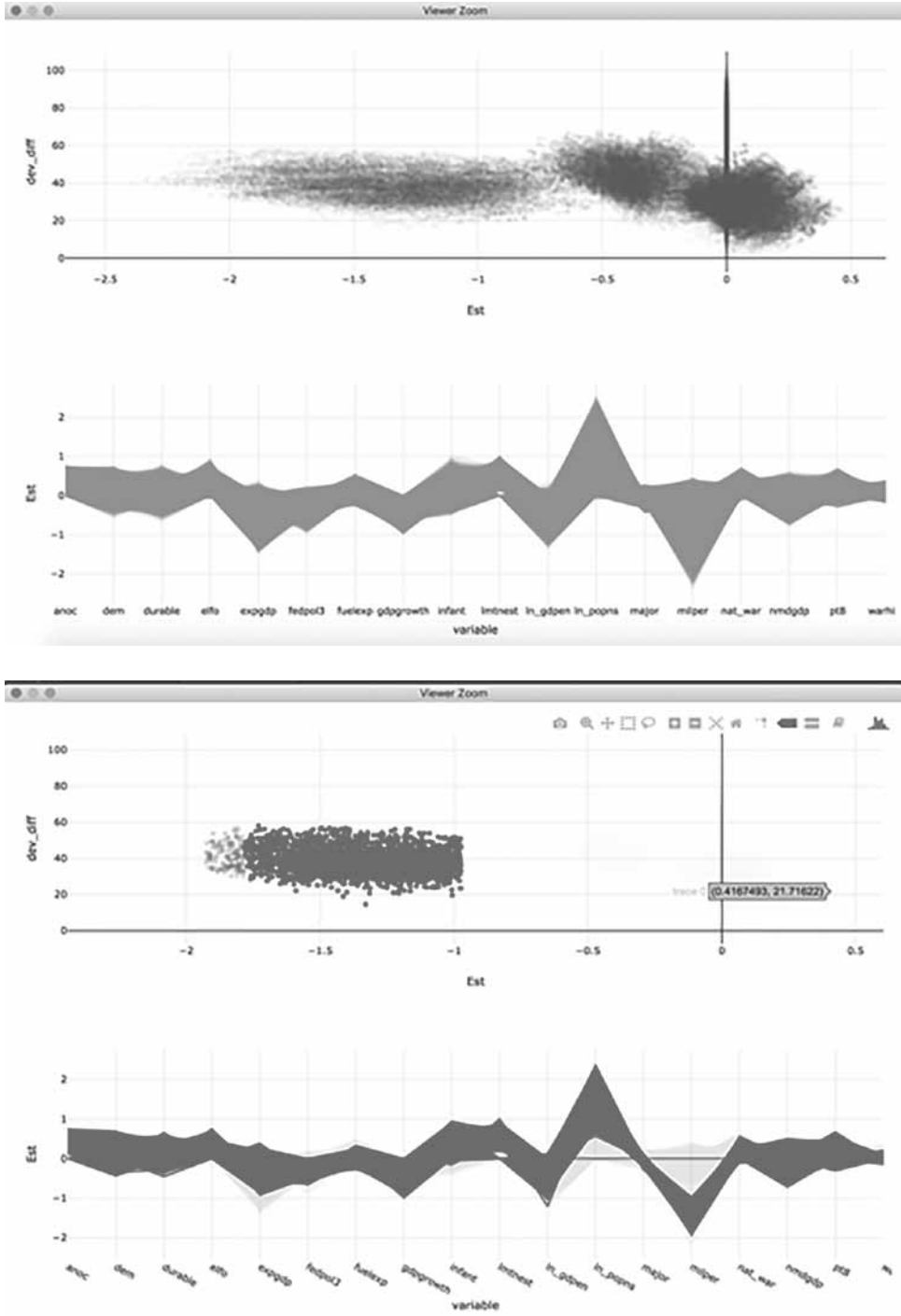


Figure 25.14 Plot ensemble for exploratory model analysis of the determinants of civil war onset. Scatter plots of model fit versus coefficient estimates for army size in over 200,000 models are interactively linked to parallel coordinate plots of coefficient estimates of all 18 explanatory variables

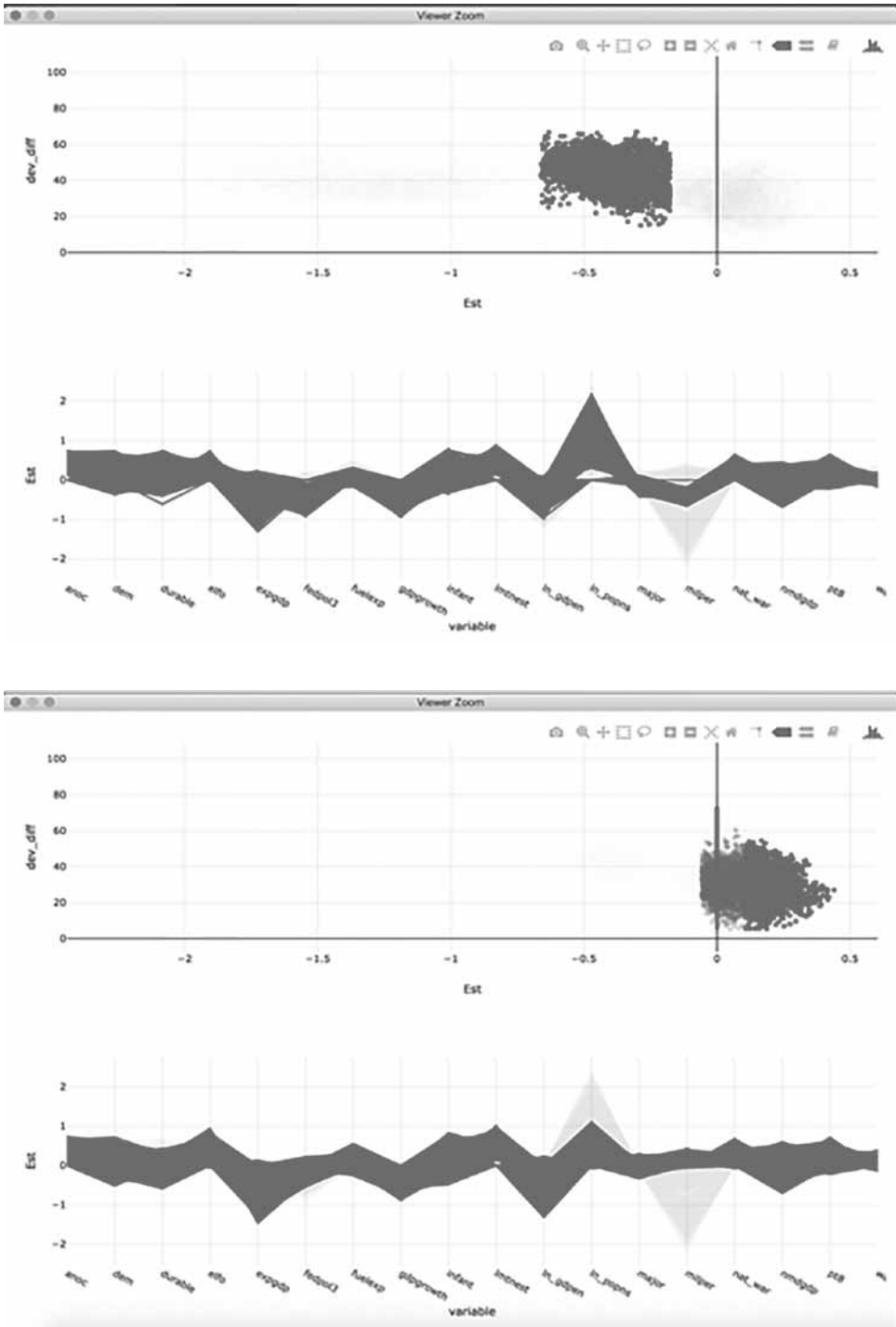


Figure 25.14 Continued

2009). This technique refers to the selection and color highlighting of a subset of the data in one plot. This subset is then simultaneously highlighted in another plot showing a different view of the same data. In this way a link between both views is established and the data is seen from different perspectives. In our EMA example, we find that the coefficient of army size is related to the behavior of log population. Models with large negative coefficients for military personnel tend to also have large positive effects of log population on civil war onset (upper right panel). Conversely, models with no or even positive effects of military size also tend to show no or smaller effects for log population. Clearly, the model exploration should now look more deeply into each of these specifications.

### **Visual Inference**

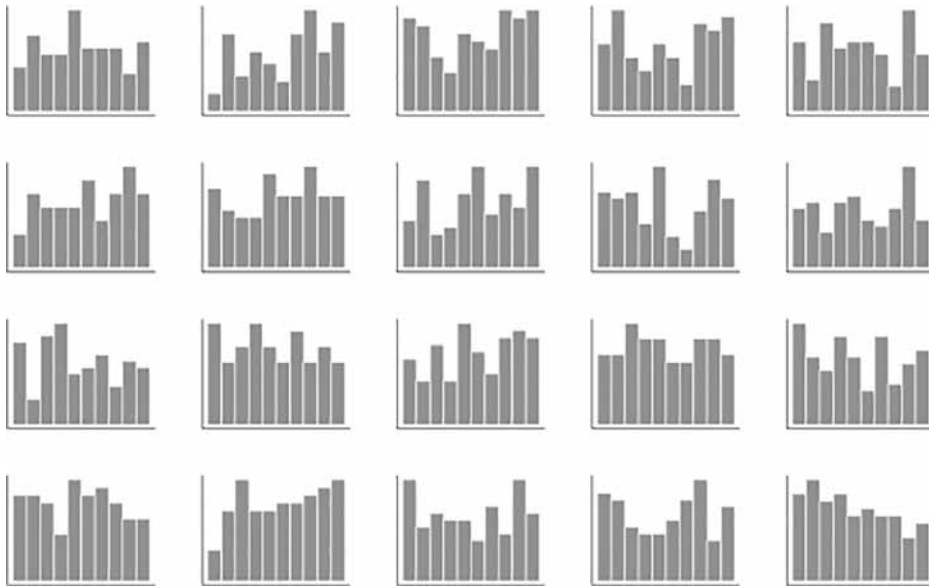
Despite the clear benefits of turning abstract data structures into visible patterns, a long-standing reservation against data visualization holds that it is merely an ‘informal’ approach to data analysis (cf. Best et al., 2001; Healy and Moody, 2014). The fear expressed in this view is that beautiful pictures may not correspond to any meaningful patterns of substantive scientific interest. Instead, it is argued, serious scientists should base their inferences on more ‘formal’ methods of hypothesis testing to discern signal from noise. Indeed, exploratory data analysis, according to one of its founders, ‘is about looking at data to see what it seems to say. It concentrates on [...] easy-to-draw pictures [...] Its concern is with appearance, not with confirmation’ (Tukey, 1977: v). Consequently, a criticism that frequently arises is that graphical displays lead researchers to over-interpret patterns that are in fact due to mere randomness.

One approach to overcome these reservations is *visual inference*, a new visual method that was only recently developed in statistics and information visualization (Buja et al., 2009; Wickham et al., 2010; Majumder et al.,

2013). The basic idea of visual inference is that graphical displays can be treated as ‘test statistics’ and compared to a ‘reference distribution’ of plots under the assumption of the null hypothesis. The null hypothesis usually posits that there is no systematic structure in the data and that any pattern is really the result of randomness. If the null hypothesis were indeed correct, the plot of the true observed data should not look any different from the plots showing random data. If, however, the plot of the true data clearly stands out from the rest, this could be taken as a rejection of the null hypothesis of no structure. In other words, visual inference brings the rigor of statistical testing to data visualization. To the best of my knowledge this approach has not yet been used in political science (although Bowers and Drake (2005) hint at it).

A so-called Line-up protocol involves the simulation of  $M-1$  null plots (for instance using variable permutations) and randomly placing the plot of the real observed data among them, resulting in a total of  $M$  plots. A human viewer is then asked to choose the plot that looks the most different from the rest. Ideally this human viewer is an impartial observer who has not yet seen the true plot, such as a colleague, student research assistant or crowd worker. If the test person succeeds and picks the plot showing the actual data, then this visual discovery can be assigned a  $p$ -value of  $1/M$ . In other words, the probability of picking the true plot just by chance is  $1/M$ . Setting  $M=20$  and thus simulating  $M-1=19$  null plots thus yields the conventional Type I error probability of  $\alpha = .05$ . We can further decrease the probability of making Type I errors by either increasing the number of null plots,  $M-1$ , or by increasing the number of observers,  $Q$ . Figure 25.15 gives an example of how this inferential process works. Try it for yourself: which of the 20 histograms stands out from the rest, and why?

How about the histogram in the last row and the last column? In fact, none of the histograms is the true plot showing actual data.



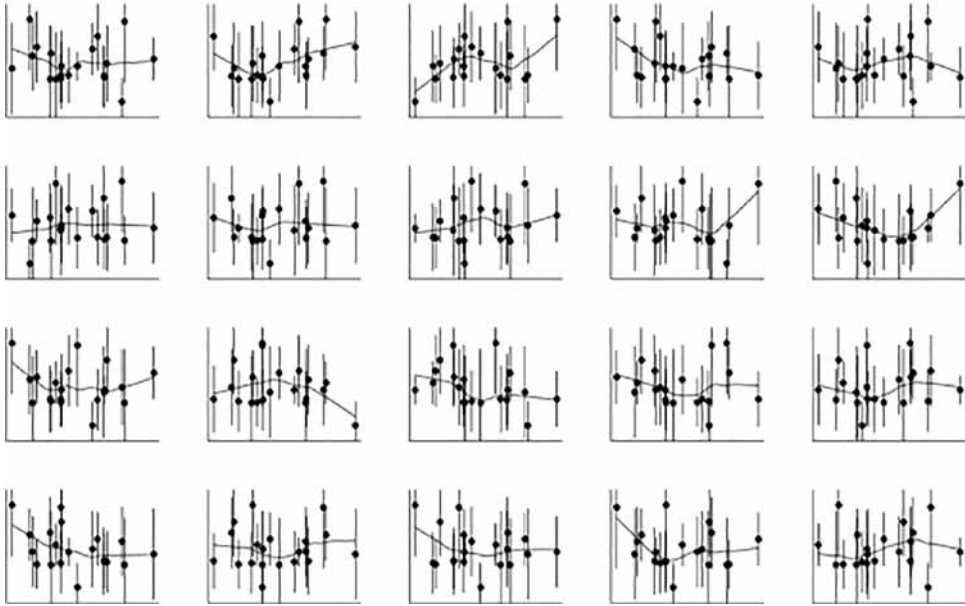
**Figure 25.15** Line-up visual inference with 20 histograms. Which plot is the most different?

All 20 histograms show 100 random draws from a uniform distribution  $U(0, 1)$ . Clearly, this demonstrates how easy it is to over-interpret patterns that are in fact due to mere randomness.

A real application follows Bowers and Drake (2005) and looks at the relation between education and political participation in the US and how this individual-level relation is conditioned by state-level educational context. A typical concern with this kind of analysis is that the number of contextual units is too small to rely on the asymptotic assumptions of classical statistical inference. Therefore, Bowers and Drake (2005) suggest visual methods instead of formal tests. Yet their visual inference remains informal: ‘when we detect a feature with our eyes, we will try to only report it as a feature rather than noise if we feel that any reasonable political scientist in our field would also detect this feature’ (Bowers and Drake, 2005: 317). Applying visual inference, we can swap assumptions concerning the reasonableness of political scientists for

a formal visual test. The null hypothesis in this example is that there is no relationship between the educational context in a state (i.e. the share of highly educated) and the effect of individual education on political participation. The ‘test statistic’ is a scatter plot version, where each dot is a state-specific individual-level effect of education on participation which is plotted along with vertical lines for the 95% confidence intervals. The size of this individual-level effect is on the y-axis. On the x-axis is the share of highly educated in the state. In addition, the plot includes a non-parametric scatter-plot smoother to help reveal any relation between state-level feature and individual-level effect. To construct a ‘reference distribution’ under the null, I randomly re-shuffle the state-level education variable and create 19 new data sets that will have no systematic relation between this variable and the coefficient by repeating this process 19 times. Figure 25.16 below shows the 19 null plots based on this simulated data along with the true plot. Which one stands out?





**Figure 25.16** Line-up for the relation between the individual education effect on political participation (y-axis) and state-level education (x-axis). Which plot is the most different?

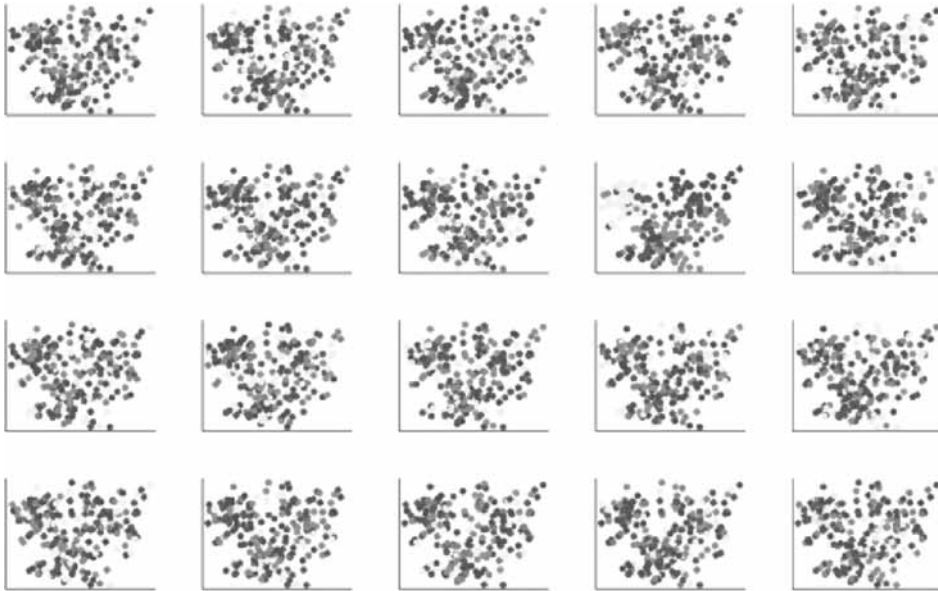
I asked nine political scientists and ten crowd workers, and not a single respondent in my sample managed to identify the plot showing the real data.<sup>4</sup> We clearly cannot reject the null hypothesis that individual educational effects are unrelated to state-level education.

Another example comes from political culture research and is inspired by the famous *World Values Survey Cultural Map*, which displays value orientations related to human development and democracy for a range of societies across the globe (see for instance Inglehart and Welzel, 2005). The ‘map’ is really a scatter plot that shows not geographic but cultural proximity by plotting countries along two value dimensions derived by factor analysis. The dimension of so-called survival versus self-expression values is plotted on the x-axis and the dimension of traditional versus secular–rational values on the y-axis. In addition, countries are colored according to their cultural zone

or civilizational heritage: African, Islamic, Latin American, South Asian, Protestant European, Catholic European, Orthodox and English-speaking.

One finding of theoretical interest suggested by the plot is that cultural zones form more or less distinct clusters with similar value orientations: culture matters. The question is whether this pattern is really systematic. The null hypothesis in this case would be that there are in fact no such civilizational clusters and that societies belonging to the same cultural zone do in fact not show similar survival versus self-expression and traditional versus secular–rational values. The reference null distribution can be constructed by a simple random permutation of the vector of cultural zones and thus the color of the dots in the scatter plot. Figure 25.17 presents 19 such null plots along with the true data plot. Can you pick the true cultural map?

The true plot clearly stands out.<sup>5</sup> Indeed, all of the political scientists and 90% of the



**Figure 25.17** Line-up for the relation between survival vs. self-expression values (x-axis) and traditional vs. secular-rational values (y-axis) clustered by cultural zone. Which plot is the most different?

crowd-sourced respondents correctly identified the observed cultural map, yielding a p-value of essentially zero. This allows us to reject the null hypothesis of no cultural value clusters around the world.

## CONCLUSION

Data visualization is an incredibly powerful method to explore, understand and communicate quantitative information. In times where political science accesses increasingly diverse and promising new data sources (e.g. text, social media and digital trace data), data visualization certainly holds a central place in the data analytic toolkit. In addition, communicating political science research to a broad, non-technical lay audience is an important skill. Data visualization is also likely to play a key role in this regard. Looking at its most important applications, key goals and central

actors, data visualization has always had a home in political science. It is hoped that the discipline will reconnect to this proud heritage and move forward to gauge the potential of data visualization for a better understanding of political processes.

## Notes

- 1 Many common visual methods for data exploration, such as histograms and box plots, are actually already abstractions from the data, due to binning decisions in the first case and the five number summary in the latter.
- 2 I thank Felix Jäger and Christian Moreau for excellent research assistance.
- 3 I thank Thomas Zittel for kindly sharing his data from the project 'Parliamentary candidates in the German states: socio-demographics, recruitment, attitudes and campaigning', funded by the German Research Foundation.
- 4 The true plot is in row three and column two.
- 5 The true cultural map is in row two and column four.

## REFERENCES

- Anscombe, Francis J. (1973). Graphs in statistical analysis. *American Statistician* 27(1): 17–21.
- Barabási, Albert-László. (2016). *Network science*. Cambridge: Cambridge University Press.
- Becker, Richard A., and Cleveland, William S. (1987). Brushing scatterplots. *Technometrics* 29(2): 127–142.
- Becker, R. A., Cleveland, W. S. and Shyu, M. J. (1996). The visual design and control of trellis display. *Journal of Computational and Graphical Statistics* 5(2): 123–155.
- Bertin, Jacques (1983). *Semiology of Graphics: Diagrams, Networks, Maps*. University of Wisconsin Press. [orig. in French, 1967.]
- Best, L. A., Smith, L. D. and Stubbs, D. A. (2001). Graph use in psychology and other sciences. *Behavioural Processes* 54(1–3): 155–165.
- Bowers, Jake (2004). Using R to keep it simple: exploring structure in multilevel datasets. *The Political Methodologist* 12(2): 17–24.
- Bowers, Jake and Drake, Katherine W. (2005). EDA for HLM: visualization when probabilistic inference fails. *Political Analysis* 13(4): 301–326.
- Buja, Andreas, Cook, Diane, Hofmann, Heike, Lawrence, Michael, Lee, Eun-Kyung, Swayne, Deborah F. and Wickham, Hadley (2009). Statistical inference for exploratory data analysis and model diagnostics. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* 367(1906): 4361–4383.
- Cleveland, William S. (1993). *Visualizing Data*. Murray Hill, NJ: Hobart Press.
- Cleveland, William S. (1994). *The Elements of Graphing Data* (Revised 2nd edition). Murray Hill, NJ: Hobart Press.
- Cleveland, William S. and McGill, Robert (1984). Graphical perception: theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association* 79(387): 531–554.
- Cook, Dianne and Swayne, Deborah F. (2007). *Interactive and Dynamic Graphics for Data Analysis with R and GGobi*. New York: Springer.
- Cook, Dianne, Lee, Eun-Kyung and Majumder, Mahbulul (2016). Data visualization and statistical graphics in big data analysis. *Annual Review of Statistics and Its Application* 3: 133–159.
- Few, Stephen (2009). *Now You See It: Simple Visualization Techniques for Quantitative Analysis*. Oakland: Analytics Press.
- Few, Stephen (2012). *Show Me the Numbers: Designing Tables and Graphs to Enlighten* (2nd edition). Oakland: Analytics Press.
- Friendly, Michael (2008). A brief history of data visualization. In Chun-houh Chen, Wolfgang Karl Härdle and Antony Unwin *Handbook of Data Visualization*. Berlin, Heidelberg: Springer, 15–56.
- Friendly, Michael (2009). *Milestones in the history of thematic cartography, statistical graphics, and data visualization*. Unpublished manuscript.
- Gelman, Andrew (2003). A Bayesian formulation of exploratory data analysis and goodness-of-fit testing. *International Statistical Review* 71(2): 369–382.
- Gelman, Andrew (2004). Exploratory data analysis for complex models. *Journal of Computational and Graphical Statistics* 13(4): 755–779.
- Gelman, Andrew (2009). *Hard Sell for Bayes*. Blogpost.
- Gelman, Andrew and Hill, Jennifer (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press.
- Gelman, Andrew and Unwin, Antony (2013). InfoVis and statistical graphics: different goals, different looks. *Journal of Computational and Graphical Statistics* 22(1): 2–28.
- Gill, Jeff (1999). The insignificance of null hypothesis significance testing. *Political Research Quarterly* 52(3): 647–674.
- Hawking, Stephen (1988). *A Brief History of Time*. London: Bantam.
- Healy, Kieran. (2018). *Data Visualization: A Practical Introduction*. Princeton: Princeton University Press.
- Healy, Kieran and Moody, James (2014). Data visualization in sociology. *Annual Review of Sociology* 40: 105–128.
- Heer, Jeffrey and Bostock, Michael (2010). Crowdsourcing graphical perception: using mechanical turk to assess visualization

- design. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 203–212.
- Hegre, Håvard and Nicholas Sambanis. (2006). Sensitivity analysis of empirical results on civil war onset. *Journal of Conflict Resolution* 50(4): 508–535.
- Hurley, Catherine B. (2004). Clustering visualizations of multidimensional data. *Journal of Computational and Graphical Statistics* 13(4): 788–806.
- Inglehart, Ronald and Welzel, Christian (2005). *Modernization, Cultural Change, and Democracy: The Human Development Sequence*. Cambridge: Cambridge University Press.
- Inselberg, Alfred (2008). Parallel coordinates: visualization, exploration and classification of high-dimensional data. In: Chen, C., Härdle, W. K. and Unwin, A. (eds), *Handbook of Data Visualization*. Berlin: Springer, 643–680.
- Jacoby, William G. (1997). *Statistical Graphics for Univariate and Bivariate Data*. Thousand Oaks, CA: Sage.
- Jacoby, William G. (2000). Loess: a nonparametric, graphical tool for depicting relationships between variables. *Electoral Studies* 19(4): 577–613.
- Jacoby, William G. and Schneider, Sandra K. (2010). *Graphical displays for political science journal articles*. Unpublished Manuscript. Michigan State University.
- Kastellec, Jonathan and Leoni, Eduardo (2007). Using graphs instead of tables in political science. *Perspectives on Politics* 5(4): 755–771.
- Keim, Daniel and Ward, Matthew (2003). Visualization. In: Berthold, M. and Hand, David J. (eds), *Intelligent Data Analysis. An Introduction* (2nd edition). New York: Springer, 403–428.
- Kerman, Jouni, Gelman, Andrew, Zheng, Tian and Ding, Yuejing (2008). Visualization in Bayesian data analysis. In: Chen, C., Härdle, W. K. and Unwin, A. (eds), *Handbook of Data Visualization*. Berlin: Springer, 709–724.
- King, Gary, Tomz, Michael and Wittenberg, Jason (2000). Making the most of statistical analyses: improving interpretation and presentation. *American Journal of Political Science* 44(2): 347–361.
- Kirk, Andy (2016). *Data Visualisation: A Handbook for Data Driven Design*. Thousand Oaks, CA: Sage.
- Matejka, Justin, and George Fitzmaurice (2017). Same stats, different graphs: generating datasets with varied appearance and identical statistics through simulated annealing. In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 1290–1294.
- Mitchell, Michael N. (2012). *A Visual Guide to Stata Graphics* (3rd edition). College Station, TX: Stata Press.
- Majumder, Mahbulul, Hofmann, Heike and Cook, Dianne (2013). Validation of visual statistical inference, applied to linear models. *Journal of the American Statistical Association* 108(503): 942–956.
- Murrell, Paul (2018). *R Graphics*. Boca Raton, FL: CRC Press.
- Neumayer, Eric and Plümper, Thomas (2017). *Robustness Tests of Quantitative Research*. Cambridge: Cambridge University Press.
- Playfair, William (1786). *The Commercial and Political Atlas; representing, by means of stained copper-plate charts, the exports, imports, and general trade of England, at a single view; the national debt; to which are added, charts of the revenue and debts of Ireland, done in the same manner by James Corry* (1st edition). London: J. Debrett; Robinson; and Sewell.
- Playfair, William (1801). *The Commercial and Political Atlas; representing, by means of stained copper-plate charts, the progress of the commerce, revenues, expenditure, and debts of England, during the whole of the eighteenth century* (3rd edition). London: T. Burton.
- Roberts, Margaret E., Stewart, Brandon M., Tingley, Dustin, Lucas, Christopher, Leder-Luis, Jetson, Gadarian, Shana Kushner, Albertson, Bethany and Rand, David G. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science* 58(4): 1064–1082.
- Taber, Charles S., and Lodge, Milton (2006). Motivated skepticism in the evaluation of political beliefs. *American Journal of Political Science* 50(3): 755–769.
- Talbot, Justin, Setlur, Vidya and Anand, Anushka (2014). Four experiments on the perception of bar charts. *IEEE Transactions*

- on *Visualization and Computer Graphics* 20(12): 2152–2160.
- Tennekes, Martijn, Jonge, Edwin de, and Daas, Piet J. H. (2013). Visualizing and inspecting large datasets with tableplots. *Journal of Data Science* 11(1): 43–58.
- Theus, Martin and Urbanek, Simon (2009). *Interactive Graphics for Data Analysis: Principles and Examples*. Boca Raton, FL: CRC Press.
- Tufte, Edward R. (2001). *The Visual Display of Quantitative Information* (2nd edition). Cheshire, CT: Graphics Press.
- Tufte, Edward R. (2006). *Beautiful Evidence*. Cheshire, CT: Graphics Press.
- Tukey, John W. (1977). *Exploratory Data Analysis*. Reading: Addison-Wesley.
- Unwin, Antony (2015). *Graphical Data Analysis with R*. Boca Raton, FL: CRC Press.
- Unwin, Antony, Volinsky, Chris and Winkler, Sylvia (2003). Parallel coordinates for exploratory modelling analysis. *Computational Statistics & Data Analysis* 43(4): 553–564.
- Unwin, Antony, Theus, Martin and Hofmann, Heike (2006). *Graphics of Large Datasets: Visualizing a Million*. Berlin: Springer Science & Business Media.
- Ware, Colin (1998). *Information Visualization: Perception for Design*. Elsevier.
- Ware, Colin (2013). *Information Visualization. Perception for Design* (3rd edition). Waltham, MA: Morgan Kaufmann.
- Wegman, Edward J. (1990). Hyperdimensional data analysis using parallel coordinates. *Journal of the American Statistical Association* 85(411): 664–675.
- Wickham, Hadley (2007). *Exploratory Model Analysis with R and GGobi*. Joint Statistical Meetings Proceedings.
- Wickham, Hadley (2010). A layered grammar of graphics. *Journal of Computational and Graphical Statistics* 19(1): 3–28.
- Wickham, Hadley, Cook, Dianne and Hoffmann, Heike (2015). Visualizing statistical models: removing the blindfold. *Statistical Analysis and Data Mining* 8(4): 203–225.
- Wickham, Hadley, Cook, Dianne, Hofmann, Heike and Buja, Andreas (2010). Graphical inference for InfoVis. *IEEE Transactions on Visualization and Computer Graphics* 16(6): 973–979.
- Wilkinson, Leland (2005). *The Grammar of Graphics* (2nd edition). Berlin: Springer.

# Text as Data: An Overview

Ken Benoit

## INTRODUCTION

When it comes to textual data, the fields of political science and international relations face a genuine embarrassment of riches. Never before has so much text been so readily available on such a wide variety of topics that concern our discipline. Legislative debates, party manifestos, committee transcripts, candidate and other political speeches, lobbying documents, court opinions, laws – not only are all recorded and published today, but in many cases this is in a readily available form that is easily converted into structured data for systematic analysis. Where in a previous era what political actors said or wrote provided insight for political observers to form opinions about their orientations or intentions, the structured record of the texts they leave behind now provides a far more comprehensive, complete and direct record of the implications of these otherwise unobservable states. It is no exaggeration, as Monroe and Schrodt (2009: 351) state, to

consider text as ‘the most pervasive – and certainly the most persistent – artifact of political behavior’. When processed into structured form, this textual record provides a rich source of data to fuel the study of politics. This revolution in the quantity and availability of textual data has vastly broadened the scope of questions that can be investigated empirically, as well as the range of political actors to which they can be applied.

Concurrent with textual data about politics becoming ubiquitous has been the explosion of methods for structuring and analysing this data. This wave has touched the shores of nearly all social sciences, but political science especially has been at the forefront of innovation in methodologies and applications of the analysis of text as data. This is most likely driven by a characteristic shared by some of the most important concepts in our discipline: they are fundamentally unobservable in any direct fashion, despite forming the foundation of our understanding of politics. Short of psychic powers or science fiction devices

for reading minds, we will never have access to direct, physical measures of the content or intensity of such core concepts as ideology, commitment to democracy or differing preferences or priorities for competing policies. Moreover, it is far from clear that every political actor even has such views or preferences, since many – such as political parties, coalition governments or nation-states – are hardly singular actors. Even singular actors may be unaware or self-deceptive about their intentions. Behaviour provides insights into these inner states, but what political actors *say*, more than the behaviour they exhibit, provides evidence of their true inner states.

For those facing the jungle of available textual data, navigating the thicket of different approaches and methodologies for making sense of this data can be no less challenging. Widely available computational tools combined with methods from machine learning allow unprecedented insight to be drawn from textual data, but understanding and selecting from these tools and methods can be daunting. Many recent authors have surveyed the range of methodologies and their applications (e.g. Wilkerson and Casas, 2017; Lucas et al., 2015; Slapin and Proksch, 2014; Grimmer and Stewart, 2013). Rather than retrace or duplicate the efforts of my expert colleagues, here I take a slightly different tack, focusing primarily on an overview of treating text ‘as data’ and then exploring the full implications of this approach. This involves clearly defining what it means to treat text as data, and contrasting this with other approaches to studying text. Comparing the analysis of text as data in the study of politics and international relations to the analysis of text as text, I place different approaches along a continuum of automation and compare the different research objectives that these methods serve. I outline stages of the analysis of text as data, and identify some of the practical challenges commonly faced at each stage. Looking ahead, I also identify some challenges that the field faces moving forward, and how we might meet them in

order to better turn the world of language in which we exist every day into structured, useful data from which we can draw insights and inferences for political science.

## TEXT, DATA AND ‘TEXT AS DATA’

### *Text as Text versus Text as Data*

Text has always formed the source material for political analysis, and even today students of politics often read political documents written thousands of years ago (Monroe and Schrodt, 2009). But for most of human history, the vast bulk of verbal communication in politics (as well as in every other domain) went unrecorded. It is only very recently that the cost of preserving texts has dropped to a level that makes it feasible to record them, or that a large amount of verbal activity has taken place on electronic platforms where the text is already encoded in a machine form that makes preserving it a simple matter of storage. Official documents such as the Congressional Record that transcribes what is said in the US legislature is now supplemented by records of email, diplomatic communications, news reports, blog posts, social media posts, public speeches and campaign documents, among others.

There is a long tradition of analysing texts to gain information about the actors who produced them (e.g. Berelson, 1952), dating from an era before computerised tools became available to facilitate even traditional methods of ‘content analysis’ (Krippendorff, 2013), defined as the human coding of texts into researcher-defined categories. In a different tradition, qualitative scholars may read critically into texts as discourses to uncover the patterns and connections of knowledge and power in the social structures that produced the texts (e.g. Foucault, 1972; Fairclough, 2001; see van Dijk, 1997 for an overview). Such data have always formed the empirical grist for the analytical mill of political

science, but only in the past two decades has the approach begun to shift when it comes to treating text as something not to be read, digested and summarised, but rather as inputs to more automated methods where the text is treated as data to be processed and analysed using the tools of quantitative analysis, even without necessarily being read at all.

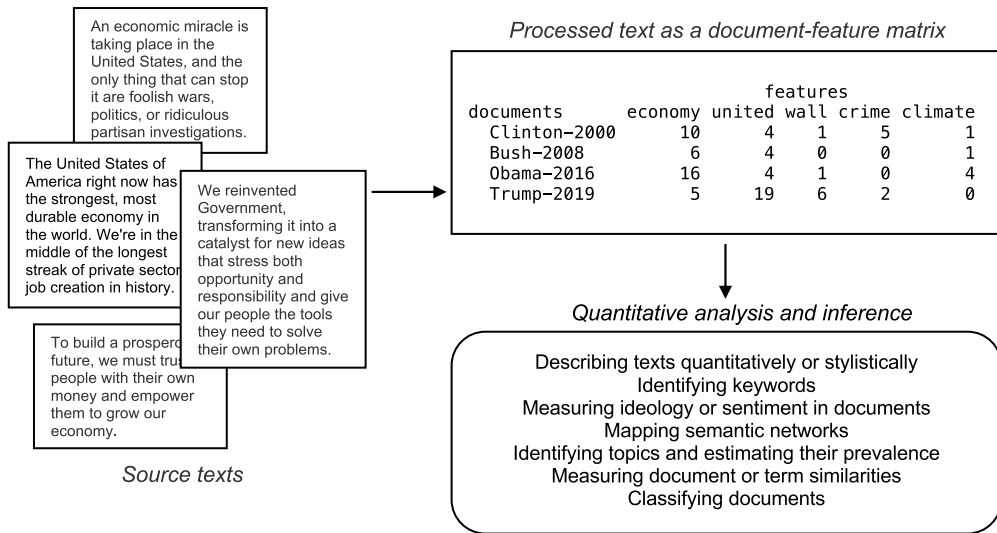
The very point of text is to communicate something, so in a sense, all forms of text contain information that could be treated as a form of *data*. Texts are therefore always informative in some way (even when we do not understand how). The primary objective of verbal activity, however, is not to record information, but to *communicate*: to transmit an idea, an instruction, a query, and so on. We can record it and treat it as data, but the purpose of formulating our ideas or thoughts into words and sentences is primarily communication, not the recording of our ideas or thoughts as a form of data. Most data is like this: the activity which it characterises is quite different from the data itself. In economics, for instance, it may be the economic transactions (exchanging goods or services using a medium of value) that we want to characterise, and the data is an abstraction of these transactions in some aggregated form that helps us to make sense of transactions using standardised measures. Through agreeing upon the relevant features to abstract, we can record and thus analyse human activities such as manufacturing, services or agriculture. The process of abstracting features of textual data from the acts of communication follows this same process, with one key difference: because raw text can speak to us directly through the language in which it is recorded, text does not first require processing or abstraction in order to be analysed. My argument here, however, is that this process of feature abstraction is the distinguishing ingredient of the approach to treating text as data, rather than analysing it directly as text.

Text is often referred to as ‘unstructured data’, because it is a (literally) literal recording of verbal activity, which is structured not

for the purposes of serving as any form of data but rather according to the rules of language. Because ‘data’ means, in its simplest form, information collected for use, text starts to become data when we record it for reference or analysis, and this process always involves imposing some abstraction or structure that exists outside the text itself. Absent the imposition of this structure, the text remains *informative* – we can read it and understand (in some form) what it *means* – but it does not provide a form of *information*. Just as with numerical data, we have to move from the act itself (speaking or writing) to a transformed and structured form of representing the act in order to turn the text into useful information. This is standard practice when it comes to other forms of data, but because we cannot read and understand raw numerical data in the same way that we can raw text, we have not yet fully equated the two processes. No one would hesitate to transform interval data such as age or income into ordinal categories of age or income ranges. (This improves accuracy at some cost of precision, as well as lessening the potential embarrassment of some survey respondents upon being asked to divulge how old they are or how little (or much) they earn.) The essence of treating text as data is that it is *always* transformed into more structured, summary and quantitative data to make it amenable to the familiar tools of data analysis.

Figure 26.1 portrays this process in three simple stages: raw texts; their processing and conversion into a quantitative form; and the analysis of this quantitative form using the tools of statistical analysis and inference. (I return in detail to the steps of this process below, but it is useful at this point to identify the essential stages of this process here.) Treating texts as data means arranging it for the purpose of analysis, using a structure that probably was not part of the process that generated the data itself. This step starts with collecting it into a *corpus*, which involves defining a sample of the available texts, out of all other possible texts that might have





**Figure 26.1** From text to data to data analysis

been selected. Just as with any other research, the principles of research design govern how to choose this sample, and should be guided by the research question. What distinguishes text from *textual data*, however, is that it has been selected for a research question to begin with, rather than simply representing a more fundamental act of communication by its producer. Once selected, we then impose substantial selection and abstraction in the form of converting the selected texts into a more structured form of data. The most common form in quantitative approaches to text as data is to extract *features* in the form of selected terms and tabulate their counts by documents: the 'document-feature matrix' depicted in Figure 26.1. This matrix form of textual data can then be used as input into a variety of analytical methods for describing the texts, measuring or mapping the targets of interest about which they contain observable implications or classifying them into politically interesting categories.

Quantitative text analysis thus moves textual data into the same domain as other types of quantitative data analysis, making it possible to bring to bear well-tested statistical and machine learning tools of analysis and prediction. By converting texts into a matrix

format, we unlock a vast arsenal of methods from statistical analysis designed for analysing matrix-type data: the comparison of distributions; scaling and measurement models; dimensional reduction techniques and other forms of multivariate analysis; regression analysis; and machine learning for prediction or identifying patterns. Many of these approaches, furthermore, are associated with well-understood properties that can be used for generating precise probability statements, such as the likelihood that an observed sample was generated from an assumed distribution. This allows us to generate insights from text analysis with precise confidence estimates, on a scale not otherwise possible.

Ironically, generating insight from text as data is only possible once we have destroyed our ability to make sense of the texts directly. To make it useful as *data*, we had to obliterate the structure of the original text and turn its stylised and oversimplified features into a glorified spreadsheet that no reader can interpret directly, no matter how expert in linear algebra. No similar lament is issued when processing non-textual data, because the form in which it can be recorded as data in the first place is already a highly stylised version of the

phenomena it represents. Such data began as a numerical table that we could not interpret directly, rather than as a direct and meaningful transcription of the act it recorded, whether it consisted of demographical data, (numerical) survey responses, conflict data, roll call votes or financial indicators. Quantitative analysis is the starting point of making sense of non-verbal data, and perhaps for these reasons has never proven controversial. With text, on the other hand, we often question what is lost in the process of extracting stylised features for the purpose of statistical analysis or machine learning, because we have a reasonable sense of what is lost in the meat grinder that turned our beautiful language into an ugly numerical matrix.

This point is so important that it warrants repeating. We hardly find it strange to be unable to make sense globally of a matrix of economic indicators, which we also recognise are imperfect and incomplete representations of the economic world involving the arbitrary selection of features from this world – such as the official definition of a basket of typical goods whose prices are used for measuring inflation. There is no controversy in acknowledging that while we might be able to interpret a specific figure in one cell of dataset by matching a column called inflation and a row with other columns whose values match ‘Canada’ and ‘1973q3’, to make sense of more general trends we need analytical synthesis using machines. With text, on the other hand, we cannot ignore the semantic violence to our raw material and its consequences of processing our raw text into textual data, with the necessarily imperfect and incomplete representation of the source language that this requires. Machines are stupid, yet treating text as data means letting stupid machines process and perhaps analyse our texts. Any human reader would know right away that *terror* has nothing to do with political violence in sentences such as ‘ending inflation means freeing all Americans from the *terror* of runaway living costs’.<sup>1</sup> We can only hope that our process

of abstraction into textual features is smart enough not to confuse the two concepts, since once our texts have become a document-feature matrix as portrayed in Figure 26.1, it will be hardly more interpretable than a set of raw inflation figures. In our discussion of the choice of appropriate procedure for analysing textual data, we return to this concern in more detail. The key point is that in order to treat text as data rather than text as text, we must destroy the immediate interpretability of source texts, but for the higher purpose of enabling more systematic, larger-scale inference from their stylised features. We should recognise this process unflinchingly, but also not lose any sleep over it, because the point in analysing text as data was never to interpret the data but rather to mine it for patterns. Mining is a destructive process – just ask any mountain – and some destruction is inevitable in order to extract its valuable resources.

### ***Latent versus Manifest Characteristics from Textual Data***

In political science, we are often most interested not in the text itself, but rather in what it tells us about a more fundamental, *latent* property of the text’s creator. In the study of politics (as well as psychology), some of our important theories about political and social actors concern qualities that are unobservable through direct means. Ideology, for instance, is fundamental to the study of political competition and political preferences, but we have no direct measurement instrument for recording an individual or party’s relative preference for (for example) socially and morally liberal policies versus conservative ones. Other preferences could include being relatively for or against a specific policy, such as the repeal of the Corn Laws in Britain in 1846 (Schonhardt-Bailey, 2003); being for or against further European integration during the debate over the Laeken Convention (Benoit et al., 2005); or being for or against a no confidence motion (Laver and

Benoit, 2002). These preferences exist as inner states of political actors, whether these actors are legislators, parties, delegates or candidates, and hence cannot be directly observed. Non-verbal indicators of behaviour could also be used for inference on these quantities, but it has been shown that what political actors *say* is more sincere than other forms of behaviour, such as voting in a legislature that is subject to party discipline and may be highly strategic (Herzog and Benoit, 2015). Textual data thus may contain important information about orientations and beliefs for which non-verbal forms of behaviour may serve as poor indicators. The field of psychology has also long used verbal behaviour as an observable implication of underlying states of interest, such as personality traits (e.g. Tausczik and Pennebaker, 2010). Absent coercive methods or mind-reading technology to discern the preferences, beliefs, intentions, biases or personalities of political and social actors, the next best alternative is to collect and analyse data based on what they are saying or writing. The target of concern is not so much what the text contains, but what its contents reveal as data about the latent characteristics for which the data serves as an observable implication.

Textual data might also focus on *manifest* characteristics whose significance lies primarily in how they were communicated in the text. Much of the field of political communication, for instance, is concerned not with the latent characteristics indicated by the texts but rather with the form and nature of the communication contained in the text itself. To take a classic example, in a well-known study of articles by other Politburo members about Stalin on the occasion of his 70th birthday, Leites et al. (1951) were able to measure differences in groups with regard to communist ideology. In this political episode, the messages signalled not only an underlying orientation but also a degree of political manoeuvring with regard to a leadership struggle following the foreseeable

event of Stalin's death. The messages themselves are significant, and these could only be gleaned from the public articles authored by each Politburo member, written in the full knowledge that they would be reprinted in the party and general Soviet press and interpreted as signals by other regime actors. To take another example, if we were interested in whether a political speaker used populist or racist language, this language would be manifest directly in the text itself in the form of populist or racist terms or references, and what would matter is whether they were used, not so much what they might represent. In their study of the party political broadcasts of Belgian political parties, for instance, Jagers and Walgrave (2007) established how much more overtly populist the language used by the extreme-right Vlaams Blok party was, compared to that of other Belgian parties.

In practice, the quality of a characteristic observable from text as being manifest versus latent is not always sharply differentiated. Stylistic features, for instance, might be measured as manifest quantities from the text but might be of interest for what they tell us about the author's more fundamental traits that led to the features' use in communication. In studies using adaptations of *readability* measures applied to political texts, for instance, we might be interested either in the latent level of political sophistication as a measure of speaker intention or in speaker characteristics, as evidenced by the observed sample of texts; alternatively, we might be interested in the manifest differences in their readability levels as more direct indicators of the medium of communication. In a study of historical speeches made in the British parliament, for instance, Spirling (2016) attributes a shift to simpler language in the late 19th century to the democratising effects of extending the franchise. Using similar measures, Benoit et al. (2019) compared a sample of US presidential State of the Union addresses delivered on the same day, by the same president, but in both spoken and written forms to show that the spoken forms used

easier language. The former study might be interested in language easiness as an indicator of a more latent characteristic about the political representation, while the latter analysis might be more focused on the manifest consequences of the medium of delivery. For many research designs using textual data, the distinction is more a question of the research objective than of some intrinsic way that the textual data is structured and analysed.

### **What 'Text as Data' Is Not**

We defined 'textual data' as text that has undergone selection and refinement for the purpose of more analysis, and distinguished latent from manifest characteristics of the text as the qualities about which the textual data might provide inference. While this definition is quite broad, it excludes many other forms of text analysis. It is useful, then, to identify the types of textual analysis that we do not consider as involving the analysis of text as data.

In essence: the study of text that does not extract elements of the text into a systematic form – into *data* – is not treating the text as data. Interpretivist approaches that focus on what a text *means* are treating the text as content to be evaluated directly, not as source material for systematic abstractions that will be used for analysis, only following which will its significance be evaluated. This is true even when the object of concern may ultimately be far greater than the text itself, such as in critical discourse analysis, whose practitioners' concern with text is primarily with social power and its abuse, or dominance and inequality as they are sustained or undermined by the text (van Dijk, 1994: 435). While this approach shifts attention to the texts as evidence of systemic injustices, the concern is more about the ability to construct a narrative of evidence for these systemic biases to be interpreted directly, rather than about extracting features from the text as data that will then be used in some analytic

procedure to produce evidence for or against the existence of injustices. The difference is subtle, but has to do with whether the interpretation of a direct reading of the text (no matter how systematic) is the end result of inquiry, versus an analysis only of extracted features of the text using a procedure that does not involve direct interpretation (such as reading) of those features. The latter treats the text as data, while the former is more focused on the text as text, to be interpreted and analysed as text.

Treating text as data is not strictly limited to quantitative approaches. Some of the most popular methods for analysing text as data in fact rely on qualitative strategies for extracting textual features. Classical *content analysis*, for instance, requires reading and understanding the text. The purpose of this qualitative strategy, however, is to use content analysis to extract features from textual data, not for analysing directly what is read and understood. In reading units of the text and annotating them with pre-defined labels, content analysis uses human judgement not to make sense of it directly, but instead only to apply a scheme to convert the text into data by recording category labels or ratings for each unit of text. Any analysis then operates on this data, and this analysis is typically quantitative in nature even if this only involves counting frequencies of keywords or category labels. But there is nothing to say that the process of extracting the features of the text into data needs to be either automated or statistical, and in thematic and content analytic approaches, they are neither. Most direct analysis of the text without systematically extracting its features as data – text as text – is by contrast almost always qualitative because raw text is inherently qualitative. This explains why text as data approaches are associated with quantitative analysis and interpretative approaches with qualitative analysis, but to equate them would be to obscure important qualitative elements that may exist as part of a text as data research design.

Many direct forms of analysing text as text exist. The analysis of political *rhetoric*, for

instance, can be characterised as the science and art of persuasive language use through ‘effective or efficient speaking and writing in public’ (Reisigl, 2008: 96). It involves a form of discourse analysis of the text, especially with respect to the use of tropes, symbols, allegories, metaphors and allusions. The study of anaphora in Martin Luther King’s ‘I Have a Dream’ speech (the repetition of ‘Now is the time...’ at the beginning of sentences), for instance, involves analysing the form of its language directly, not abstracting it into data that will only then be analysed. When elements of the speech are extracted systematically into features, however, and these features are subject to an analytic procedure whose interpretation can be used as an *indicator* of rhetorical quality, then the same input text *has* been treated as data.<sup>2</sup> This involves an act of literary brutality – the disassembly and matrix decomposition of one of the most moving speeches in US political history – but it allows us to compare Martin Luther King’s speech to other pieces of political oratory on a large scale and on a common methodological footing, in a way that would have been infeasible through direct interpretation.<sup>3</sup>

Finally, it is worth mentioning how the rapidly expanding field of natural language processing (NLP) from computer science fits within the boundaries of the text as data definition. Most computer scientists are puzzled by our use of the label, as if treating text as a form of data using quantitative tools were something new or special. This is because computer scientists’ approaches *always* involve some form of automated extraction of textual features and the processing or analysis of these using algorithmic and mathematical methods. The difference between many applications in NLP and the uses of textual data in political science lies not in whether the text is treated as data, but rather in the purposes for which this data is used. Computer scientists are frequently concerned with engineering challenges, such as categorising structure and syntax in language, classifying or summarising documents, mapping

semantic spaces, machine transition, speech recognition, voiceprint authentication, and so on. All of these are driven by textual data, but for objectives very different from the political scientist’s goal of making inferences about politics. Much research in NLP concerns the use of (big) textual data to make inference about patterns in natural language. Text for political scientists, by contrast, is just one more type of informative behaviour about politics, not something whose innate properties interest us in their own right. The main advantage and objective of analysing text as data in political science is to make inferences about the same phenomena that we have long studied using non-textual data.

The key dividing line, then, involves whether the analytic procedure – whether this is interpretation, critical discourse analysis, rhetorical analysis, frequency analysis or statistical analysis – is applied to directly to the text, or whether some intermediate step is applied to the text to extract its salient features which only then are analysed for insight. Within this broad definition, there are many forms this can take, and in what follows I contrast these along a continuum of automation and research objective, or what I call the target of concern (Table 26.1).

**Table 26.1 A map of approaches to the analysis of political text**

<i>Approach</i>	<i>Method</i>	<i>Target of concern</i>
Literary	Discourse analysis	Meaning Rhetoric Power relations Hidden biases Rhetoric Symbolism
Qualitative	Thematic analysis Content analysis	Topics Positions
Hybrid quantitative	Dictionary analysis	Affect
Purely quantitative	Statistical summary Machine learning	Authorship Intent Similarity Events

## VARIETIES OF TEXT ANALYSIS

We can distinguish three main variants of text analysis, differing in whether they treat the text as information to be analysed directly versus whether they treat the text as a source of data to be systematically extracted and analysed. Their situation along the continuum of text as text versus text as data can be contrasted on the basis of the degree of automation and on the target of concern to which the analysis is applied. This is a vast oversimplification, of course, but serves to contrast the essential differences between approaches.

### *Literary Analysis*

The first area is the one furthest from the approach of treating text ‘as data’ described here: literary analysis. This approach is aimed not at treating the text as an observable implication of some underlying target of interest, but rather as the target of interest itself: text as text. In extreme forms, this may treat the text as the sole object of interest, holding the characteristics or intention of the author of the text as irrelevant. This was the view of the ‘New Criticism’ school of literature theory advanced by Wimsatt and Beardsley (1946) in their influential essay ‘The Intentional Fallacy’, which argued against reading into author intentions or experiences, and advocated instead focusing exclusively on the text itself.

A great many other schools of literary thought exist, of course, including postmodernist approaches that do just the opposite of avoiding reading beyond the texts, and instead examine them critically as situated in their social context. What do the texts reveal about the structures of power in a social system, especially in relation to marginalised individuals or groups? Critical discourse analysis is concerned less (or not at all) with description of the text or inference from data extracted from the text, but rather with features underlying the system in which

the text occurred, even if its analysis takes place through analysing the text as text. A critical discourse of presidents’ speeches, for example, could focus on its commands and threats and how these are aimed at ‘managing the minds of others through a manipulation of their beliefs’ (van Dijk, 1994: 435; for an example in this context see Chilton, 2017). Treating presidential speeches as data, by contrast, could consist of a computerised analysis of the words used to contrast sentiment across time or to compare different individuals (e.g. Liu and Lei, 2018). The former is interested in the text as evidence for a philosophical and normative critique of power, while the latter is concerned with supplying more empirical data on the ability to describe and compare the preferences or styles of political actors in the context of open-ended scientific propositions. Discourse analysis may be very systematic, and indeed this was a key contribution of Fairclough (2001), who developed a sophisticated methodology for mapping three distinct dimensions of discourse onto one another. The key point here is with respect to the role of the text in the analysis, whether it forms the end object of inquiry as a text versus whether it will be used as a source of data, with the text itself of secondary or instrumental value.

### *Qualitative Text Analysis*

What I have labelled as *qualitative* approaches to the analysis of political text are distinguished from discourse analysis by focusing not on what the texts mean, either about the authors or about their attempts to influence the audience or to shore up or wear down the structures of the social system, but instead on gaining more neutral empirical data from the texts by using qualitative means to extract their features. ‘Qualitative’ is used here in its simplest form, to mean that the analytical tool does not involve statistical or numerical analysis and, at its core, involves human judgement and decision rather than machines.

These methods include *content analysis* and *thematic analysis*.

Sometimes called ‘qualitative content analysis’, content analysis is the human annotation of textual content based on reading the texts and assigning them categories from a pre-defined scheme. Many of the most widely cited comparative political datasets are generated from content analytic schemes of this type, such as the Manifesto Project (e.g. Budge et al., 1987, 2001) and the Comparative Policy Agendas Project (Walgrave and De Swert, 2007; Baumgartner et al., 2008). Both employ human coders to read a text, segment that text into sentences or phrases and apply fixed content codes to the segments using a pre-defined scheme that the coders have been trained to use.

Thematic analysis is essentially the same procedure, but involving a more iterative process whereby the annotation scheme can be refined during the process of reading and annotating the texts. These two approaches are closely related, since most content analytic schemes are developed by starting with a core idea and then are refined through a thematic process of attempting to apply it to a core set of texts. Thematic analysis resembles discourse analysis, and may even involve the same computer assisted tools for text annotation. It differs however in that both it and content analysis aim at a structured and more neutral and open-ended empirical approach to categorising, in the words of early political scientist Harold Lasswell (1948), who says what, to whom, and to what extent. Qualitative text analysis in this tradition aims not at a critique of discourse, but rather as ‘a research technique for the objective, systematic and quantitative description of the manifest content of communication’ (Berelson, 1952: 18).

Qualitative text analysis is labour intensive, but leverages our unique ability to understand raw textual data to provide the most *valid* means of generating textual data. Human judgement is the ultimate arbiter of the ‘validity’ of any research exercise, and if human judgement can be used to generate

data from text, we tend to trust this procedure more than we would the results of a machine – just as we would typically trust a bilingual human interpreter to render a correct translation more than we would Google Translate. This conclusion belies the unfortunate fact that humans are also notoriously unreliable, in the sense of not usually doing things in the exact same way when confronted with the same situation. (There are special psychological designations for those who do, including autism and obsessive-compulsiveness.) Two different human annotators, moreover, have naturally different perspectives, judgements, proclivities and experiences, and these invariably cause them to apply an analytic scheme in different ways. In tests to replicate the Manifesto Project’s scheme for annotating the sentences of manifestos, even among trained expert coders, Mikhaylov et al. (2012) found levels of inter-rater agreement and reliability so low that had the coders been oncologists, their levels of tumour misdiagnosis would have been professionally and financially catastrophic. Methods exist to increase coder reliability, such as formulating explicit rules and carefully training coders, but even these are imperfect. Machine methods, by contrast, may generate results that are invalid or systematically wrong (if poorly designed), but at least they will be perfectly reliably wrong. This allows valuable and scarce human effort to remain focused on testing and calibrating machine-driven methods, without the frustration of knowing that wrong answers might be due to random and uncontrollable factors.

### ***Hybrid Quantitative: Dictionary Analysis***

Dictionary analysis provides a very good example of a method in between qualitative content analysis and fully automated methods. The spread of computerised tools has made it possible to replace some or all of the analytic process, using machines that are perfectly reliable (but that don’t know Karl

Marx from Groucho Marx, much less what they are doing). One of the pioneering projects in what are known as *dictionary* approaches, the *General Inquirer* (Stone et al., 1966), arose in the late 1960s as an attempt to measure psychological qualities through texts as data, by counting words in electronic texts according to their membership in pre-defined psychological categories including positive and negative affect or ‘sentiment’. Because the field of psychology also has the problem that many of its most important concepts are inner states that defy direct measurement, psychology has also long been concerned with the use of language as observable implications of a speaker or author’s inner states, and some of the earliest and most ambitious dictionary-based projects have arisen in that field (e.g. also Martindale, 1975; Tausczik and Pennebaker, 2010).

In the ‘dictionary’ approach to analysing text as data, a canonical concept or label (the dictionary ‘entry’ or, in the terminology I prefer, *key*) is identified with a series of patterns to which words in a text will be matched. These patterns, which I will call *values*, are usually considered equivalent instances of the dictionary key. A key labelled *posemo* (for positive emotion) might contain the values *kind*, *kindly* and *kindn\**, for instance, to match references to the emotional characteristic of ‘having or showing a friendly, generous, and considerate nature’.<sup>4</sup> The last value (with the ‘\*’) is an example of a ‘glob’ pattern match, where the *wildcard* character will match any or no additional characters up to the end of the term – for instance, *kindness* or *kindnesses*. The false positives – words we detected but should not have – of *kindred* or *kindle* are excluded by these patterns, but so are *kindliness* and its variants – what we could call ‘false negatives’, or terms we should have detected but failed to do so.

This illustrates the key challenge with dictionary approaches: calibrating the matches to dictionary concepts in a valid fashion, using only crude fixed patterns as indicators of semantic content (meaning). The difficulty

lies in constructing a text analysis dictionary not only so that all relevant terms are matched (no false negatives), but also that any irrelevant or wrong terms are not (no false positives). The first problem is known as *specificity*, and is closely related to the machine learning performance measure known as *precision*. The second problem is known as *sensitivity*, and relates to the machine learning concept of *recall*. Match too broad a set of terms, using for instance the pattern *kind\**, and the matches attributed to positive emotion could wrongly include references to a popular electronic book reader. Match too specific a set of terms, such as *kind* only, and we would fail to match its adverbial form ‘kindly’.

Thus far we have focused on variants distinguished by spelling, but the problem can be even more fundamental because many words spelled identically may have completely different meanings. This quality is known as *polysemy*, and especially afflicts text as data approaches in English. To continue our example, *kind* may also be a noun meaning ‘a group of people or things having similar characteristics’, such as ‘more than one *kind* of text analysis’, or an adverb meaning ‘to some extent’, such as ‘dictionary calibration can get *kind* of tricky’. To illustrate, I used a part-of-speech tagger and some frequency analysis to distinguish the different meanings from the State of the Union corpus of presidential addresses. Of the 318 uses of *kind*, nearly 95% were the noun form while only 4% referred to the adjective denoting positive emotion (three more matches were to the ‘kind of’ usage). It is unlikely that human annotators would confuse the noun form with the adjective indicating a positive emotion, because their qualitative data processing instruments – their human brain, with its higher thoughts structured by language itself – would instantly recognise the difference. Human judgement is also inconsistent, however, and in some rare cases a qualitative annotator could misinterpret the word, might follow their instructions differently or might simply make a mistake. The computer, on the



other hand, while mechanistically matching all occurrences in a text of the term *kind* with the category of positive emotion, will produce 95% false positive matches by including the term's non-emotional noun form homograph, but do so with perfect consistency.

This discussion is not meant to disparage dictionary approaches, as they remain enormously popular and extremely useful, especially for characterising personality traits or analysing political sentiment. They also have the appeal of easy interpretability. While building the tools to efficiently count matches of dictionary values to words in a text might require some deft engineering, the basic idea is no more complicated than a counting exercise. Once counted, the analysis of these counts uses the same simple scales that are applied to content analysis counts, such as the percentage of positive minus negative terms. Conceptually, dictionary matches are essentially the same as human-coded content analysis, but in a cruder, more mechanistic way. Content analysis uses human judgement to apply a set of category labels to units of texts using human judgement after reading the text. Dictionary analysis replaces this with automated pattern matching to count category labels using automatic matching of the values defined as matches for those labels with words or phrases in the text. Both methods result in the construction of a matrix of texts by category counts, and from that point onward, the methods of analysis are identical. The target of concern of both approaches, as well as of the purely quantitative approaches discussed below, may be topics, positions, intentions or affective orientations, or even simple events, depending on the coding or dictionary scheme applied and the methods by which the quantitative matrix is scaled.

Dictionary methods are listed as 'hybrid' approaches because while they involve machines to match the dictionary patterns to the texts, constructing the set of matches in the dictionary is entirely a matter for human judgement. At some point, some human analyst made the judgement call to put *kind* as a

match for 'positive emotion' rather than (for instance) *kind*\*, but decided not to include (or simply overlooked) *altruistic* and *magnanimous*. Depending on the educational level of the texts to which this dictionary is applied, it will fail, to various degrees, to detect these more complicated, excluded synonyms. Many dictionaries exist that have been used successfully in many highly cited publications, but this is no guarantee that they will work for any untested application. In their attempt to use the venerable Harvard Psychosociological Dictionary to detect negative sentiment in the annual reports of public corporations, Loughran and McDonald (2011) for instance found that almost three-fourths of their matches to the unadjusted Harvard dictionary category of 2,010 negative words were typically not negative in a financial context: words such as *tax*, *cost*, *liability* and *vice*. Only through a careful, qualitative process of inspection of the word matches in context were they able to make significant changes to the dictionary in a way that fit their application, before trusting the validity of results after turning the machine loose to apply their dictionary to a corpus of 50,000 corporate reports.

### ***Purely Quantitative: Statistical Summaries***

Statistical summary methods are essentially quantitative summaries of texts to describe their characteristics on some indicator, and may use statistical methods based on sampling theory for comparison. The simplest identify the most frequently occurring words, and summarise these as frequency distributions. More sophisticated methods compare the differential occurrences of words across texts or partitions of a corpus, using statistical association measures, to identify the words that belong primarily to sub-groups such as the words associated with male- versus female-authored documents, or Democratic versus Republican speeches.

Measures of similarity and distance are also common in characterising the relationships between documents or terms. By treating each document as a vector of term occurrences – or, conversely, each feature as a vector of document occurrences – similarity and distance measures allow two documents (or features) to be compared using bivariate measures such as the widely used cosine similarity measure or Pearson’s correlation coefficient, or one of the many other distance measures such as the Euclidean or Jaccard distance. Such metrics form the backbone of the field of information retrieval but also allow comparisons between documents (and authors) that might have a more substantive political measure, such as ideological proximity. When generalised by comparing ‘local alignments’ of word sequences, similarity measures also form the basis of *text reuse* methods, which have been used to study the origins of legislation by Wilkerson et al. (2015) and the influence of interest groups by Hertel-Fernandez and Kashin (2015). Other quantitative summary measures are designed to characterise specific qualities of texts such as their *readability* – of which the Flesch (1948) reading ease measure is probably the best known – or *lexical diversity*, designed to measure vocabulary diversity across a text. While such indexes are not traditionally associated with stochastic distributions, it is possible to compute confidence intervals for these based on bootstrapping (Benoit et al., 2019) or averaging measures computed across moving windows of fixed text lengths (Covington and McFall, 2010), to judge statistically whether an observed difference between texts is significant.

### ***Purely Quantitative: Machine Learning***

#### ***Supervised machine learning***

In the final step along the continuum of automation versus human judgement, we have machine learning methods that require no

human analytical component, and are performed entirely by machine. Of course, human judgement is still required to select the texts for input or for training the machine, but this involves little more than a choice of which texts to input into the automated process. In purely quantitative approaches to text as data, there are choices about the selection and processing of inputs to be made, but not in the design of the instrument for processing or analysing the data in the way that dictionary approaches involve.

In purely quantitative approaches, it may not only be unnecessary to read the texts being analysed, but also be unnecessary for it to be *possible* to read them. Provided we have the means to segment the texts (usually, into words), then unsupervised approaches to scaling positions, identifying topics or clustering texts can happen without any knowledge of the language itself. Even supervised methods do not require the training texts to be read (although it is reassuring and preferable!), provided that we are confident that the texts chosen are good representatives of the extremes of the positions we would like to scale. For unsupervised scaling methods, no reading knowledge is required, *if we are confident that the texts are primarily about differences over well-defined issues*. For topic modelling, not even that is required. Of course, validation is crucial if we are to trust the results of automated methods, and this almost always involves human judgement and interpretation. Having skipped human judgement as part of the analytical process, in other words, we bring back our judgement at the conclusion of the process in order to make sense of the results. If our better judgement indicates that something is askance, we may choose to adjust the machine or its inputs and repeat the process until we get improved results. This cycle is often repeated several times, perhaps with different model parameters, for such tasks as classification, topic models (especially for choosing number of topics), document selection for unsupervised scaling or more fine-grained adjustment

such as feature selection. The choice of machine and its settings are important, but the ability to make sense of the words has become unimportant. This approach works with any language, because in stark contrast to the literary methods in which the meaning of words is the target of concern, 'it treats words simply as data rather than requiring any knowledge of their meaning as used in the text' (Laver et al., 2003: 312). In fully automated and quantitative approaches, the words are merely signals to help us interpret the political phenomena that gave rise to them, much as astronomers interpret minute variations in light wavelengths to measure the fundamental targets of concern that influence them, such as planetary sizes and orbits.

*Supervised machine learning* is based on the idea that a procedure will 'learn' from texts about which the analyst declares some external knowledge, and the results of this learning are then mapped onto texts for which the analyst lacks this knowledge. The objective is inference or prediction about the unknown texts, in the same domain as the input knowledge. Classifiers based on supervised examples start with a *training set* of texts with some known label, such as *positive* or *negative*, and learn from the patterns of word (feature) frequencies in the texts to associate orientations of each word. These orientations are used for projections onto a *test set* of documents whose label is unknown, based on some aggregation of the learned word feature orientations given the observed frequencies of the words in the unknown documents. While they perform this learning and prediction in different ways, this basic process is common to classifiers such as Naive Bayes (Pang et al., 2002), SVMs (Joachims, 1999), random forests (Fang and Zhan, 2015), neural networks (Lai et al., 2015) and regression-based models (e.g. Taddy, 2013).

When applied to estimating quantities on a continuously output scale rather than class prediction, supervised machine learning techniques may be adapted for *scaling* a dimension that is 'known' by virtue of the

training examples used to fit the model. This is the approach of the *Wordscores* model (Laver et al., 2003) that has been widely used in political science to scale ideology, as well as its more modern descendant, *class affinity* scaling (Perry and Benoit, 2018). Both methods learn word associations with two contrasting 'reference' classes and then combine these with word frequencies in texts whose positions are unknown, in order to estimate their positions with respect to the reference classes.

Supervised *scaling* differs from supervised *classification* in that scaling aims to estimate a position on a latent dimension, while classification aims to estimate a text's membership in a latent class. The two tasks differ in how greedily they demand input data in the form of more features and additional documents. Typically, classification tasks can be improved by adding more training data, and some methods, such as convolutional neural networks (Lai et al., 2015), require very large training sets. To minimise classification error, we may not care what features are used; as long as the model is not overfit, the primary goal is simply to correctly predict a class *label*. Scaling, on the other hand, is designed to isolate a specific dimension on which texts are to be compared and provide a point estimate of this quantity, on some continuous scale. Validating this quantity is much harder than in class prediction, and typically involves comparison to external measures to establish its validity.

Unlike classification tasks where accuracy is the core objective, supervised scaling approaches have been shown capable of producing valid and robust scale estimates even with relatively small training corpora (see Klemmensen et al., 2007; Baek et al., 2011). The key in scaling applications is more the quality of training texts – making sure they contain good textual representations of the opposing poles of a dimensional extreme (Laver et al., 2003: 330) – rather than their quantity. For scaling applications, training texts only need to contain strong examples

of lexical usage that will differentiate the dimensional extremes, such as strong ‘conservative’ language in one set contrasting with strong ‘liberal’ language in another, using the same lexicon that will be used in the out-of-sample (or, in the words of Laver et al. (2003), ‘virgin’) texts. One advantage of not being concerned with classification performance is that scaling is robust to irrelevant text in the virgin documents. Training texts that contain language about two extremes of environmental policy, for instance, are unlikely to contain words about health care. Scaling an unknown text using a model fitted to these environmental texts will therefore scale only the terms related to (and hence only the dimension of) environmental policy, even if the document being scaled contained out-of-domain text related to health care. For unsupervised methods, by contrast, irrelevant text will seriously affect unsupervised scaling approaches.

Most political scientists are interested more in measurement and scaling than in classification, which is typically of only instrumental value in estimating or augmenting a dataset for additional testing. In their study of echo chambers on the Twitter social media platform, for instance, Colleoni et al. (2014) used supervised learning trained on Tweets from around 10,000 users known to be Republican or Democrat, to predict the party affiliation of an additional 20 million users. They used the supervised classifier to augment their dataset of social media with a label of party affiliation, which is not part of the social media data but which was nonetheless central to their ability to measure partisan homophily in communication networks. Classification in social science is generally more useful in augmenting data rather than representing an interesting finding in its own right. While classifying a legislator’s party affiliation might be an interesting engineering challenge for a computer scientist, this typically yields no new insight for a political scientist, as this information is already known (which does not mean that it has not

been done, however: see Yu et al., 2008). Estimating the sincere political preference of a legislator whose vote is uninformative because of party discipline, by comparison, is typically of great interest in political science.

### *Unsupervised machine learning*

Unsupervised learning approaches are similar to supervised methods, with one key difference: there is no separate learning step associated with inputs in the form of known classes or policy extremes (if scaling). Instead, differences in textual features are used to infer characteristics of the texts, and these characteristics are interpreted in substantive terms based on their content or based on their correlation with external knowledge. A grouping might be labelled based on its association with different political party affiliations of the input documents, for instance, even though the party affiliations did not form part of the learning input. Examples of unsupervised methods associated with text are clustering applications, such as *k*-means clustering (see Grimmer and Stewart, 2013: §6.1), designed to produce a clusters of documents into *k* groups in a way that maximises the differences between groups and minimises the differences within them. These groups are not labelled, and so must be interpreted *ex post* based on a reading of their content or the association of the documents with some known external categories. Because this is primarily a utility device for learning groups, it has few applications in political science outside of a data augmentation tool, although it has been used as a topic discovery tool in some applications, such as in the case of Sanders et al. (2017), who used clustering as one method to identify economic policy topics from UK select committee oversight hearings.

An unsupervised learning method that has received wide application is the Latent Dirichlet allocation (LDA) *topic model* (Blei et al., 2003). Topic models provide a relatively simple, parametric model

describing the relationship between clusters of co-occurring words representing ‘topics’ and their relationship to documents which contain them in relative proportions. By estimating the parameters of this model, it is possible to recover these topics (and the words that they comprise) and to estimate the degree to which documents pertain to each topic. The estimated topics are unlabelled, so a human must assign these labels by interpreting the content of the words most highly associated with each topic, perhaps assisted by contextual information. No human input is required to fit the topics besides a document–feature matrix, with one critical exception: the number of topics must be decided in advance. In fitting and interpreting topic models, therefore, a core concern is choosing the ‘correct’ number of topics. There are statistical measures (such as *perplexity*, a measure based on comparing model likelihoods, or *topic coherence*, based on maximising the typical pairwise similarity of terms in a topic) but a better measure is often the interpretability of the topics. In practice the precise choice of topics contains a degree of arbitrariness, and often, to recover interpretable topics, some extra ones are also generated that are not readily interpretable.<sup>5</sup>

Political scientists have made widespread use of topic models and their variants, including some novel methodological innovations driven by the specific demands of political research problems. Quinn et al. (2010) shifts the *mixed membership* model of topics within documents to a time unit (days in the US Senate) and estimates the membership of texts with each time unit (speeches made on that day) as representing a single topic. Combined with some prior information, this model can produce estimates of the daily attention to distinct political topics, to track what the Senate is talking about over a long time series. Another variation is Grimmer’s (2010) *expressed agenda model*, which measures the attention paid to specific issues in senators’ press releases, based on the idea that each senator represents a mixture of

topics and will express these through individual press releases. Another innovation of which political scientists should be proud is the *structural topic model* (Roberts et al., 2014), which introduces the ability to add covariates in the form of categorical explanatory variables to explain topic prevalence. In their paper introducing this method, Roberts et al. (2014) apply it to open-ended survey responses on immigration questions to show differences in the estimated proportions of topics pertaining to fear of immigration, given the treatment effect of a survey experiment and conditioning variables related to whether a respondent identified with the Democratic or Republican party. In each of these innovations, political scientists have adapted a text mining method to specific uses enabling inference about differences between time periods, individuals or treatments, turning topics models from an exploratory tool into a method for testing systematic propositions that might relate to fundamental political characteristics of interest.

Another unsupervised method that is not only widely applied but also developed by political scientists is the unsupervised *Wordfish* scaling model (Slapin and Proksch, 2008). This model assumes that observed counts in a document–feature matrix are generated by a Poisson model combining a word effect with a parameter representing a position on a latent dimension, conditioned by both document and feature fixed effects. It produces estimates of a document’s latent position, which can be interpreted as left–right ideology (Slapin and Proksch, 2008), preference over environmental policy (Klüver, 2009) support or opposition to austerity in budgeting (Lowe and Benoit, 2013) or preferences for the level of European integration (Proksch and Slapin, 2010). One limitation of this model, however, is that it permits estimation on only a single dimension (although other dimensional estimates using similar methods are possible, as Monroe and Maeda (2004) and Däubler and Benoit (2018) have demonstrated). In a

detailed comparison of scaling model estimates to ratings of the same texts by human coders, Lowe and Benoit (2013) showed that an anti-system party appeared wrongly (according to the human raters) in the middle of the scale of support or opposition to the budget, because of its differences on a dimension of politics not captured in a single government–opposition divide. Because they are anchored according to extremes identified by the user, supervised scaling methods such as Wordscores can extract different positional estimates from the same texts (provided the training inputs for these texts were different). Unsupervised scaling, however, will always produce only one set of estimates for the same texts. When an analyst wants to estimate multiple dimensions, the only recourse is to input different texts. When Slapin and Proksch (2008) used the method to scale policy positions from German party manifestos on three separate dimensions of economic, social and foreign policy, they first had to segment each manifesto into new documents containing only text relating to these themes (which required reading the texts, in German, and then manually splitting them). To control the outputs from unsupervised methods, one must control the inputs.

Poisson scaling (e.g. the Wordfish method) is very similar to older methods to project document positions onto a low-dimensional space, after singular value decomposition (and some additional transformation) of the high-dimensional document–feature matrix. Such older methods include *correspondence analysis* (CA: Greenacre, 2017) and *latent semantic analysis* (LSA: Landauer et al., 1998), both forms of metric scaling that can be used to represent documents in multiple dimensions (although LSA is more commonly used as a tool in information retrieval). These lack some advantages of parametric approaches, such as the ability to estimate uncertainty using outputs from the estimation of statistical parameters, but have nonetheless seen some application in political science because of their ease of computation

and ability to scale multiple dimensions (e.g. Schonhardt-Bailey, 2008).

Because unsupervised scaling methods take a matrix as input, and this matrix might just as easily have been transposed (swapping documents for features), these methods also permit the measurement and scaling of word features as well as documents. The metric scaling from CA, for instance, allows words' locations to be identified in the same dimensional spaces as documents (see Schonhardt-Bailey, 2008, for instance). Wordfish scaling also allows us to estimate the policy weight and direction, similar to a discrimination parameter from an item-response theory model, for each feature. When features are policy categories, this can provide information of substantive interest in its own right, such as how different policies form the left–right 'super-dimension' and how these might differ across different political contexts (Däubler and Benoit, 2018).

### ***Distributional Semantic Models and 'Word Embeddings'***

A final exciting area deserving mention are text as data approaches based on matrices of observed words but weighted by their 'word vectors', estimated from fitting a *distributional semantic model* (DSM) to a large corpus of text, often a corpus separate to the text to be analysed as data in a given application. The notion of distributional semantics was famously articulated by the linguist John Firth, who stated that 'You shall know a word by the company it keeps' (Firth, 1957: 11). Using a 'continuous bag-of-words model' to estimate word co-occurrences within a specified context (for instance, a window of five words before and after), models can be fit to estimate a vector of real-valued scores for each word representing their locations in a multi-dimensional semantic space. Known collectively as *word embedding* models, such methods provide a way to connect words according to their usages in a way that offers

potentially vast improvements on the context-blind ‘bag-of-words’ approach.

Relatively new methods for fitting DSMs include the ‘word2vec’ model (Mikolov et al., 2013), which uses a ‘skip-gram’ neural network model to estimate the probability that a word is ‘close’ to another given word; the ‘GloVe’ (‘gloval vectors of words’: Pennington et al., 2014) model, which predicts surrounding words using a form of dynamic logistic regression; and the ELMo model (‘Embeddings from Language Models’: Peters et al., 2018). All of these methods are widely available in open-source software implementations.

Word embedding models are usually not thought of as methods on their own for analysing text as data, but rather as extremely useful complements to representations of text as data based on word counts. They have been shown to greatly improve performance for applications such as text classification, sentiment analysis, clustering or comparing documents based on their similarities or document summarization. Estimated from a user’s own corpus, furthermore, word embeddings allow the direct exploration of semantic relations in their own verbal context, to determine the associations of terms far more closely related to their meanings than possible using only simple clustering or similarity measures from bag-of-words count vectors.

For users that cannot fit local embedding models to a corpus, pre-trained word vectors are available that have been estimated from large corpora, such as that trained on six billion tokens from Wikipedia and the ‘Gigaword’ corpus (Pennington et al., 2014). This allows a user analysing text to represent his or her texts not just from the corpus at hand, but also augmented with quantitative measures of the words’ semantic representations fitted from other contexts. This provides an interesting twist on the discourse analytic notion of *intertextuality* (e.g. Fairclough, 1992), a process in which the meaning of one text shapes the meaning of another. Incorporating semantic representations fitted from large corpora into the analysis of text is

a literal recipe for reinforcing the pre-dominant social relations of power as expressed in language, a problem that has not gone unnoticed. Bolukbasi et al. (2016) and Caliskan, Bryson and Narayanan (2017) show that word embeddings encode societal stereotypes about gender roles and occupations, for instance that engineers tend to be men and that nurses are typically women. Data and quantification do not make our textual analyses neutral, and we should be especially aware of this when incorporating semantic context into text as data approaches.

## THE STAGES OF ANALYSING TEXT AS DATA

We have described the essence of the approach of treating text as data as involving the extraction and analysis of features from text to be treated as data, either about the manifest characteristics of the text itself or latent characteristics for which the text provides observable implications. In this section I describe this process in more detail, outlining the steps involved (Table 26.2) and the key choices and issues faced in each stage.

### *Selecting Texts: Defining the Corpus*

A ‘corpus’ is the term used in text analysis to refer to the set of documents to be analysed, and that have been selected for a specific

**Table 26.2** Stages in analyzing text as data

1. Selecting texts and defining the corpus.
2. Converting the texts into a common electronic format.
3. Defining documents and choosing the unit of analysis.
4. Defining and refining features.
5. Converting textual features into a quantitative matrix.
6. Analyzing the (matrix) data using an appropriate statistical procedure.
7. Interpreting and reporting the results.

purpose. Just as with any other research design, research built on textual data begins with the analyst identifying the corpus of texts relevant to the research question of interest and gathering these texts into a collection for analysis. Texts are generally distinguished from one another by attributes relating to the author or speaker of the text, perhaps also separated by time, topic or act. A year of articles about the economy from *The New York Times*, for instance, could form a sample for analysis, where the unit is an article. A set of debates during (one of the many) votes on Brexit in the UK House of Commons could form another corpus, where the unit is a speech act (one intervention by a speaker on the floor of parliament).

German-language party election manifestos from 1949 to 2017 could form a corpus, where a unit is a manifesto. A set of Supreme Court decisions from 2018 could form a corpus, where the unit is one opinion. In each example, distinguishing external attributes, chosen by the researcher for the purpose of analysing a specific research question, are used to define the *document* distinguishing one unit of textual data from another.

In many political science applications using textual data, the ‘sample’ of texts may, in fact, be every known text generated by the political universe for that application. In tracking the words spoken on abortion per day in the US Congress, for instance, a study might examine every spoken utterance in the Senate from 1997 to 2004. Yet even in such situations where a researcher may not face overt decisions to sample texts from a larger population that is too large to cover in its entirety, such as how many newspaper articles to select from which set of days, it is still important to be aware of selection issues that shape what sort of text becomes a recorded feature of the social system. Such ‘social bookkeeping’ has long been noted by historians seeking texts to gain leverage on events long past, but it may also feature in many forms of observed political text, especially spoken text in structured settings such

as legislatures. Historical coalition political manifestos, for instance, are notoriously difficult to obtain because they tend to disappear once a coalition has broken down, creating a potential sample bias slanted toward more stable coalitions. The key is to be aware of the mechanisms governing the generation of text, with the aim of making sure that the observable text provides representative coverage of the phenomenon that it will be used to investigate.

Some sampling choices may be motivated on practical grounds, especially resource limitations. In text as data approaches pre-dating the availability of computerised tools, it was not uncommon to suggest examining 100-word samples from a text for measuring such quantities as the readability of a text (e.g. Gunning, 1952) or taking ‘all the words in 16 two-page groups spread uniformly throughout the book’ for a measure of lexical diversity (Herdan, 1955: 332). In the modern era, by contrast, down-sampling may be required due to access limitations or because of the sheer volume of available data. The Twitter streaming API, for instance, has an overall rate limit of 1% of all Tweets for those without access to Twitter’s exclusive ‘firehose’ of all Tweets. Even researchers who have captured the tens of millions of daily Tweets available within this rate limit may decide to work on a random sub-sample of this dataset, because of the computational and storage costs involved in trying to analyse the larger dataset.

### ***Converting the Texts into a Common Electronic Format***

This step is purely technical, involving no research design decisions, but it can nonetheless pose one of the stickiest problems in text analysis. Strictly speaking, it is not necessary to work with computers to treat texts as data. The old-school methods for computing textual readability – for instance Gunning’s FOG index referenced above, or applying the



complex rules from Spache (1953) to match words in a text to a list of ‘familiar’ terms – could involve working with pen, printed text, an abacus and a lot of coffee (possibly while working by candlelight and wearing a hair-shirt). In almost all contemporary applications, however, texts are sourced and processed using computers. The problem is that there are vast differences in the formats for recording electronic texts, including Adobe’s ‘pdf’ format, which is actually a collection of different variants and versions; markup languages (such as HTML or XML); word processing formats (such as Microsoft Word, which also exists in many different versions); spreadsheet formats (such as Microsoft Excel); key-value schemes (such as JSON); or, if one is really lucky, plain text (.txt) files requiring no special handling. Even plain text files, however, can require a form of conversion, since the machine *encoding* of text has many different forms, especially in the pre-Unicode era from about 1970–2000 when the same set of 8-bit numeric values were mapped to different characters depending on the platform and the national context.<sup>6</sup> Unicode has replaced this, by providing a single, comprehensive mapping of unique code points to every known character in the world’s writing systems, present and past, including emoji and special symbols. Because Unicode is a standard, however, rather than an *encoding*, it still needs to be implemented on machines, and Unicode also covers standards for this encoding, such as UTF-8 (the most common).

Conversion of images into text is another possible headache, especially for older documents that may have been scanned. To convert these ‘image-only’ documents, which may exist in pdf form but not contain actual text, *optical character recognition* may be needed: the conversion of images of characters into electronically encoded text. Depending on the quality of the images, this can require a great deal of manual correction and cleaning. To the human eye, there may be no essential difference between *OCR* and

*OCR*, but to a computer these are completely different words. Other challenges can involve typographic ligatures (such as the ‘fi’ often used in such words as *find*) and other typographic relics such as the medial *s*, printed as *f*, which was disused by around 1800 but widely found in 19th-century printing. Most *OCR*, however, will not recognise that *Congrefs* is the same as *Congress*.<sup>7</sup>

### **Defining Documents and Choosing the Unit of Analysis**

This step is a refinement from the selection of the corpus in that prior to extracting textual features for analysis, the unit of analysis may need further definition, through selection or sampling or through aggregating documents into larger units or splitting them into smaller ones. The attributes that differentiate source texts, in other words, may not form the ideal units for analysing the text as data. (Note that by ‘units’ here we refer to the document units, not textual features, which are covered next.)

For example, while we might have a corpus of social media posts, these might be better aggregated over some time period, such as a day, or by user. This not only ameliorates a possible problem with overly short documents, but also focuses attention on the unit of interest. Whether this is time or a user (or speaker or other unit of authorship) will depend on the research problem. For other problems, segmenting a document into smaller units might be the answer. These could be structural, such as sentences or paragraphs, or some fixed-length chunk of tokens (segmented words). Fixed-length chunks are especially useful for sampling schemes, for instance in measuring textual characteristics using a moving average across a fixed window size of a text (e.g. Covington and McFall, 2010). Some schemes may combine these approaches, such as Labbé et al.’s (2004: 209) analysis of Charles de Gaulle’s broadcast speeches from June 1958 to April

1969: first they combined these into a single ‘document’ where each speech was in the order of broadcast, and then they applied a form of moving average measure of lexical diversity on segments that overlapped the original document boundaries defined by the speech dates.

Identifying units of analysis may also be done qualitatively, based on reading the texts and identifying politically relevant units of text. The best known example in political research is the ‘quasi sentence’ that forms the unit of analysis for the long-running Comparative Manifesto Project. Quasi sentences are textual units that express a policy proposition and may be either a complete natural sentence or part of one. Because some authors may express two distinct policy statements within a single natural sentence, the use of quasi sentences supposedly permits a more valid and complete representation of the content of the textual data. The trade-off, however, is that the same human decision process that interprets the sentence structure to identify text units also causes the procedure to be unreliable and often difficult or impossible to replicate (Däubler et al., 2012). This trade-off between reliability – whether repetition of a procedure produces stable results – and validity – whether the measurement or analysis reflects the truth of what is being measured or represented by the textual data – is a recurrent theme in research involving textual data. This affects not just the identification and preparation of units for analysis but also the design of coding frames and measurement and scaling models for analysing textual data.

The ability to redefine documents in terms of the smaller textual units they contain illustrates a curious feature of textual data: that the units of analysis are defined in terms of collections of the features. If we think of this data in matrix form (such as the intermediate stage of Figure 26.1), then the units of analysis are represented by the rows of documents and the features as columns derived from terms – indeed, this matrix is usually

called a *document–term matrix*. Since a document is just an arbitrary collection of terms, however, it means that the more we segment our document into smaller collections, the more it approaches being the unit of a feature defined by the column dimension of the data. Grouping documents does the opposite. Redefining the boundaries of what constitutes a ‘document’, therefore, involves shifting data from columns into rows or vice versa. This ability to reshape our data matrix because one dimension is defined in terms of a collection of the other is unique to text analysis. We could not perform a similar reshaping operation on, say, a survey dataset where we would not spread an individual’s observed responses across additional rows, because we cannot split an individual as a unit and because that individual is defined in terms of a sampled, physical individual, not as an arbitrary collection of survey questions.

Ultimately, how we reshape our documentary units by grouping or splitting them will depend on our research question and the needs of our method for analysing the data. Knowing how the sampling procedure for the textual data selection relates to the sampling units and the units of analysis may have implications for subsequent inference, given that the units of analysis are not randomly sampled textual data, irrespective of the sampling units. Determining which are most suitable will depend on the nature of the analytical technique and the insight it is designed to yield, and sometimes the length and nature of the texts themselves.

### ***Defining and Refining Features***

Features start with the basic semantic unit of text: the word. There are many forms of ‘words’, however, and these typically undergo a process of selection and transformation before they become features of our textual dataset. Words might also simply form the basis for recording an abstraction

triggered by the word, such as a dictionary key, or even a category assigned by a human annotator (in manual content analysis).

First, we should become familiar with some terms from linguistics. Words as they occur in a text are commonly known as *tokens*, so that the text ‘one two one two’ contains four tokens. *Tokenization* is the process of splitting a text into its constituent tokens, as in the second column of Figure 26.2 (which includes punctuation characters as tokens). Tokenisation usually happens by recognising the delimiters between words, which in most languages takes the form of a space. In more technical language, inter-word delimiters are known as *whitespace*, and include additional machine characters such as newlines, tabs

and space variants.<sup>8</sup> Most languages separate words by whitespace, but some major ones such as Chinese, Japanese and Korean do not. Tokenising these languages requires a set of rules to recognise word boundaries, usually from a listing of common word endings. Smart tokenisers will also separate punctuation characters that occur immediately following a word, such as the comma after *word* in this sentence. To introduce another term, *word types* refer to uniquely occurring words. So in our example, the four-token text contains only two word types, *one* and *two*. Comparing the rates of types and tokens forms the foundation for measures of lexical diversity (the rate of vocabulary usage), with the most common such measure comparing

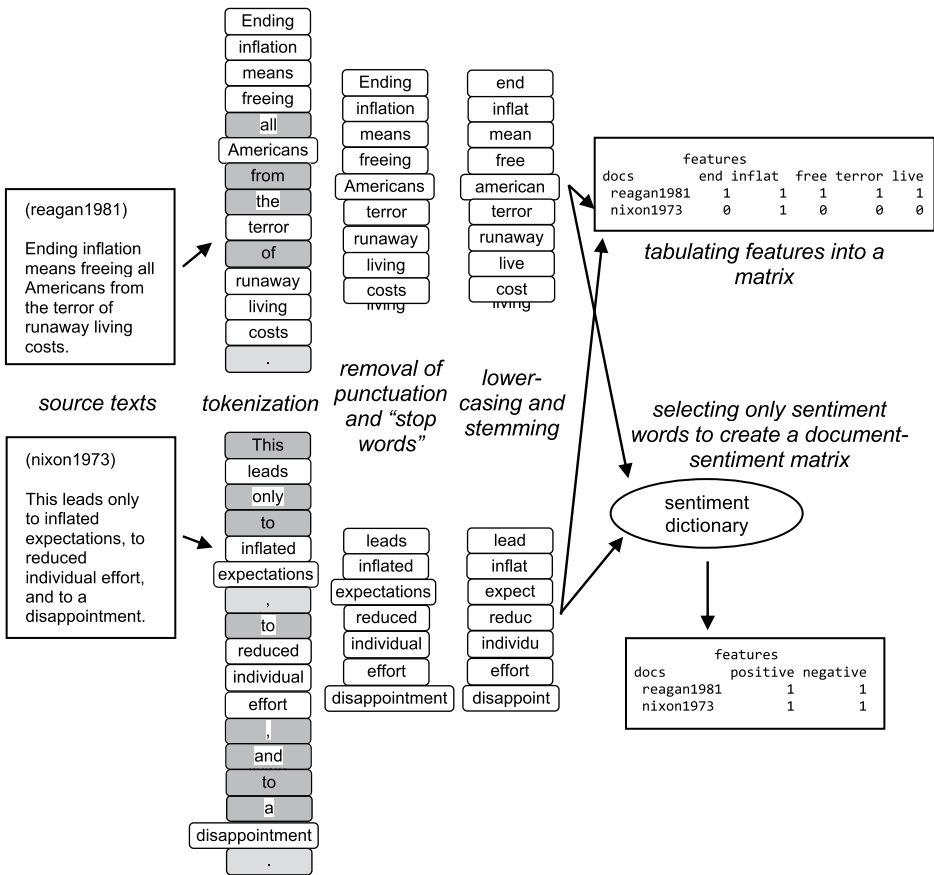


Figure 26.2 From text to tokens to matrix

the number of types to the number of tokens (the ‘type–token ratio’).

For a token to become a *feature* of textual data, it typically undergoes transformation in a step often called ‘pre-processing’ (although it should really be called processing). The most common types of token processing are *lower-casing*, which treats words as equivalent regardless of how they were capitalised; *stemming*, which reduces words to their stems, which is a cruder algorithmic means of equating a word with its *lemma*, or canonical (dictionary) form; and the elimination of words either through the use of pre-defined lists of words to be ignored or based on their relative infrequency. The first form of textual data pre-processing treats words as equivalent when they differ only in their inflected forms, so that, for example, the different words *taxes*, *tax*, *taxation*, *taxing*, *taxed* and *taxable* are all converted to their word stem *tax*. The second common textual pre-processing practice is to remove words that are considered unlikely to contribute useful information for analysis. These words, commonly called *stopwords*, are usually function words such as conjunctions, prepositions and articles that occur in the greatest frequency in natural language texts but add little specific political meaning to the text that would be deemed useful to analyse from textual data. (See Figure 26.2.) The problem with excluding words from a pre-set list, however, is that there exists no universally suitable list of words known to contribute nothing useful to all textual data analyses. For instance, the pronoun *her*, as Monroe et al. (2008) found, has a decidedly partisan orientation in debates on abortion in the US Senate. For these reasons, it has been noted that a general trend in preparing textual data for analysis has been gradually to reduce or eliminate reliance on stopword lists (Manning et al., 2008: 27). Another approach to restricting the focus of textual data analysis from all words to only potentially informative words is to filter words by indices constructed from their relative frequency across as well as

within documents, through a weighting or trimming scheme based on frequencies (discussed below), but this first requires a matrix of all eligible features to be formed.

Other methods of processing tokens include converting text to ‘*n*-grams’, defined as sequences of *n* consecutive tokens to form not words but phrases. This is a brute force method of recovering politically meaningful *multi-word expressions* that might contain identical unigrams but that as phrases mean exact opposites, such as *command economy* versus *market economy*. Also known as *collocations*, such expressions can be detected by statistical methods (e.g. Dunning, 1993). Detecting specific multi-word expressions is generally preferable to simply forming all *n*-grams, since the *n*-gram approach increases the number of features by (nearly) a multiple of *n*, and most of these will occur very rarely or represent frequently occurring but uninteresting combinations such as *let us*.

*Types* represent unique words, but we should remember that this uniqueness is typically based on their forming unique combinations of characters. Especially in English, *homographs* (words that are different but that spelled identically) will appear falsely as the same word type, at least to the machines we are using to process them. We could be more specific in distinguishing these by using a *part-of-speech* (POS) tagger that will at least distinguish homographs that are not the same parts of speech. In the example we cited earlier of the different uses of the term *kind*, for instance, a part-of-speech tagger could have annotated our tokens to distinguish these types (and this is indeed how I computed the proportions of its different forms in that example). Annotating tokens using a POS tagger can help us distinguish terms with opposite meanings such as *sanctions* in the sentence: *the President sanctions the sanctions against Iran*, by treating these as ‘sanctions/VERB’ and ‘sanctions/NOUN’, one meaning permission and the other meaning a penalty. Despite the obvious advantages,

however, differentiating word types using POS taggers in bag-of-words approaches to text as data is done seldomly, if ever.

It is not uncommon to read in a published application based on the analysis of text as data, perhaps in a footnote, that the authors took ‘the standard pre-processing steps’ to prepare their input texts. In truth there is no standard, and without details of the specific steps a researcher took, such summary references as to what invasive procedures were applied to the text are uninformative. Each application will have different needs for feature processing, with different consequences as a result of the choices made at this stage. In one of the few systematic studies of feature processing choices and their consequences, Denny and Spirling (2018: 187) replicated several published text analyses from political science using a variety of alternative feature processing steps. Their results show that ‘under relatively small perturbations of preprocessing decisions ... very different substantive interpretations would emerge’. Researchers in practice should be aware of these decisions, critically examine the assumptions of their methods and how these relate to feature selection and, test the robustness of these results.

### ***Converting Textual Features into a Quantitative Matrix***

This is mainly a mechanical step, resulting in a matrix whose dimensions are determined by the choices relating to the definitions of documents and features. We have already mentioned that some schemes call it a *document–term matrix*. (Some might even call it a *term–document matrix*, but there are great advantages in fixing the ‘documents’ to be row units and saving our efforts to promote diversity for more important problems.) We have been using the term *feature* thus far, but it is worth noting why and how this is different from just speaking about ‘terms’. Computer scientists use *feature* to refer to

what social scientists have long called *variables*: attributes of our units of analysis that differ across units. Because calling them *features* emphasises how they differ from terms or words (and may no longer even be words), I use this term to denote the selections and transformations made from token-based units that become the data used for analysis. I prefer the term *features* since the tokens have invariably been transformed in some way before they are shaped into a matrix, or may be abstractions from tokens such as annotations or dictionary keys rather than even transformed tokens.

Most matrices containing feature frequencies are characterised by a high degree of *sparsity*, meaning that they are mostly zeroes. Document–feature matrices are affected by what is known in machine learning as the *curse of dimensionality*: new observations also tend to grow the feature set, and each new term found in even a single document adds a new column to the matrix. There is even a ‘law’ named for this in linguistics: *Heap’s Law*, which states that the number of types grows exponentially with the number of tokens.<sup>9</sup> Forming a matrix of the (lower-cased) word features from the pre-2020 US presidential inaugural address corpus, for instance, creates a matrix of 58 inaugural speech documents by 9,273 features, but nearly 92 percent of the cells in this matrix of 537,834 cells are zeros. In fact, more than 41% of the features in this matrix are *hapax legomena*, defined as words that occur only a single time, such as the term *aborigines* in Ulysses S. Grant’s (politically incorrect) promise ‘to bring the aborigines of the country under the benign influences of education and civilisation’.

One strategy for mitigating the problem of exponentially increasing dimensionality is to trim or to weight the document–feature matrix. Trimming can be done on various criteria, but usually takes the form of a filter based on some form of feature frequency. Weighting schemes convert a matrix of counts into a matrix of weights. The

most common of these is relative term frequency, a weighting process also known as document *normalisation* because it homogenises the sum of the counts for each document. Since documents in a typical corpus vary in length, this provides a method for comparing frequencies more directly than counts, which are inflated in longer documents (although these frequencies are also subject to length effects related to Heap's law). Other popular weighting schemes are *tf-idf*, or term frequency-inverse document frequency, popular as a method in information retrieval for down-weighting the terms that are common to documents. In addition to the term frequency already discussed, *tf-idf* adds a weight that approaches zero as the number of documents in which a term appears (in any frequency) approaches the number of documents in the collection.<sup>10</sup> When we have selected our texts because they pertain to a specific topic – as we usually will – then inverse document frequency weighting means zeroing out most of our topical words, since these will appear in most or all documents. In texts of debates over health care, for instance, *tf-idf* weighting is likely to eliminate all words related to health care, even when they might occur at very different rates across different documents. If we think that it is not the occurrence, but rather the relative frequencies of words that are informative, then using *tf-idf* weighting is the opposite of what we want. While it will automatically remove 'stop words' without using a list, *tf-idf* weighting will also throw out the substantive baby with the linguistic bathwater. Except for classification tasks where removing all but the most discriminating features can improve performance, *tf-idf* weighting is usually inappropriate for the analysis of political texts.<sup>11</sup>

Because the rows and columns of the document–term matrix are unordered, the features that were originally carefully ordered words, in carefully ordered sentences, are now stored in a matrix object with no representation of order. In natural language

processing, this approach is known as 'bag-of-words', because it has disassociated the words from their context. For this reason, some text as data analyses use a different representation of documents based on token *vectors*, since these preserve order. For token vectors to be used in most analyses, however, such as computing a similarity score between token counts, these need to be aligned into what is effectively a matrix representation. Other forms of analysis, such as forming co-occurrence matrixes, require iterating over the token streams and tabulating counts that are later combined into matrix form.

We have already noted the curious inter-relationship between features, and documents as collections of features. Some matrix representations do away with the notion of documents altogether, forming feature-by-feature matrices counting how features co-occur within a defined context. This context might be the original document, or a moving local window for each target feature, for instance the five tokens found before or after the target feature. (Note here that I am very specifically using *token* to refer to a word when it exists as a segmented textual unit, but *feature* when it has been shaped into a matrix.) Known as a *feature co-occurrence matrix*, this matrix is a special variant of our document–feature matrix, where the documents have been redefined as features themselves, and the counts are tabulated within a context that we define. This is the basis for input into network analysis, for instance the inter-relationships of words based on their co-occurrence.

For simplicity, the focus here is on features based on a bag-of-words approach, but matrix representations can be generalised to include weights based on word embedding vectors, possibly redefining documents as new units such as sentences or paragraphs. We have already mentioned the popularity of vector representations of term features estimated from word embedding models. One option at the stage of creating the document–feature matrix is to combine the counts with weights or scores from these word vectors,

especially for comparing documents (for semantic similarity, for instance) or for text classification using predictive models. Methods exist for combining word vectors with *tf-idf* weights to turn documents into more semantically meaningful matrix representations, extending the notion of the document–feature matrix into a more complex representation than the simpler version depicted in Figure 26.2.

### **Analyzing the Textual Data Using an Appropriate Quantitative or Statistical Procedure**

The key here is *appropriate*: does the procedure for analysing the textual data produce reliable and valid insights into the question motivating the analysis? It is worth keeping in mind that by the time we have reached this stage of the analysis, we have already proceeded on the basis of some strong assumptions, namely:

- 1 The texts accurately represent the underlying target of concern.
- 2 Our sample of texts are a typical or at least complete representation of the phenomena that is our target of concern.
- 3 Our conversion of the texts into data has retained the essential information we need to provide insight on our target of concern.

The first assumption is by no means obvious in politics, where much verbal activity could be dismissed as ‘cheap talk’ or as insincere promises or false or misleading claims,<sup>12</sup> but we have good reason to think that text is more sincere than other forms of behaviour, especially in a legislative setting (Herzog and Benoit, 2015). Our selection from these also needs to be based on sound principles, just as data selection does in any research exercise. The third choice is something we have just discussed but involves many additional and deeper issues. It also interacts with a fourth strong assumption made at the analysis stage:

- 4 The analytic procedure yields a reliable and valid basis for inference on our target of concern.

The main risks with respect to reliability come when human judgement forms part of either the process of extracting data features or performing the analysis. In content analysis, for instance, human coders may be responsible for both defining the units of textual data and for assigning them annotations (‘codes’) based on their reading the textual units and judging the most applicable category from a set of instructions. The former process is known as *unitization* and the second as *coding* (Krippendorff, 2013 – although computer scientists typically call this text *annotation*). Both processes can pose severe challenges for even trained and highly educated human coders to apply at conventionally acceptable rates of reliability and inter-coder agreement (Mikhaylov et al., 2012). With respect to the potential unreliability of the analytic procedure, this is seldom a problem in text as data designs, because even the simplest procedures – such as comparing the relative rates of negative versus positive sentiment – involve quantitative comparisons that would not differ according to the judgement or personality of the analyst.

The validity of the analytic procedure in terms of providing insight on the target of concern is strongly influenced by the choices made at the feature extraction stage. Often, identical choices might be suitable for one analytic purpose but unsuitable for others. Consider the following three sentences, which we might wish to compare using a measure of textual similarity, such as *cosine similarity*, a measure that ranges (for text counts) between 0.0 to indicate the absence of any correlation and 1.0 to indicate two texts with identical feature proportions.

- (a) *Party X embraces protection of citizens through universal health care.*
- (b) *Party Y prioritises economic growth, even at the cost of environmental protection.*
- (c) *Party Y prioritises environmental protection, even at the cost of economic growth.*

The cosine similarity of text (a) with the second text is fairly low (at 0.32), as we might expect given that it concerns a different area of policy (but both are still statements about a party and both use the term *protection*). If we wanted to measure differences in the policy areas receiving *attention*, a measure of similarity based on vectors of word occurrences might suit our purposes well. But if we wished to measure differences in policy *position*, then cosine similarity in this example would be a poor instrument, as it indicates perfect similarity (1.0) between texts (b) and (c), despite these indicating exactly opposite political priorities. We can think of ways to differentiate them that might involve using sentence structure rather than simple bag-of-words approaches, but this only underscores the point that the appropriate choice of analytical procedure is influenced by choices made at the feature extraction stage.

Often there is an iterative process between the feature extraction and analysis stages, in which, following a preliminary analysis, we need to return to the feature extraction and processing stage in order to make adjustments before repeating the analysis. Sometimes, this might result from examining unintended or anomalous results of an analysis and deciding that these would be better avoided through different feature processing choices. Observing clusters of the same root terms with different inflections, for instance, could motivate stemming the tokens and repeating the analysis. Likewise, anyone who has plotted a *word cloud* of unselected features (where these appear in sizes proportional to their relative frequency) will quickly return to more aggressive feature selection when they see the words *the* and *and* dominating the plot. Other feature processing decisions can especially influence unsupervised methods such as topic models, because unsupervised approaches necessarily attempt to learn from all supplied information. Or, observing a set of topics sharing high proportions of stopwords might be cleaned up by removing

stopwords from features prior to fitting the topic model (and removing stopwords almost always improves the interpretability of topics fit using topic models). In their study of the effects of these choices, Denny and Spirling (2018: 187) found that key ‘modelling choices, such as the optimal number of topics, were also startlingly dependent’ on decisions made at the feature processing stage. Other techniques may be more robust to this, especially supervised methods or those that automatically down-weight uninformative features through their conditional probabilities or by applying a regularisation penalty. The best fine-tuning will be a combination of theoretically motivated choices of feature processing, confirmed by careful inspection following the analysis.

Should we be concerned that this cycle might encourage dishonesty, by tweaking our feature extraction until we get the results we want? In short no, although of course we should not contrive results. Residual diagnostics have long been a feature of basic statistical analysis, and often these serve to detect anomalies that indicate errors to be corrected before re-running the results, or fixes to be applied to get our data to conform more closely to the assumptions of our model (such as applying a log transformation to skewed variables or applying weights to heteroskedastic residuals in least-squares regressions). In working with textual data, this process is all the more important. Natural language often shows a slippery resistance to neat transformation into numerical data, because of features such as polysemy or the fact that many words in non-compounding languages lose an important part of their meaning when separated from the multi-word expressions in which they occur. Or it might be a simple matter of spelling or OCR mistakes indicating we have a cluster of words that should be the same but whose characters need correcting because an ‘i’ was rendered as an ‘l’ or a ‘o’ as a ‘0’, or because we did not remove running page footers from texts converted from pdf format. We should never underestimate



just how messy can be the process of converting text, no matter how clearly we can read it, into clean features of textual data. Often, the best – or the first – stage at which this becomes fully apparent is during analysis, and when detected it often means returning to earlier stages, cleaning things up or making better choices and repeating the analysis. This is a far more valid and honest approach than sticking with results that we know are wrong and could have fixed had we only gotten cleaner electronic texts to begin with or been better informed about the full consequences of our feature processing decisions.

Many of the analytic procedures we apply to textual features take the form of advanced statistical models that impose strong assumptions on the data-generating process, such as assuming that conditional word counts are identical and independently distributed as a Poisson (e.g. Slapin and Proksch, 2008), a negative binomial (e.g. Lo et al., 2016) or a multinomial process (e.g. Roberts et al., 2013). We know with certainty that words are not conditionally or positionally independent and that the degree of this will vary from mild to extreme in non-systematic ways, depending on the stylistic choices of a speaker or writer as well as characteristics of the language being used. To apply the tested and well-known properties of statistical data analysis to text, we must impose assumptions about the data-generating and stochastic processes that come with statistical approaches. The problem is, there exists no neat, parsimonious model of the data-generating process for natural language, so we rely on models whose assumptions are violated in sometimes painfully obvious ways. Fitting models that violate statistical assumptions is hardly new in social science, but because we so directly and intimately understand the nature of the source data (natural language) we are likely to be more acutely aware of these problems.

The good news is that even when violating statistical assumptions wholesale, we still get a tremendous amount of useful juice from models that are highly simplistic from a

linguistic point of view. The ‘naive’ in ‘Naive Bayes’, after all, is an overt recognition that its class conditional probabilities are wrong, because the assumption of independence required to compute the joint probabilities from word counts is blatantly fictional. Yet, Naive Bayes remains a highly useful tool for classifying texts (Zhang, 2004). It is hard to summarise this better than have Grimmer and Stewart (2013: 4):

The complexity of language implies that all methods *necessarily* fail to provide an accurate account of the data-generating process used to produce texts. Automated content analysis methods use insightful, but wrong, models of political text to help researchers make inferences from their data ... Including more realistic features into quantitative models does not necessarily translate into an improved method, and reducing the assumptions used may not imply more productive analyses. Rather, subtleties of applying the methods to any one data set mean that models that are less sophisticated in the use of language may provide more useful analysis of texts. (emphasis in original)

Two additional considerations often guide our choice of analytical method for analysing the features of textual data. One is interpretability, something we discuss more in our final stage. A second consideration is computational efficiency. Even with cheap, efficient computing resources, some models can be enormously expensive to fit. The advantages in low computational cost of fitting simpler, efficient models such as linear SVMs or Naive Bayes might well outweigh marginal gains in classifier performance from more advanced, but more computationally expensive models such as recurrent or convolutional neural network models. In addition, simpler methods often prove more robust in the sense of avoiding overfitting, a risk which every computer scientist acknowledges but which few explore in published applications (which typically aim at demonstrate how a new method has outperformed all other methods at some specific task for a specific dataset). As social scientists, we must give far greater priority to robustness and its

transparent demonstration in our choice of method for analysing text as data.

### ***Interpreting and Summarising the Results***

Summarising and communicating findings forms the end stage of any analysis, and the analysis of text as data is no different. Because it involves making sense following abstraction and analysis of raw input data that we could make direct sense of to begin with, however, interpreting the results of textual data analysis can involve some special challenges. Because the analytical stage involved using a trusted methodology, we typically stake our claim of validity of results on the basis that they inherit the trusted procedural properties of the methodology. But because the application of text analysis methods always involves choices at earlier stages, there is an additional measure of trust to establish upon interpreting results, namely that the researcher has appropriately processed the texts and correctly applied the analytic method. This is usually established through additional tests showing the robustness of the conclusions to different choices or demonstrating that the parameters of one's model (such as the number of topics) are optimal. Robustness checks are common in econometric analyses of non-textual data, but only recently have begun to form parts of textual data analyses in the political and social sciences.

Especially when text analysis is exploratory, such as demonstrating a new application or methodology, validation is a crucial part of interpreting one's results. For supervised scaling methods, this is tricky because there is seldom an objective measure with which text-based point estimates can be compared. Instead, we typically rely on comparison to some external measures obtained through alternative, often non-textual means, such as expert survey estimates of policy positions in the case of scaling ideology. Validating

supervised classification methods is easier, because we could have objective labels for verifying predicted classes (such as observed party affiliation), numeric scores (from a survey question) or labels assigned through human annotation.

Interpreting the results of unsupervised methods is trickier, because these results often involve reading into the textual contents of topics or word weights and deciding whether they accord with some reasonable interpretation of the world. Point estimates from unsupervised scaling can be compared to the same sorts of external measures as supervised scaling estimates, or to summaries of detailed human readings of the scaled texts (e.g. Lowe and Benoit, 2013). Topic models are trickier, but typically involve reading the word features that are the most frequent in each topic and assigning a label to that topic. Roberts et al. (2014: 1073) for instance interpreted their 'Topic 1' as 'the "crime" and "welfare" or "fear" topic', because its most frequently used word features included *illeg*, *job*, *immigr*, *welfar* and *crime*. Their second topic, which they interpreted as emphasising 'the human elements of immigrants', also contained among its most frequent word features *immigr*, *illeg*, *border* and *worri*. These distinctions are hardly clear-cut, and any labels attached to topics are ultimately subjective. Model-based diagnostics for setting an optimal number of topics, furthermore, may be unrelated or even negatively correlated with topics' semantic coherence (Chang et al., 2009). The best application of unsupervised methods will produce results that are semantically coherent with our understandings of the texts and with the world that our analysis of them aims to represent.

As analytical tools become increasingly sophisticated, we now have access to powerful methodologies whose procedural workings may be non-transparent. No one has figured out the data-generating process of language, but with modern approaches for classification, this has become unnecessary. Some of the best performing classification

methods for text, for instance, use ‘deep learning’ models fit to the level of characters. When fed with enormous amounts of data, convolutional neural network models can outperform other approaches (Zhang et al., 2015) but it is impossible to assess their operation in any application in the way one would diagnose even an advanced computational method such as fitting a Markov chain Monte Carlo model. Because social science eschews black boxes, we often stick to interpretable models even when better-performing alternatives exist, especially if we have large quantities of data. If a huge amount of training data are available, ‘then the choice of classifier probably has little effect on your results’ (Manning et al., 2008: 337), and we should be guided by the principle of parsimony to prefer more transparent and simpler models over more opaque and complex ones, even at the cost of small trade-offs in performance. Just as our concern in social science is explanation rather than prediction, we generally prefer model specification based on theory and isolating the effect specific explanatory factors, not attempting to include every possible variable to maximise variance explained. Because the goals of explanation or measurement differ from the (typical machine learning) objective of prediction, it is worth reminding ourselves of this preference.

Communicating the results of text analysis in a compact and effective way is practically challenging because numerical tables only poorly capture the full nuances of language, and we typically have too many features and documents (or topics) to fit these easily into a format that will not overwhelm a reader. Graphical presentation of text analytic results is especially important, and should offer special opportunities given that we can read and interpret word features when they form the elements of a plot. Despite this potential, however, innovation in visual presentation of text analytic results has been slow to non-existent, moving little beyond the ‘word cloud’ and its variations. Designed to show the most frequent terms, the word cloud plots

features in sizes proportional to their relative frequency in the textual dataset, producing a plot with some visual appeal but often no clear communication of any particular result. This is slightly improved by using a ‘comparison’ word cloud that partitions word plots according to external categories, such as the Twitter hashtags used in Figure 26.3 according to whether the user was predicted to support Brexit or not (Amador Diaz Lopez et al., 2017). Other methods exist, of course, especially for characterizing the semantic content of topics from topic models, probably the area in which the most innovations of visual presentation in text analysis have occurred (e.g. figure 5 from Reich et al., 2015). Given the unique interpretability of word features, however, it is justified to feel that we should have developed more imaginative graphical ways to include words in our plots (and not just on the axis labels).

A final word on presentation and interpretation concerns how we characterize the *uncertainty* of our text analysis results. In addition to inheriting procedural validity established by decades of statistical theory, the quantitative analysis of text as data also makes it possible to quantify the uncertainty of our results. In the analysis of text as data, this can take two forms: parametric and non-parametric. Parametric methods rely on the assumptions we have imposed on the data through some model of its stochastic process, in the context of an established procedure for producing estimates – such as maximum likelihood or simulations from Bayesian posterior distributions. These are typically too small because of unmodeled heterogeneity in our model of text data, but even this bias can be quantified. Another approach is non-parametric, through bootstrapping a text by resampling from its elements. In exploring different methods of characterising uncertainty for measurement models of text, Lowe and Benoit (2013) advocate repeating the analysis with different versions of a text that have been reassembled after resampling their sentences



**Figure 26.3** Word cloud of influential hashtags from a sample of Tweets about Brexit

Source: Amador Diaz Lopez et al. (2017).

with replacement – bootstrapping documents from their sentences, in other words. This method has begun to appear in different applications, such as Benoit et al.’s (2019) use of it to compute confidence intervals for document-level readability statistics, but it has been slow to catch on despite its almost universal applicability to text as data analyses. Measuring uncertainty in the analysis of text as data remains one of the most important challenges in this field (Grimmer and Stewart, 2013: 28), and should be a requirement if we are to give the quantitative analysis of text full methodological status alongside that of non-textual data.

## CONCLUSION AND FUTURE DIRECTIONS

Treating text as data means converting it into features of data and analysing or mining

these features for patterns, rather than making sense of a text directly. This process turns text from something directly meaningful into data that we cannot interpret in its raw form, but whose analysis produces meaningful insights using structured rules in ways and at magnitudes that would be impossible without having treated the text as data. This approach to text analysis has become increasingly mainstream in the political and social science, and the methods and applications increasingly innovative. This trend, which is likely to continue, has been driven by several forces.

First, as in so many other areas of human activity, in textual analysis the rise of the machines has enabled scholars to automate key parts of the analytic process, a process formerly performed using qualitative methods by unreliable humans who actually knew what they were doing. With text analytic methods, humans can now mine large quantities of textual data, using sophisticated

methods, implemented by perfectly reliable computer automatons.

A second force driving the textual revolution has been one of scale: the incredible volume of texts available today requires automated, quantitative approaches if we are to analyse more than a small subset of this data. The growth of electronic publications has made machine-readable text ubiquitous, and along with it comes the promise of a huge wealth of information about the characteristics of the political and social actors that generate these texts. Such texts include legislative speeches, political party manifestos, legal decisions, election campaign materials, press releases, social media posts, correspondence and television and radio transcripts, to name but the key ones. Resource limitations may still cause us to sample from available texts, but this involves much larger samples than in previous eras. Miners want to extract all, not just a sample, of the rich resources available, and the logic of text mining points to the same conclusion. Methods that require reading a text, or determining what it 'means', are simply not applicable to a scale of tens or hundreds of thousands or more texts. Instead, we need tools that can turn unstructured text into structured information, using inexpensive and efficient methods for parsing, annotating and categorising the elements of text to prepare it for analysis and then to perform this analysis.

Of course, access to big textual data and the machines to process it are only as useful as the methodologies that the machines can implement. A final enabler (and driver) of the shift to treating text as data has been the development and application of sophisticated statistical learning methods for extracting information and generating inferences from textual data. These are extensions of statistics and machine learning but with specific applications to textual data.

Many challenges lie ahead, and these should be met in the same way as most other breakthroughs in social science methodology: through innovations required to solve specific

research problems as part of our agenda to understand the political and social world. Some challenges have already been identified, such as a need for improved validation of our models of textual data under a broader range of circumstances, and a more realistic way to characterise uncertainty in textual data analysis. Some are just emerging, such as how to incorporate named entity recognition and part-of-speech tagging to distinguish alternative meanings, or how to identify and make use of multi-word, non-compositional phrases (how to distinguish, in other words, *Homeland Security* from *social security*). Other recent innovations include the merger of human qualitative input for processing textual data with statistical scaling or machine learning for analysis, possibly using crowd-sourcing for text annotation through an agile, 'active learning' process. As advances continue in other fields such as machine learning and natural language processing in computer science, we must keep a firm grip on political science research objectives and standards while at the same time borrowing what is useful to our discipline. As we gain experience and understanding through both theory and applications, textual data analysis will continue to mature and continue to produce valuable insights for our understanding of politics.

## ACKNOWLEDGEMENTS

I am deeply grateful for the patience shown by the editors of this *Handbook* and for their invaluable feedback on earlier versions. For feedback and comments I also thank Audrey Alejandro, Elliott Ash, April Biccum, Dan Hopkins, Cheryl Schonhardt-Bailey, Michael Laver, Stephan Ludwig, Ludovic Rheault, Arthur Spirling and John Wilkerson. The code required to produce all of the analysis in this chapter can be viewed at <https://gist.github.com/kbenoit/4593ca1deeb4d890077f3b12ba468888>.

## Notes

- 1 From Ronald Reagan's 1981 inaugural address: see <https://www.presidency.ucsb.edu/documents/presidential-documents-archive-guidebook/inaugural-addresses>.
- 2 Exactly such an analysis has been applied by Nick Beauchamp to the 'I Have a Dream' speech. Beauchamp's 'Plot Mapper' algorithm segments the text into sequential chunks, creates a chunk-term count matrix, computes the principal components of this matrix, standardises the resulting scores and plots the first two dimensions to show the rhetorical arc of a speech. See <http://www.nickbeauchamp.com/projects/plotmapper.php>.
- 3 For another cringeworthy example of procedural barbarity committed against a great political text, see Peter Norvig's 'The Gettysburg Powerpoint Presentation', <https://norvig.com/Gettysburg/>.
- 4 This example is taken from a very widely used psychological dictionary known as the Linguistic Inquiry and Word Count (2015 version): Tausczik and Pennebaker (2010).
- 5 For a deeper general discussion of these issues, see Steyvers and Griffiths (2007).
- 6 The history of encoding is a long and complicated saga that most practitioners of text analysis would happily ignore. It has to do with how the original 7-bit (containing 128 characters, or  $2^7$ ) 'ASCII' character set needed adaptation to new languages and symbols by adding an eighth bit. There was very little standardisation in how the resulting additional 128 characters were mapped, so that text sent in, for example, encoded in Windows-1250 (for Central and East European languages) would look garbled on a system using the similar, but not identical ISO-8859-2 for words like źródło.
- 7 This also explains the apparently widespread usage in the 1700s of the 'f-word': not even Google Books has been able to distinguish it from the work *suck*. See [https://books.google.com/ngrams/graph?content=fuck&year\\_start=1700&year\\_end=2000](https://books.google.com/ngrams/graph?content=fuck&year_start=1700&year_end=2000)
- 8 Not Klingons, but rather the variations on the simple space character included in the Unicode 'Separator, Space' category, such as U+205F, the 'Medium Mathematical Space'.
- 9 Technically speaking, Zipf's Law states that  $M = kT^b$ , where  $M$  is the vocabulary size (the number of unique word types),  $T$  is the number of tokens, and  $k$  and  $b$  are constants for computational linguists to estimate and argue about (but that are usually  $30 \leq k \leq 100$  and  $b \approx 0.5$   $30 \leq k \leq 100$ ). Manning et al., 2008: 88).
- 10 Perhaps surprisingly, there is no universal definition of *tf-idf* weighting, and formulas may differ

depending on whether the *tf* is a count or a proportion, what sort of constant may be added, or what logarithmic base and constant are applied to the inverse document frequency. A good measure, however, is  $tf_{ij} * \log_{10} \frac{N}{df_j}$ , where  $tf_{ij}$  is the count of feature  $j$  in document  $i$ ,  $N$  is the number of documents in a collection, and  $df_j$  is the number of documents in which feature  $j$  occurs (Manning et al., 2008: 118). A feature occurring in all  $N$  documents thus receives a weight of zero since  $\log(1) = 0$ .

- 11 We could also add that many models commonly used in political science – such as the 'Wordfish' Poisson scaling model or variants of Latent Dirichlet allocation (topic) models – only work with counts as inputs, so that *tf-idf* or other weighting schemes are inapplicable.
- 12 See Kessler, Rizzo and Kelly (2019), 'President Trump made 8,158 false or misleading claims in his first two years', *Washington Post*, 21 January. <https://www.washingtonpost.com/politics/2019/01/21/president-trump-made-false-or-misleading-claims-his-first-two-years>

## REFERENCES

- Amador Diaz Lopez, Julio Cesar, Sofia Colignon-Delmar, Kenneth Benoit and Akitaka Matsuo. 2017. 'Predicting the Brexit Vote by Tracking and Classifying Public Opinion Using Twitter Data.' *Statistics, Politics and Policy* 8(1):210–220.
- Baek, Young Min, Joseph N. Cappella and Alyssa Bindman. 2011. 'Automating Content Analysis of Open-Ended Responses: *Wordscores* and *Affective Intonation*.' *Communication Methods and Measures* 5(4):275–296.
- Baumgartner, Frank R., Christoffer, Green-Pedersen and Bryan D. Jones. 2008. *Comparative Studies of Policy Agendas*. Routledge.
- Benoit, Kenneth, Kevin Munger and Arthur Spirling. 2019. 'Measuring and Explaining Political Sophistication through Textual Complexity.' *American Journal of Political Science* 63(2):491–508.
- Benoit, Kenneth, Michael Laver, Christine Arnold, Madeleine O. Hosli and Paul Penning. 2005. 'Measuring National Delegate Positions at the Convention on the Future of Europe Using Computerized Wordscoring.' *European Union Politics* 6(3):291–313.

- Berelson, Bernard. 1952. *Content Analysis in Communications Research*. New York: Free Press.
- Blei, D. M., A. Y. Ng and M. I. Jordan. 2003. 'Latent Dirichlet Allocation.' *The Journal of Machine Learning Research* 3:993–1022.
- Bolukbasi, Tolga, Kai-Wei Chang, James Zou, Venkatesh Saligrama and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS'16)*. Curran Associates Inc., Red Hook, NY, 4356–4364.
- Budge, Ian, David Robertson and Derek Hearl, eds. 1987. *Ideology, Strategy and Party Change: Spatial Analyses of Post-War Election Programmes in 19 Democracies*. Cambridge: Cambridge University Press.
- Budge, Ian, Hans-Dieter Klingemann, Andrea Volkens, Judith Bara and Eric Tanenbaum. 2001. *Mapping Policy Preferences: Estimates for Parties, Electors, and Governments 1945–1998*. Oxford: Oxford University Press.
- Caliskan, Aylin, Joanna J. Bryson and Arvind Narayanan. 2017. 'Semantics derived automatically from language corpora contain human-like biases.' *Science* 356(6334): 183–186.
- Chang, Jonathan, Jordan Boyd-Graber, Sean Gerrish, Chong Wang, and David M. Blei. 2009. Reading tea leaves: how humans interpret topic models. In *Proceedings of the 22nd International Conference on Neural Information Processing Systems (NIPS'09)*. Curran Associates Inc., Red Hook, NY, pp. 288–296.
- Chilton, Paul. 2017. "'The People" in Populist Discourse: Using Neuro-Cognitive Linguistics to Understand Political Meanings.' *Journal of Language and Politics* 16(4):582–594.
- Colleoni, Elanor, Alessandro Rozza and Adam Arvidsson. 2014. 'Echo Chamber or Public Sphere? Predicting Political Orientation and Measuring Political Homophily in Twitter Using Big Data.' *Journal of Communication* 64(2):317–332.
- Covington, Michael A. and Joe D. McFall. 2010. 'Cutting the Gordian Knot: The Moving-Average Type-Token Ratio (MATTR).' *Journal of Quantitative Linguistics* 17(2):94–100.
- Däubler, Thomas and Kenneth Benoit. 2018. 'Estimating Better Left–Right Positions Through Statistical Scaling of Manual Content Analysis.' London School of Economics typescript.
- Däubler, Thomas, Kenneth Benoit, Slava Mikhaylov and Michael Laver. 2012. 'Natural Sentences as Valid Units for Coded Political Texts.' *British Journal of Political Science* 42(4):937–951.
- Denny, Matthew J. and Arthur Spirling. 2018. 'Text Preprocessing for Unsupervised Learning: Why It Matters, When It Misleads, and What to Do about It.' *Political Analysis* 26(2):168–189.
- Dunning, Ted. 1993. 'Accurate Methods for the Statistics of Surprise and Coincidence.' *Computational Linguistics* 19:61–74.
- Fairclough, Norman. 1992. 'Intertextuality in Critical Discourse Analysis.' *Linguistics and Education* 4(3–4):269–293.
- Fairclough, Norman. 2001. *Language and Power*. Harlow, UK: Pearson Education.
- Fang, Xing and Justin Zhan. 2015. 'Sentiment Analysis Using Product Review Data.' *Journal of Big Data* 2(5):1–14.
- Firth, John R. 1957. A Synopsis of Linguistic Theory, 1930–1955. In *Studies in Linguistic Analysis*. Oxford: Blackwell pp. 1–32.
- Flesch, Rudolph. 1948. 'A New Readability Yardstick.' *Journal of Applied Psychology* 32(3):221–233.
- Foucault, Michel. 1972. *The Archaeology of Knowledge and the Discourse on Language*. New York: Pantheon Books.
- Greenacre, Michael. 2017. *Correspondence Analysis in Practice*. 3rd ed. Boca Raton, FL: Chapman and Hall/CRC.
- Grimmer, Justin 2010. 'A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases.' *Political Analysis* 18(1):1–35.
- Grimmer, Justin and Brandon M. Stewart. 2013. 'Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts.' *Political Analysis* 21(3):267–297.
- Gunning, Robert. 1952. *The Technique of Clear Writing*. New York: McGraw-Hill.
- Herdan, Gustav. 1955. 'A New Derivation and Interpretation of Yule's "Characteristic" K.' *Zeitschrift für angewandte Mathematik und Physik ZAMP* 6(4):332–334.
- Hertel-Fernandez, Alexander and Konstantin Kashin. 2015. 'Capturing Business Power

- across the States with Text Reuse.' Presented at the Annual Meetings Midwest Political Science Association, Chicago, 16–19 April.
- Herzog, Alexander and Kenneth Benoit. 2015. 'The Most Unkindest Cuts: Speaker Selection and Expressed Government Dissent during Economic Crisis.' *The Journal of Politics* 77(4):1157–1175.
- Jagers, Jan and Stefaan Walgrave. 2007. 'Populism as Political Communication Style: An Empirical Study of Political Parties' Discourse in Belgium.' *European Journal of Political Research* 46(3):319–345.
- Joachims, Thorsten. 1999. Transductive inference for text classification using support vector machines. In *Proceedings of the Sixteenth International Conference on Machine Learning (ICML 1999)*, Bled, Slovenia, June 27–30. pp. 200–209.
- Klemmensen, Robert, Sara Binzer Hobolt and Martin Ejnar Hansen. 2007. 'Estimating Policy Positions Using Political Texts: An Evaluation of the Wordscores Approach.' *Electoral Studies* 26(4):746–755.
- Klüver, Heike. 2009. 'Measuring Interest Group Influence Using Quantitative Text Analysis.' *European Union Politics* 10(4):535–549.
- Krippendorff, Klaus. 2013. *Content Analysis: An Introduction to Its Methodology*. 3rd ed. Thousand Oaks, CA: Sage.
- Labbé, Cyril, Dominique Labbé and Pierre Hubert. 2004. 'Automatic Segmentation of Texts and Corpora.' *Journal of Quantitative Linguistics* 11(3):193–213.
- Lai, Siwei, Liheng Xu, Kang Liu and Jun Zhao. 2015. Recurrent Convolutional Neural Networks for Text Classification. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI'15)* pp. 2267–2273. Austin, Texas.
- Landauer, Thomas K., Peter W. Foltz and Darrell Laham. 1998. 'An Introduction to Latent Semantic Analysis.' *Discourse Processes* 25(2–3):259–284.
- Lasswell, Harold Dwight. 1948. *Power and Personality*. New York: W. W. Norton and Company.
- Laver, Michael and Kenneth Benoit. 2002. 'Locating TDs in Policy Spaces: Wordscoring Dáil Speeches.' *Irish Political Studies* 17(1):59–73.
- Laver, Michael, Kenneth Benoit and John Garry. 2003. 'Estimating the Policy Positions of Political Actors Using Words as Data.' *American Political Science Review* 97(2):311–331.
- Leites, Nathan, Elsa Bernaut and Raymond L. Garthoff. 1951. 'Politburo Images of Stalin.' *World Politics* 3(3):317–339.
- Liu, Dilin and Lei Lei. 2018. 'The Appeal to Political Sentiment: An Analysis of Donald Trump's and Hillary Clinton's Speech Themes and Discourse Strategies in the 2016 US Presidential Election.' *Discourse, Context & Media* 25:143–152.
- Lo, James, Sven-Oliver Proksch and Jonathan B. Slapin. 2016. 'Ideological Clarity in Multiparty Competition: A New Measure and Test Using Election Manifestos.' *British Journal of Political Science* 46(3):591–610.
- Loughran, Tim and Bill McDonald. 2011. 'When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks.' *The Journal of Finance* 66(1):35–65.
- Lowe, William and Kenneth Benoit. 2013. 'Validating Estimates of Latent Traits From Textual Data Using Human Judgment as a Benchmark.' *Political Analysis* 21(3):298–313.
- Lucas, C., R. A. Nielsen, M. E. Roberts, B. M. Stewart, A. Storer and D. Tingley. 2015. 'Computer-Assisted Text Analysis for Comparative Politics.' *Political Analysis* 23(2): 254–277. doi:10.1093/pan/mpu019.
- Manning, C. D., P. Raghavan and H. Schütze. 2008. *Introduction to Information Retrieval*. Cambridge: Cambridge University Press.
- Martindale, Colin. 1975. *Romantic Progression: The Psychology of Literary History*. New York: Hemisphere Publishing Corporation.
- Mikhaylov, Slava, Michael Laver and Kenneth Benoit. 2012. 'Coder Reliability and Misclassification in the Human Coding of Party Manifestos.' *Political Analysis* 20(1):78–91.
- Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2 (NIPS'13)* pp. 3111–3119. Red Hook, NY: Curran Associates Inc.
- Monroe, Burt L. and Ko Maeda. 2004. 'Talk's Cheap: Text-Based Estimation of Rhetorical Ideal-Points.' POLMETH Working Paper.
- Monroe, Burt L., Kevin M. Quinn and Michael P. Colaresi. 2008. 'Fightin' Words: Lexical



- Feature Selection and Evaluation for Identifying the Content of Political Conflict.' *Political Analysis* 16(4):372–403.
- Monroe, Burt L. and Philip A. Schrodt. 2009. 'Introduction to the Special Issue: The Statistical Analysis of Political Text.' *Political Analysis* 16(4):351–355.
- Pang, Bo, Lilian Lee and Shivakumar Vaithyanathan. 2002. 'Thumbs Up? Sentiment Classification Using Machine Learning Techniques.' In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)* pp. 79–86. Association for Computational Linguistics.
- Pennington, J., R. Socher and C. D. Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* pp. 1532–1543. Association for Computational Linguistics.
- Perry, Patrick O. and Kenneth Benoit. 2018. 'Scaling Text with the Class Affinity Model.' *arXiv preprint arXiv:1710.08963*.
- Peters, Matthew E., Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee and Luke Zettlemoyer. 2018. 'Deep contextualized word representations.' In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics* pp. 2227–2237. New Orleans, Louisiana.
- Proksch, Sven-Oliver and Jonathan B. Slapin. 2010. 'Position Taking in the European Parliament Speeches.' *British Journal of Political Science* 40(3):587–611.
- Quinn, Kevin M., Burt L. Monroe, Michael Colaresi, Michael H. Crespin and Dragomir R. Radev. 2010. 'How to Analyze Political Attention with Minimal Assumptions and Costs.' *American Journal of Political Science* 54(1):209–228.
- Reich, Justin, Dustin H. Tingley, Jetson Leder-Luis, Margaret Roberts and Brandon Stewart. 2015. 'Computer-Assisted Reading and Discovery for Student Generated Text in Massive Open Online Courses.' *Journal of Learning Analytics* 2(1):156–184.
- Reisigl, Martin. 2008. Analyzing Political Rhetoric. In *Qualitative Discourse Analysis in the Social Sciences*, ed. Ruth Wodak and Michal Krzyżanowski. Basingstoke: Palgrave Macmillan pp. 96–120.
- Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson and David G. Rand. 2014. 'Structural Topic Models for Open-Ended Survey Responses.' *American Journal of Political Science* 58(4):1064–1082.
- Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley, Edoardo M. Airoidi. 2013. The structural topic model and applied social science. In *Advances in Neural Information Processing Systems Workshop on Topic Models: Computation, Application, and Evaluation*. pp. 1–20.
- Sanders, James, Giulio Lisi and Cheryl Schonhardt-Bailey. 2017. 'Themes and Topics in Parliamentary Oversight Hearings: A New Direction in Textual Data Analysis.' *Statistics, Politics and Policy* 8(2):153–194.
- Schonhardt-Bailey, Cheryl. 2003. 'Ideology, Party and Interests in the British Parliament of 1841–47.' *British Journal of Political Science* 33(4):581–605.
- Schonhardt-Bailey, Cheryl. 2008. 'The Congressional Debate on Partial-Birth Abortion: Constitutional Gravitas and Moral Passion.' *British Journal of Political Science* 38(3):383–410.
- Slapin, Jonathan B. and Sven-Oliver Proksch. 2008. 'A Scaling Model for Estimating Time-Series Party Positions from Texts.' *American Journal of Political Science* 52(3):705–722.
- Slapin, Jonathan B. and Sven-Oliver Proksch. 2014. Words as Data: Content Analysis in Legislative Studies. In *The Oxford Handbook of Legislative Studies*, ed. Shane Martin, Thomas Saalfeld and Kaare W. Strøm. Oxford: Oxford University Press pp. 126–144.
- Spache, George. 1953. 'A New Readability Formula for Primary-Grade Reading Materials.' *The Elementary School Journal* 53(7):410–413.
- Spirling, Arthur. 2016. 'Democratization and Linguistic Complexity: The Effect of Franchise Extension on Parliamentary Discourse, 1832–1915.' *Journal of Politics* 78(1):120–136.
- Steyvers, Mark and Tom Griffiths. 2007. Probabilistic Topic Models. In *Handbook of Latent Semantic Analysis*, ed. Thomas K. Landauer, Danielle S. McNamara, Simon Dennis and Walter Kintsch. Vol. 427. New York: Routledge pp. 424–440.

- Stone, Philip J., Dexter C. Dunphy and Marshall S. Smith and Daniel M. Olgvie. 1966. *The General Inquirer: A Computer Approach to Content Analysis*. Cambridge, MA: MIT Press.
- Taddy, Matt. 2013. 'Multinomial Inverse Regression for Text Analysis.' *Journal of the American Statistical Association* 108(503):755–770.
- Tausczik, Yla R. and James W. Pennebaker. 2010. 'The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods.' *Journal of Language and Social Psychology* 29(1):24–54.
- van Dijk, Teun A. 1994. 'Critical Discourse Analysis.' *Discourse & Society* 5(4):435–436.
- van Dijk, Teun. 1997. 'What Is Political Discourse Analysis.' *Belgian Journal of Linguistics* 11(1):11–52.
- Walgrave, Stefaan and Knut De Swert. 2007. 'Where Does Issue Ownership Come From? From the Party or from the Media? Issue-Party Identifications in Belgium, 1991–2005.' *The Harvard International Journal of Press/Politics* 12(1):37–67.
- Wilkerson, John, David Smith and Nicholas Stramp. 2015. 'Tracing the Flow of Policy Ideas in Legislatures: A Text Reuse Approach.' *American Journal of Political Science* 59(4):943–956.
- Wilkerson, John and Andreu Casas. 2017. 'Large-Scale Computerized Text Analysis in Political Science: Opportunities and Challenges.' *Annual Review of Political Science* 20:529–544.
- Wimsatt, William K. and Monroe C. Beardsley. 1946. 'The Intentional Fallacy.' *The Sewanee Review* 54(3):468–488.
- Yu, Bei, Stefan Kaufmann and Daniel Diermeier. 2008. 'Classifying Party Affiliation from Political Speech.' *Journal of Information Technology & Politics* 5(1):33–48.
- Zhang, H. 2004. The Optimality of Naive Bayes. In *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference, (FLAIRS)*.
- Zhang, Xiang, Junbo Zhao and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1 (NIPS'15)* pp. 649–657. Cambridge: MIT Press.

# Scaling Political Positions from Text: Assumptions, Methods and Pitfalls

Benjamin C.K. Egerod and Robert Klemmensen<sup>1</sup>

## WHY DO WE WANT TO SCALE TEXTS?

Virtually all instances of political conflict can be thought of in spatial terms. In everyday language as well in academic discourse we use metaphors relating to space when describing politics. Indeed, it is difficult to even talk about politics without ‘using the notions of position, distance, and movement’ (Benoit and Laver, 2006: 12). In politics, the left–right distinction – by definition a spatial notion – may be the most enduring organizing principle (Bobbio, 1996), and the underlying conceptualization of political preferences distributed along different latent dimensions is closely linked to the spatial models of politics often associated with Downs (1957), Smithies (1941) and Hotelling (1990). While the prevalence of the left–right distinction has made it natural to focus on political ideology, most instances concerning differences between preferences can be thought of in spatial terms. For instance, interest group scholars often conceptualize

the degree to which special interests attain their preferences in terms of how policy proposals move relative to the stated positions of groups (Dür, 2008; Bernhagen et al., 2014; Klüver, 2009). Indeed, there is good reason to believe that spatial models are not just good ways of representing multidimensional data, but good approximations of how humans think about preferences (Armstrong et al., 2014). Therefore, considerable amounts of energy have been devoted to developing methods which can reliably and validly place actors in political spaces. Scaling methods are devoted to precisely that, and have a long history of successfully placing legislators, political parties and judges in ideological spaces (e.g. Martin and Quinn, 2002; Poole and Rosenthal, 1985, 2000). More recently, the surge in computational power and availability of new forms of data has provided the possibility of scaling political preferences of extremely diverse sets of actors (Bonica, 2013, 2014; Barberá, 2015; Bond and Messing, 2015; Crosson et al., 2018).

It is in this landscape that the scaling of political positions from text fits. The techniques that are used to scale texts are in large part parallel to the ones used for estimating positions from other data sources, and the use of text scaling has evolved in a similar fashion. While scaling positions from hand-coded databases has a long and successful history (Klingemann et al., 2006), we have seen a surge in the application of computationally intensive methods for scaling texts without first manually coding them. This is due both to the explosion of texts available, and to an increase in the techniques and the computational power that allows us to use them (Martin and Yurukoglu, 2017; Monroe and Maeda, 2004; Lo et al., 2016; Slapin and Proksch, 2008; Laver et al., 2003). Computational text scaling offers an extremely wide range of potential applications, and while its use is constantly growing and new estimators are continuously invented and applied to innovative data sources, there is no doubt that the field will continue its development for many years to come.

This chapter is dedicated to introducing computational techniques for scaling policy positions from political texts. Because text scaling is similar in theory to other forms of scaling, but also presents its own challenges, we introduce the reader not only to how common models can be applied, but also to the particular assumptions they make – and how they can be broken.

We start by discussing how text scaling relates to the broader field of measurement theory as it has evolved in political science, and the core assumptions that are needed to scale a set of texts. This structures our review of specific methods for text scaling. We discuss how techniques vary in their assumptions, and illustrate their use with a diverse set of political texts. We discuss the work done by Laver et al. (2003), which introduced the use of automatic scaling of text to political science. We proceed to review the Poisson scaling model (Monroe and Maeda, 2004; Slapin and Proksch, 2008), which

scales policy position with practically no input from the researcher. The techniques we discuss in this chapter represent but a small subset of the universe of potential scaling techniques. Therefore, we also discuss how each estimator has been extended. While it is impossible to cover all possible estimators in a single chapter, we hope that this chapter can serve as a starting point for the reader. Finally, to illustrate some of the potential pitfalls when scaling texts, we include two simulation studies investigating (a) how short texts can be, and (b) how differently they can use their words before common methods are no longer able to meaningfully place them in space. The final section concludes with a discussion of how our existing toolbox can be extended by incorporating new methods for taking word context into account, and how we can think about measurement error more productively than by simply discarding models.

## **TEXT SCALING AS INFERENCE ABOUT LATENT POSITIONS**

The goal of methods for scaling positions is to use some observed set of outcomes to draw inferences about an actor's (in the widest sense of the word) unobservable position on a latent dimension relative to other actors. Position is here to be understood as the political preference on some dimension. To get at such a position, the observed outcomes must reveal some kind of preference on the part of the actor. While this holds regardless of the nature of the observed data and the context in which it was produced, different types of data obviously require varying models of the data generating process. Without a good theoretical model of how the observed outcome can discriminate between different latent positions, it simply becomes unclear what exactly it is that is being scaled. The spatial model of politics is probably the model most widely used to

relate behaviors – including textual behaviors – to positions. While it obviously is not the only possible model of any single data generating process, it is highly appealing because it is well tested through years of refinement. In scaling techniques that rely on variations of the spatial model, actors are assumed to choose the outcomes that are most closely aligned with their ideal point (their political preference).<sup>2</sup> For example, donors contribute to campaigns of candidates, legislators cast their vote for policies and Facebook users follow political pages – and all choose the ones that are most closely aligned with their preferences. Actors receive monotonically decreasing utility from choosing outcomes (candidates, policies, Facebook follows) as they increase in distance from their ideal point. The process of choosing a candidate, a roll call vote or a page follow, however, is inherently random in nature, which is typically modeled with some distributional assumption (for a more thorough review, see Armstrong et al., 2014).

When scaling the political positions of a corpus of texts, similar assumptions are needed, and the spatial model generalizes well to this setting. Here, we can view the choice of words as the outcome. Whenever certain statements are associated with particular political positions, we can use them to discriminate between positions in a certain political space. In other words, the use of a particular (set of) word(s) provides us with a revealed preference for a specific (kind of) policy. Whenever we can think of the data generating process in these terms, the spatial model of politics is likely to provide a good approximation, and scaling a set of documents might be feasible.

### ***Which Assumptions Are Needed to Scale a Text?***

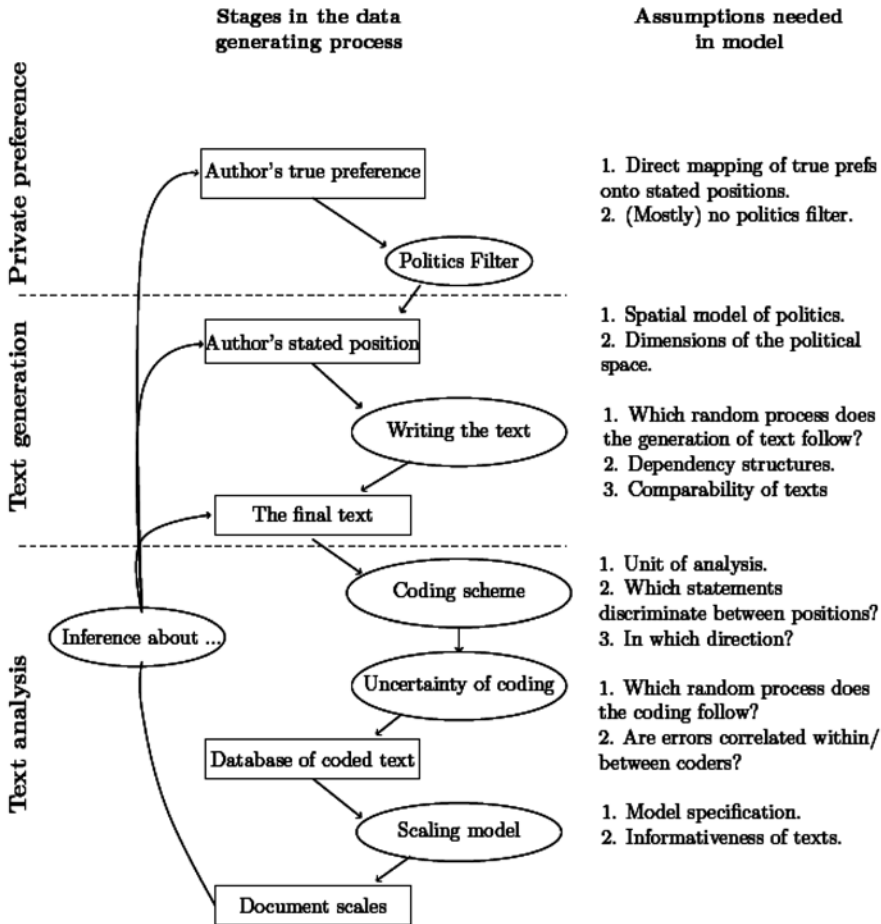
To infer from the text in a document to a political position, we need assumptions about three stages in the data generating process,

which together operationalize the spatial model of politics in the setting of text scaling. First, we make assumptions about the author's private preference and intent with the analyzed document. The second set of assumptions is about how those preferences are expressed in any particular document, and how that document relates to the others in the corpus, which in combination make up the relative political space we want to estimate.

While the first two sets of assumptions are about modeling the causal process that relates text generation to political position, we also need the third set of assumptions to translate text into data and from data to scales. Specifically, we need a statistical model that operationalizes the considerations in the underlying causal model about how preferences are communicated in text, and how this relates to the latent position of the document. Table 27.1 presents the stylized version of the full text and data generating process proposed in Benoit et al. (2009).

**True Preferences and the Politics Filter:** First, even though it is called text scaling, what we most commonly want to draw inferences about is not the political position articulated in the text, but the preference held by the author. But if the cost to articulating a position is low, authors might engage in cheap talk. Conversely, if costs are high, they might choose not to articulate the position for strategic reasons. All of the techniques we review here assume that authors do not censor their statements for political reasons. Therefore, we sideline this discussion, even though it obviously can cause significant measurement error.

**Model of the Political Space:** Second, we need to make assumptions about how any given author translates her position into text, and how that relates to the other authors in the corpus. Specifically, the language used in the texts must discriminate between the intended messages of different authors. In other words, the authors should receive varying levels of utility from their choice of words, and this



**Table 27.1 A stylized model of political text generation and model assumptions in each step**

Note: The stylized model of data generating process is based on Benoit et al. (2009).

variation should be related to the political space we want to measure. If authors of different preferences receive the same utility from similar choices of words, we cannot use the texts to discriminate between their positions. The documents should be informative about the political differences we seek to estimate (Slapin and Proksch, 2014). Particularly in contexts where there are strong common norms about how to phrase a document (as with highly technical legislative or legal documents) or the texts do not communicate any preference at all, it can be difficult to scale documents. An interesting special

case of incomparability is when authors simply use different languages. Importantly, De Vries et al. (2018) have shown that – at least when it comes to topic models – automated translation software (e.g. Google Translate) provides a good way of making languages comparable.

With regard to the relation between documents, a set of texts is only scalable if the texts can be placed in the same Euclidian space. This is an often undiscussed assumption, and it can be broken in three ways. One way is if political preferences are discrete views, not matters of continuous differences.

Another violation would be if the language used in the documents is incomparable in the way meaning is ascribed to words. Analyzing text that is produced under very different conditions or in varying contexts; that is from different time periods or actors; or that has very different audiences in mind would make it difficult to place the texts relative to each other – let alone in the same space (Slapin and Proksch, 2014). Finally, we need to make assumptions about the dimensionality of the space.

**Stochastic Writing Process:** Third, whereas the particular spatial model constitutes the assumptions about the systematic component of the data generating process, scaling also requires assumptions regarding the stochastic part. An important assumption in most scaling techniques is that the analyzed units are conditionally independent. When the unit of analysis is word frequencies, this is labeled ‘the bag of words’ assumption (Grimmer and Stewart, 2013). Conditional on the statistical model, any relation between the use of words is purely noise, and their ordering is inconsequential to the positions that we obtain through our scaling techniques. This assumption allows us to determine differences across texts based on the relative frequencies with which words occur in a text. The main difference in the scaling techniques we review in this chapter is how these frequencies are converted into positions. While this assumption is certainly wrong, it provides an effective simplification, and often scaling techniques work well despite the obviously wrong assumption (Grimmer and Stewart, 2013).

Finally, the writing process is inherently stochastic: the same writer would not write the same text if she sat down to communicate the same message repeatedly. Even if a perfect model of political text could be constructed, and no assumptions were violated at all, the randomness of the writing process would still produce uncertainty in the resulting position estimates. We cannot observe all possible texts the actors could have written, but we can estimate the variance in their

positions. While this is clearly not the same as uncertainty, it can help us identify a range within which the author’s intended message is likely to be.

**The Model Specification:** Finally, each assumption about the data generating process has to be incorporated into a combined scaling model. This requires assumptions about how the use of words is related to the latent policy position through a functional form and a statistical distribution. This is the step which operationalizes the theoretical considerations in the text generation process. If we use frequencies of individual words, we need to specify our expectations about how a change in frequency helps us discriminate the underlying position. For example, is the relationship log-linear? How does utility decrease as words are further removed from the author’s ideal point? We should also consider which other parameters (such as controls, prior information, etc.) to include in the model. Together, the assumptions about functional form, distribution and relations between words represent principal statistical assumptions, which implement the theoretical model of text generation.

Most of the estimators which we will discuss here are highly greedy in their data requirements, however. Thus, an additional and crucial assumption is that the texts to which we apply scaling techniques contain sufficient information for the technique to pick up differences across texts. The limiting factor is of course how words are distributed in the document-term matrices which are the analytic underpinnings of the methods we discuss. There have to be different distributions of words, and these distributions have to be meaningfully linked to the latent dimensions that we are trying to estimate.

## Summary

In this section, we discussed how *text* scaling in particular relates to the broader discipline of estimating latent political preferences.

We saw how this requires assumptions about the political context, and how it shapes the manner in which authors state their preferences through text. We discussed which assumptions are needed to implement a theoretical model of the political space in the context of text data, and how this forms the basis of scaling the positions of a set of documents. It is worth reiterating that while all of these assumptions are wrong, the model might still be useful in the sense that the estimate of the latent variable may correlate well with the true political preference of an author.

In the next sections, we review the scaling techniques that have been used by political scientists. We relate each technique to the spatial model of politics, and discuss its assumptions about how observed text is related to the author's latent position. This continuously reminds us that while most of the *visible* discretion when scaling a text is contained in the choice of unit of analysis and preprocessing, the choice of even off-the-shelf scaling techniques involves making a number of model assumptions.

## USING MACHINE LEARNING TO SCALE DOCUMENT POSITIONS

Before analysis we need to make the choice of scaling technique, each of which embodies a set of assumptions about the data generating process. While most techniques resemble statistical ideal point models, we can further distinguish between supervised and unsupervised models. Supervised models use human input, typically in the form of a set of training texts. These estimates can then be used to predict the positions of texts the model has not encountered previously. The training set also serves to define the policy space that the researcher seeks to estimate. Unsupervised techniques simultaneously learn about the latent space and estimate document positions in it, without input from the researcher.

When using any kind of computational technique, preprocessing of the documents is the first step in the modeling process, as it involves making the decision about where in the text we expect the signal about policy positions to be located. A first problem that any scaling needs to deal with is how to translate words into numeric values. That is, what is our unit of analysis, and how do we quantify the prevalence of certain phrases? The techniques we review in the following sections have traditionally used counts of individual words (unigrams) as their units of analysis. Pairs of consecutive words (bigrams) and words in all possible three contiguous sequences (trigrams) have also been used in political science. But one could use any length of word string (n-grams), which are more common in the broader field of natural language processing. Additionally, noise is often sifted out by removing extremely common words (so-called stopwords), numbers, punctuation and symbols as well as by reducing words to their stem. Especially for unsupervised models, it can be a good idea to remove highly infrequent words. It is extremely important to note that all of these choices regarding preprocessing have consequences, and should not reflect some standardized procedure but rather be seen as a step in building the model (Denny and Spirling, 2018). The output of this is typically a document-term matrix, where documents are identified along the rows, terms (words, bigrams, n-grams etc.) are in columns and the frequencies are in the cells. This is the input on which we estimate our statistical scaling models.<sup>3</sup> We use the *quanteda* package (Benoit et al., 2018), which is available in R and comes with an excellent online tutorial that walks the reader through each step in a computational scaling model.

An important caveat when using ideal point models for scaling is that even if all assumptions held, and we could think of a perfect statistical model of text, the incidental parameters problem – that there are too many parameters to estimate relative to



observations – would still mean that our position estimates would be wrong. In this regard, it is important to note that the use of computational techniques is no substitute for careful reading of the text and understanding of the subject matter. Automated scaling serves to amplify human ability, not replace it, and its use should be subject to careful subject-specific validation (Grimmer and Stewart, 2013).

## SUPERVISED TECHNIQUES: WORDSCORES

The Wordscores algorithm is a one-step approximation of a reciprocal averaging estimator for correspondence analysis on words (Lowe, 2008). Originally developed by Laver, Benoit and Garry (LBG) (2003), it was pioneering because it was one of the first attempts to introduce computational scaling techniques to a wider political science audience. And the model has been hugely successful for a number of reasons. First of all, the model is easy to implement because the authors made their software available to the wider public. Second, it relies on prior information in the form of reference texts with known positions. This makes it very stable compared to, e.g., unsupervised techniques. This also partially defines the latent political space before estimation, which makes it extremely flexible. Third, the algorithm is very clear and simple, which further broadens the group of potential users.

### *The Wordscores Model and Assumptions*

Wordscores begins from the premise that we have access to a set of texts  $R$  with known positions on the dimension we are interested in. Hence the precondition for a Wordscores model is that we have reliable and valid measures of the positions in a set of reference texts. Wordscores works through the core

assumption that each word  $w$  has a specific political position, and that the position of a document can be found by averaging over these word scores. The simple idea is that if we first ascribe positions to each word  $w$  by observing their frequencies in our reference texts  $r$ , where document positions are known, then we can use those word scores to predict the positions of out-sample texts by simply observing frequencies of words that also occurred in the reference texts. This is done by developing a measure of the probability of observing a given word  $w$  in our reference texts  $r$ , and using this to infer the positions of a set of out-of-sample texts from their word frequencies.

Specifically, LBG propose to calculate a score  $S$  for each individual word in the text using the following equation.

$$S_{wd} = \sum_r P_{wr} \cdot A_{rd}, \quad (1)$$

where  $P_{wr}$  is the probability ( $P$ ) of word ( $w$ ) occurring in text ( $r$ ), and  $A_{rd}$  is the position given to reference text  $r$  on the dimension  $d$ . Now we have values for each word in our text and we can therefore use the in-sample word scores to infer the position of the out-of-sample texts by using the frequencies of the word, whose positions we know:

$$S_{vd} = \sum_w F_{wv} \cdot S_{wd}, \quad (2)$$

where  $F_{wv}$  is the word frequency in the out-of-sample texts  $v$ . Lowe (2008) outlines the conditions under which bias in policy positions estimated in this fashion is minimized:

- 1 Positions of reference texts are equally spaced and extend over the range of the positions of individual words.<sup>4</sup>
- 2 Positions of individual words in the reference texts are equally spaced and extend *past* document positions in both direction.
- 3 All words are equally informative.

While it is obvious that the first two conditions cannot hold simultaneously in any

real-world setting, they provide guidance when choosing reference texts in a way that minimizes bias. Specifically, the conditions suggest that there should be sufficient overlap between distributions of words in the reference texts, and that they should include a sufficient range of potential word positions in the out-of-sample texts.

As mentioned previously, a strong assumption when scaling in general is that the vocabulary does not change radically over texts. When using Wordscores alongside a good choice of reference texts (defined by the above conditions), estimates are generally less sensitive to differences in the meanings and uses of words. We illustrate this point later.

### ***Using Wordscores***

To illustrate the use of Wordscores, we draw on data from Baturo and Mikhaylov (2013), who use speeches by Russian governors to estimate levels of alignment with Putin and Medvedev, respectively. By leveraging the fact that the main policy dimension in a Wordscores estimation is defined a priori through the use of reference texts, they are able to estimate where each governor's address to the local parliament falls on a scale from Medvedev to Putin. This use of prior information is what makes supervised techniques like Wordscores extremely flexible. In terms of the spatial model of politics, we can think of the underlying policy space as one in which two leaders compete for control, and state slightly different policy preferences. The assumed utility function of the authors is one in which the governors prefer to converge on the policy position of the most powerful national leader. Thus, we have defined a coherent policy space, and have a clear idea about how written words reveal a preference. In combination, this provides us with a foundation for mapping words onto a latent position in this particular space. This is an interesting case, in part because it shows how broadly we can construe the spatial

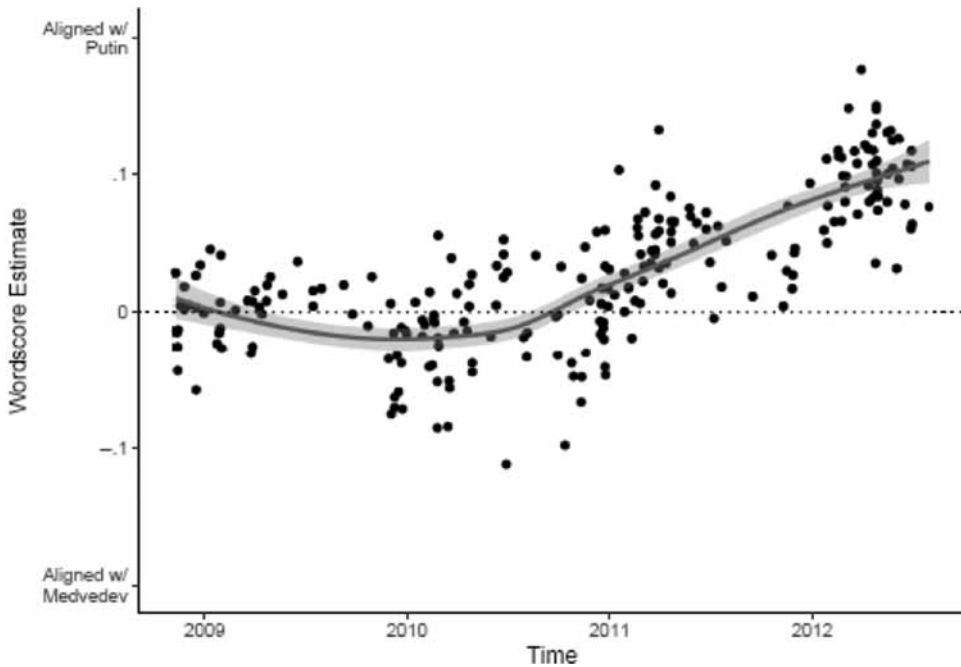
model of politics. An additional interesting feature is that it does not necessarily assume that authors communicate their sincere preferences.

We use the texts made available by Baturo and Mikhaylov (2013), which exclude segments on foreign policy. Otherwise, the only preprocessing we do is to remove punctuation. We set the reference scores of Putin and Medvedev to be 1 and  $-1$ , respectively. Thus, we fully replicate the original study. Figure 27.1 shows how alignments estimated through Wordscores changed over time.

To validate the Wordscore estimates, we follow Baturo and Mikhaylov (2013) and use monthly expert evaluations of how powerful Putin and Medvedev are, respectively. To get a direct estimate of partisan alignment, we compute the difference between the Wordscores estimate and Putin's reference position. We use the monthly averages of this difference to facilitate comparison with the expert survey. Figure 27.2 shows the results. It is clear that when the average governor's speech is more aligned with Putin, expert perception of Medvedev's influence is lower. The correlation is strong and precisely estimated. The correlation between expert perception and the alignment of governor speeches with Putin's position is somewhat lower. This is likely because there is relatively little variation in both estimates of Putin's influence – he remains continuously powerful by both estimates.

### ***Extensions to Wordscores***

The original incarnation of Wordscores was strong in its simplicity, but made many of its assumptions implicitly. Lowe (2008) clarifies the underlying assumptions and provides a statistical model for Wordscores, which also serves to relate it to the broader family of statistical ideal point estimators and correspondence analysis. Perry and Benoit (2017) implement a scaling technique using an affinity class model, which is highly similar



**Figure 27.1 On a scale from Medvedev to Putin**

*Note:* Each point represents the Wordscores estimate of a governor's speech at a given point in time. The dimension is identified using the most recent parliamentary address by Medvedev (reference score = -1) and Putin (reference score = 1), respectively. The solid line is a Loess smoother, and the shaded region is a 95% pointwise confidence interval.

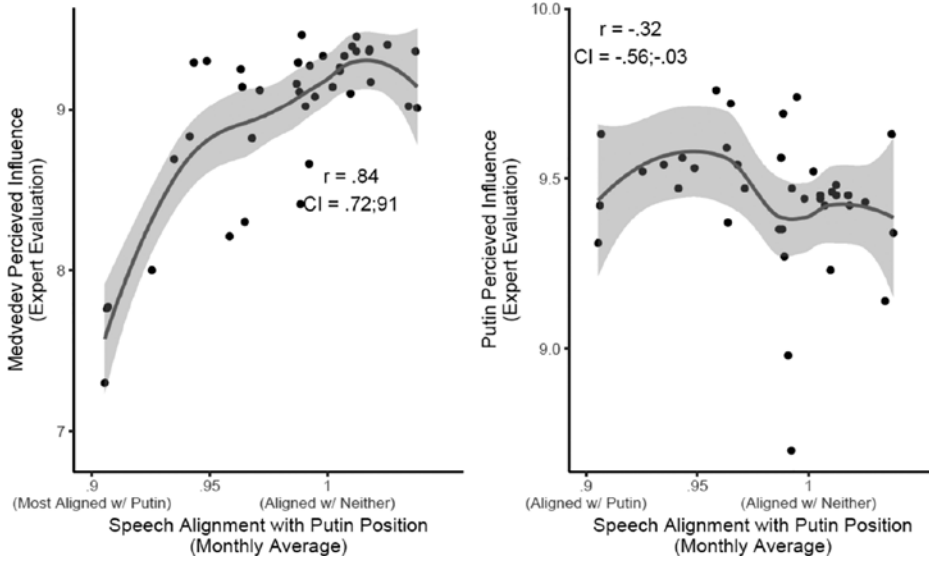
to the original Wordscores model, and solves some of the problems identified in Lowe (2008).

A second issue that has drawn some attention is how to transform the raw scores obtained by the procedure described above to the same scale used to score the reference texts. In their original article LBG assume that the raw scores for the reference texts have the correct mean but that the variance is incorrect. Lowe (2008) argues that this assumption might lead to biased scores because of the shrinkage discussed above. Martin and Vanberg (2008) criticize the original transformation, arguing that there the original transformation is dependent on the choice of reference text, leading to the uncomfortable position that any desired result could be obtained provided that the right combination of reference texts are chosen. Consequently,

they propose a transformation of the raw scores which builds on the relative distance ratios using two anchoring texts which serve as the unit in relation to which all other positions are expressed. Lowe (2008) argues that researchers are then confronted with a choice when deciding which transformation to use. The original LBG transformation is dependent on the reference texts and is indifferent to the virgin texts; the opposite is true of the Martin–Vanberg transformation.

## UNSUPERVISED TECHNIQUES: WORDFISH

Wordfish (introduced in Slapin and Proksch, 2008) is an unsupervised machine learning algorithm, which is based on a Poisson item



**Figure 27.2 Perception and governor speeches**

Note: The figure shows the correlations between expert perception of Medvedev’s and Putin’s power, respectively, and the Wordscores estimate of the average governor’s alignment with Putin. The solid lines are Loess smoothers, and the shaded regions are 95% pointwise confidence intervals.

response theory (IRT) model (Lowe, 2015). Being unsupervised, it simultaneously estimates policy positions and learns the policy space using only the texts provided, with no external information in the form of virgin texts or anchoring (Grimmer and Stewart, 2013). While this is a strength in many aspects, it requires strong modeling assumptions and presents challenges, particularly in regard to validation of the particular policy scales and the dimension as such (Grimmer and Stewart, 2013; Lowe and Benoit, 2013). In this section, we briefly introduce the statistical model underlying Wordfish, its assumptions and they can be broken, and how a Wordfish model can be estimated and interpreted.

**The Wordfish Model and Assumptions**

The Wordfish estimator assumes the data generating process to be as follows:

$$y_{ij} \sim \text{Poisson}(\lambda_{ij})$$

$$\lambda_{ij} = \exp(\alpha_i + \psi_j + \beta_j * \omega_i)$$

Where  $y$  is the count of word  $j$  in the position document of actor  $i$ .  $y$  is assumed to be drawn from a Poisson distribution and connected through its mean,  $\lambda$ , to the systematic component, where  $\omega$  is the estimated position of document  $i$ .  $\psi$  is a word fixed effect, which signifies the frequency of word  $j$ , when a document expresses a center position on the estimated scale (the difficulty parameter in IRT language).  $\beta$ , a word’s weight in estimating positions, is an estimate of how sensitive the use of a word is to the political position. In IRT it is called the discrimination parameter, because it measures how the latent position parameter changes in response to word frequencies. It is parallel to a variable’s loading in factor analysis (Jackman, 2001).  $\omega$  is the position of actor  $i$  as estimated through its position document. Finally, this leaves  $\alpha$ , a set of document fixed effects. ‘While

estimation was originally done by iterating over conditional maximum likelihoods, equivalent – but faster – implementations relying on expectation maximization have since been suggested’ (Lauderdale and Herzog, 2016).<sup>5</sup>

Wordfish requires a number of the previously introduced assumptions about the underlying spatial model. However, dimensionality in particular is key to the Wordfish model, and it uses a different statistical operationalization than Wordscores.

### *The statistical model*

Wordfish operates under the assumption that the generation of words in a text – conditional on the model – follows a Poisson process. This has consequences for estimation of both the latent position parameter and the uncertainty of all quantities in the model. First, regarding parameter estimation, it translates into an assumption about the functional form of the relation between word frequency and the latent parameter being log-linear. The model specification introduced above implies assuming monotonicity and that the weight of each word must be the same in any subgroup with the same latent parameter. The former would be violated in situations where word weights do not always increase or decrease with the latent parameter. The latter violation occurs if two groups with the same policy position use the same word differently.

Second, assuming a Poisson process implies that the variance of the rate of the word count is equal to its expected rate. This assumption will be broken in the presence of both under- and overdispersion as well as structural zeros (when there is zero probability of a word occurring in a text). This induces well-known problems of underestimation of the uncertainty in Poisson models for count data (King, 1998), which translate directly into the Wordfish setting (Lowe and Benoit, 2013). In practice, violations of the distributional assumption will also lead to poor coverage (Lowe and Benoit, 2011).

### *(Uni)dimensionality of the policy space*

While all scaling techniques require assumptions about dimensionality, the unsupervised ones – like Wordfish – are particularly vulnerable, because they learn about the policy space without input from the researcher.<sup>6</sup> If a researcher misspecifies dimensionality, she risks estimating a policy dimension which is either meaningless or not the one she is interested in. There are a number of reasons why this is a risk. When the generative model specifies a unidimensional policy space, when it really is multidimensional or when the word weights are misspecified, we risk misspecifying the policy dimension. But even if all modeling assumptions hold, the dimension identified by Wordfish might simply be wrong. First, word use in texts addressing the same concerns are likely to be highly correlated. Wordfish will recognize differences in word use between two texts as indicative of their different political positions, but in reality these differences could be due to the topics addressed by the authors (Lowe and Benoit, 2013). A notable such case is in situations where texts use completely incomparable language or do not address similar topics at all. In these situations they cannot be scaled together, and if they are, it will often result in the main policy dimension being misspecified. Finally, the Wordfish estimator’s likelihood function is prone to have many local minima, and estimation can easily get stuck in an uninteresting one. This problem could be compounded if there is not enough data for the curvature of the likelihood to be estimated correctly. In this situation, the algorithm might capture noise and a meaningless policy dimension.

### *Using Wordfish*

To illustrate the use of Wordfish, we investigate its performance in estimating the policy positions of European interest groups on

three specific issues. Here, we draw on data from Egerod (2016). We use texts from the European Commission's online consultations regarding *Reinforcing Sanctioning Regimes in the Financial Services Sector*, *A New European Regime for Venture Capital* and *Review of the Investor Compensation Scheme Directive*. We will refer to them simply as Sanctions, Venture Capital and ICSD, respectively. For the present purposes, we include only a subset of the interest group responses. Below, when we examine the consequences of violated assumptions, we include all groups. Alongside the interest group position papers, we include the Commission's original Green Paper, which outlines the issues within each consultation, and the final policy proposal.

The main fracture between the interested parties in all three consultations was whether the EU should impose *more* or *fewer* rules. Therefore, we can think of the underlying spatial model as one in which actors are placed along a continuum ranging from wanting more to fewer supranational rules. This is the underlying political space in which we wish to place actors. To gauge Wordfish performance, we compare its estimates to hand-coded positions, which aim directly at capturing this space. See Egerod (2016) for more information on the hand-coding.

To prepare documents for Wordfish scaling, we reduce words to their stem and remove stopwords, numbers and punctuation. Figure 27.3 shows the positions estimated through Wordfish, and how they correlate with the hand-coded positions for documents in each of the three online consultations we investigate here. The two correlate highly in all three cases, although by far most strongly in the case of National Sanctions and clearly the least in the Venture Capital case. To save space, we do not discuss the reasons for discrepancies between automated and human scaling which are present in all three cases.

We can use the word weights, or  $\beta$  parameters (i.e. the word discrimination parameter),

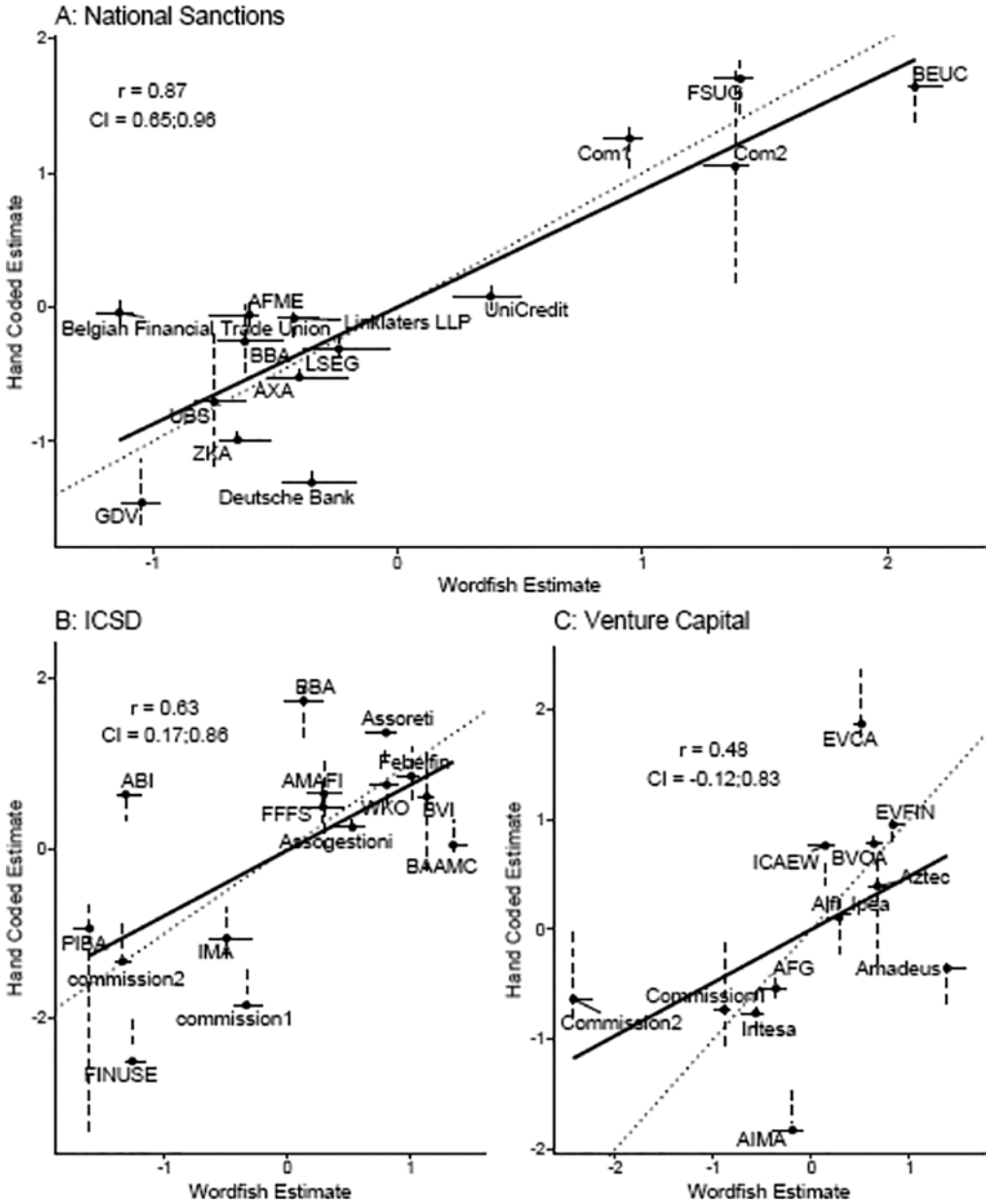
to analyse the substantive content of the dimension recovered by Wordfish. This can potentially be used to explain why the two sets of scales diverge for some documents. Figure 27.4 shows the 21 words with the, respectively, highest and lowest weights for each consultation.

We can take the National Sanctions consultation as an example of how to diagnose divergence between human and machine based scales. There, we can observe that 'labour', 'clause' and 'employ' all have very negative weights – far out in the tail of the full distribution of word weights. This can help us explain why the Belgian union of employees in the financial sector is estimated to be a stronger advocate for fewer EU rules. A thorough reading of the union's position document reveals that, while it is relatively positive overall, it expends many words on strongly arguing against employees of financial institutions being liable to prosecution when laws are broken. This seems to be the aspect Wordfish has caught.

### **Extensions to Wordfish**

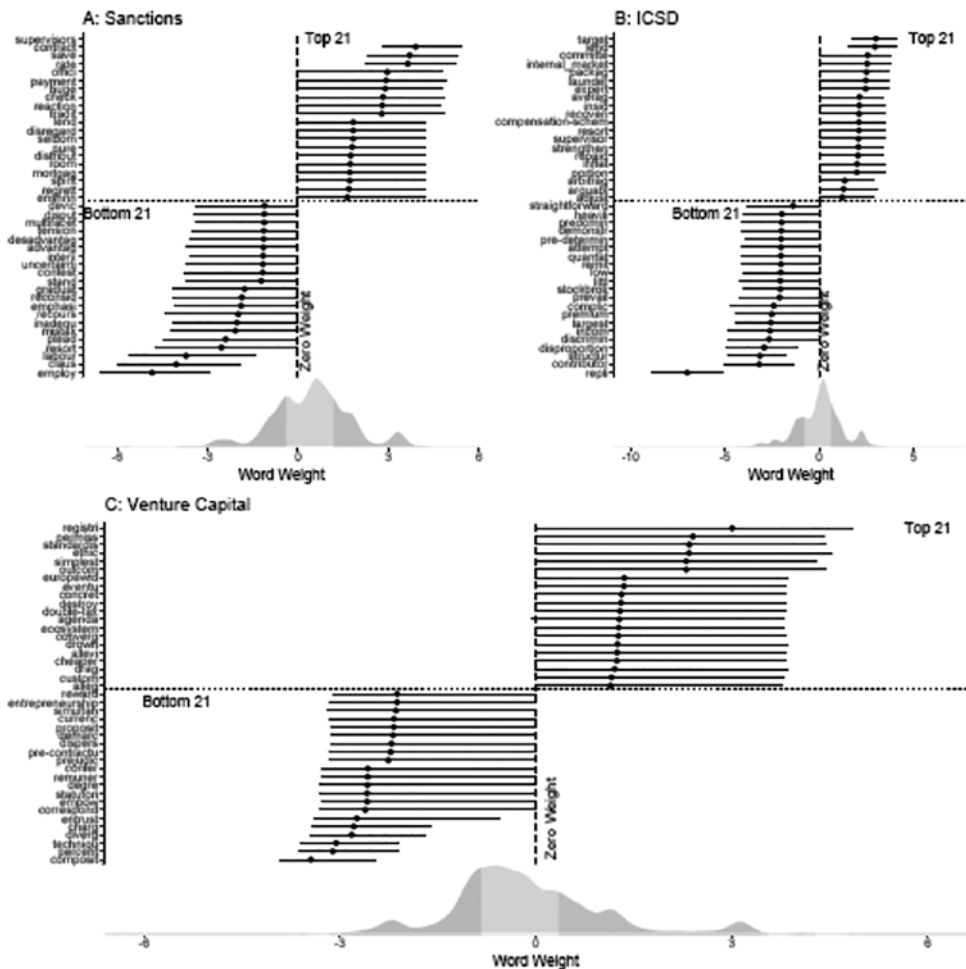
While Wordfish builds on item response theory, it is closely related to correspondence analysis (like Wordscores) (Lowe, 2008). Therefore, correspondence analysis will often provide similar position scales at a lower computational cost. In an early implementation of the Poisson scaling model, Monroe and Maeda (2004) use a Bayesian setup to estimate a two-dimensional model. This is one way of dealing with some dimensionality issues. Slapin and Proksch (2008) have done so by manually separating out the parts of a text that are most closely aligned with predefined dimensions.

Besides these issues of dimensionality, there are a number of relevant extensions to the Wordfish model itself. Lowe and Benoit (2013) introduced the use of asymptotic standard errors, instead of the very computationally intensive Poisson bootstrapped



**Figure 27.3 Validating Wordfish estimates against a human benchmark**

Note: Horizontal lines around points are 95% parametric Poisson bootstrapped confidence intervals (CIs) around the Wordfish estimate. Vertical dashed lines are 95 percent CIs from non-parametric bootstraps of the hand-coded scales. 500 resamples used. The solid line is the best linear fit; the dotted diagonal line shows what a perfect fit would look like.



**Figure 27.4 Which words define the policy space?**

Note: Points show the point estimate for each word weight. Lines are 95 percent CIs based on 500 resamples from a conditional Poisson distribution. Density plots show the marginal distribution of word weights in the full corpora. Dark shaded areas are below the 25th and 75th percentiles, respectively.

standard errors used in Slapin and Proksch (2008). The Lowe and Benoit (2013) implementation also allows for varying levels of dispersion. Both the analytical and the original technique, however, rely heavily on the model being correctly specified. As a way of obtaining uncertainty estimates with weaker assumptions, Lowe and Benoit (2013) also introduced a non-parametric bootstrap procedure. The `quanteda` package in R supplies functionality for random sampling of

words, which can be used to implement the bootstrap with relative ease. Lauderdale and Herzog (2016) deal with problems of comparability between corpora by using `Wordfish` to estimate issue specific positions, which they aggregate using Bayesian linear factor analysis to get estimates of overall ideology from text data.

Finally, Lo et al. (2016) exploit the fact that as the rate of failure converges to the limit, a Poisson distribution can be reparameterized



as a negative binomial one. This allows them to incorporate a document level dispersion parameter, which can be interpreted as the clarity associated with a document's stated position.

## WHEN ASSUMPTIONS ARE BROKEN

Assumptions about the generative model and the use of prior information vary between models. Thus, when one model is obviously misspecified and performs poorly, a researcher can choose another, more suitable one. However, it is not well understood how to handle violations of assumptions that are common across models, or how to proceed when no algorithm performs satisfactorily. In this section, we will investigate the consequences of violations of two basic assumptions that all common scaling techniques rely on: the comparability of language use and the length of the texts.<sup>7</sup>

We use real-world texts and simulate changes to the corpora to quantify the effect of marginal changes to document length and word use. This allows us to inspect what happens when core assumptions are broken in realistic but controlled settings.

### *Language Differences*

When we previously illustrated the use of Wordfish by using position documents from interest groups in EU consultations, we only relied on a subset of the actual position papers, consisting of 14, 13 and 15 documents, respectively, in the National Sanctions, Venture Capital and ICSD consultations. In reality, however, each corpus consists of 42, 44 and 57 documents from a very wide range of different actors, ranging from individuals, through government branches and NGOs, to different kinds of corporations and capital funds. These three corpora present an extremely hard test for both the Wordscores

and Wordfish algorithms, in particular regarding dimensionality and the comparability of the authors' use of words.

To gauge the impact of language differences, we ran both scaling techniques several times on different subsets of consultation documents. We began by running them on all documents in each consultation, then we randomly removed five to eight documents. In each consultation, we chose not to remove the type of actor which was most active in the consultation, which ensures that the included documents become more comparable with each iteration. In the case of Sanctions, we only removed non-corporations. In Venture Capital, we removed documents from actors that were not venture capital funds. In the case of the ICSD, we removed all documents other than those from national employer associations. After numerous iterations, this left only one or few types of organization and the Commission. The strength of this framework lies in its approximation of counterfactual scenarios – as documents are removed in a semi-random way, and the consultations otherwise remain the same, we hope to estimate the causal impact of altering the composition of the different corpora. Additionally, the dimensionality of the policy space is close to pre-defined by the Commission in its original policy paper, since it directs interested parties to comment on specific topics.

To save space, we do not show the performance of scalers within each iteration. However, we do find that both Wordfish and Wordscores are highly sensitive to the subset of documents being used, and that both perform best in the smallest, most homogeneous sets of texts. To quantify the degree to which these improvements are driven by decreased differences in word use, we use the correlation between the recovered scales and hand-coded positions as the dependent variable in two linear regressions – one for each algorithm. We measure differences in language with two proxies: the average correlation in word use and the number of unique words

in each iteration. Because other quantities change besides similarity of the documents, we include as controls the number of positions to be estimated, the average document length, an inverted Herfindahl-Hirschman index capturing how many different types of interest group were included in the estimation and fixed effects for consultations. The results for our variables of interest are presented in Figure 27.5.

As we can see, the improvements in algorithm performance follow predictable patterns. For both algorithms, the correlation between the human benchmark and the computer-based scales decreases by almost .1 for each percentage point the number of unique words increases. Note that because we control for average document length, the increase in the number of unique words captured here is for an unchanged document length. For Wordfish, performance improves by more than .1 every time the average correlation of word use in the corpus increases by .1. This effect is smaller for Wordscores, where the improvement in performance is .02. While the latter estimate is not statistically significant, it is still a strong correlation.

This illustrates that as word use in documents becomes more dissimilar, automatic scaling becomes less feasible. The fact that Wordscores is less vulnerable to differences in word use correlations illustrates a key difference between the two algorithms. Wordfish relies heavily on documents addressing the same concerns using the same words. If they do not, the algorithm is likely to pick up differences in the topics the authors address, not in their political positions. For Wordscores, performance relies more on the reference texts being representative of the broader universe of texts in the corpus. As long as that is the case, differences in word frequencies matter less (although they are not irrelevant), but as they become less representative (e.g. because the number of unique words increase), the performance of Wordscores decreases markedly.

### ***Document Length and Informativeness***

To get at the effect of document length on the performance of scalers, we use a text corpus where we know that Wordscores and Wordfish provide good estimates of policy positions – the Lowe and Benoit (2013) data on parliamentary speeches during the debate on the 2010 budget in the Irish Dáil. This data also includes estimates of the position of each speech based on human judgment. We simulate changes in document length by randomly reducing the number of times an actor articulates each word in the corpus by between 0 and 3. We reduce the word frequencies in the corpus in this way 100 times, estimating both Wordscores and Wordfish models in each iteration. To measure the performance of the algorithms, we predict the expert coded benchmark using scales recovered from each algorithm and compute the root mean squared error (RMSE). Because reductions in word frequencies are random, we can get estimates of uncertainty through randomization inference repeating the entire process 100 times. The results are presented in Figure 27.6.

The results show that the performance of Wordfish and Wordscores decreases dramatically as the average document length decreases from the baseline of approximately 3,900. To begin with, scales from both models predict the positions of speeches with a relatively small RMSE of between .3 and .4 – corresponding to approximately one-third of the standard deviation. The error increases quickly and stabilizes at about 80% of a standard deviation for Wordfish, when it hits an average document length of 1,800 words. For Wordscores, the RMSE stabilizes at approximately 75 percent of a standard deviation, when the average document length is around 700.<sup>8</sup> With an error of that size, the recovered scales are close to useless.

As in the previous simulation, the use of reference texts to guide the estimates of the Wordscores model proves an important

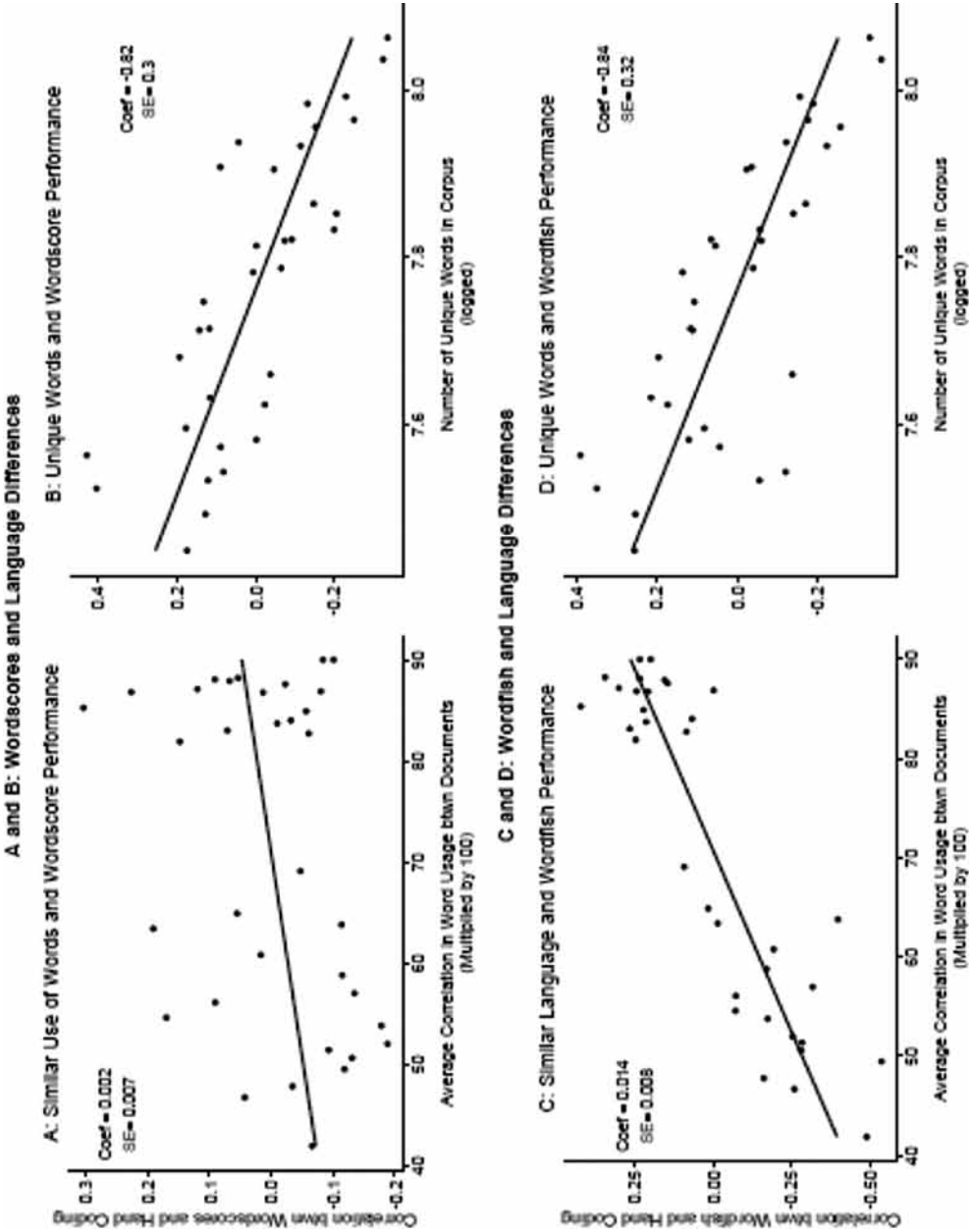
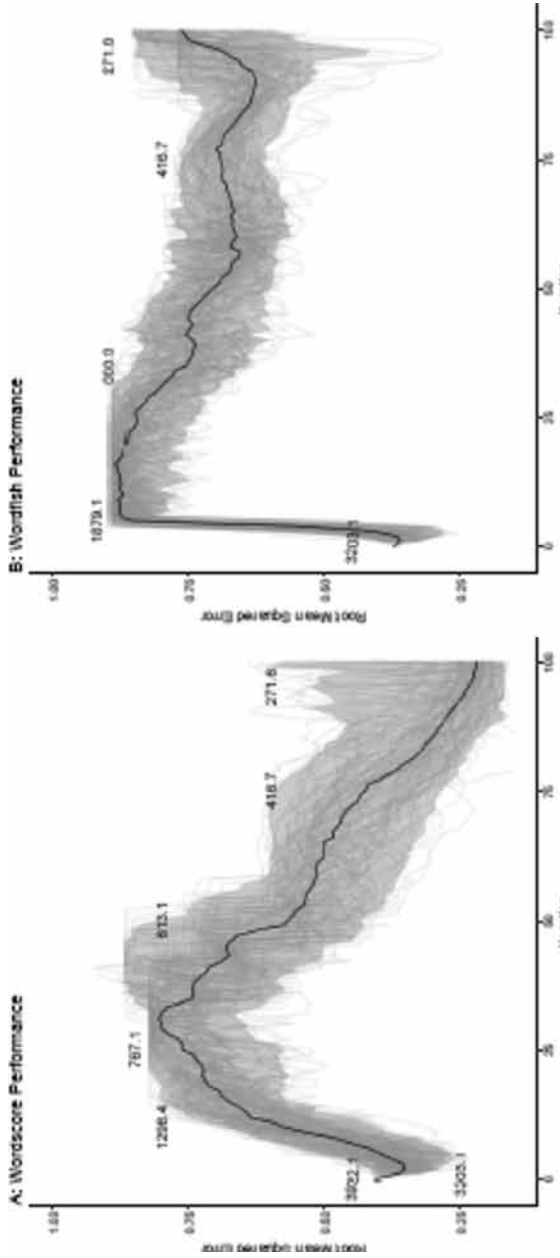


Figure 27.5 How the performance of scaling algorithms vary with the comparability of texts



**Figure 27.6 Document length and performance of scaling models**

Note: In each iteration, we reduce the frequency with which each actor uses a word by a random integer between 0 and 3. We run 100 iterations, and rerun the algorithm 100 times to get uncertainty. Shaded lines represent the root mean squared error in each iteration, the solid line shows the average across different random removals. The shaded area is the 95% confidence interval.

feature. As the average document length decreases below 600 words, Wordscores' performance actually starts to improve again. This happens because the word usage in the reference texts becomes more representative as the length of the left-out texts decreases. While this obviously hinges on a good choice of reference texts, it suggests that the performance of the Wordscores model is a non-monotonic function of document length. With the right reference texts, the algorithm may perform equally well in small and large corpora. The obvious caveat is, of course, that because the texts are very short in the final 20 iterations of each chain, the uncertainty around the average root mean squared error is relatively high.

### ***Guidelines for Constructing Your Corpus***

With the results from these two simulation studies, we can provide some tentative guidelines for researchers. While it is difficult to give precise advice beyond the particular cases we have investigated here, the results show that the performance of the two scaling techniques is strongly influenced by observable characteristics of the documents being scaled. While we probably cannot infer precise thresholds from these cases to the universe of potential texts to be scaled, it does tell us what we should be aware of when we construct those corpora.

In our cases, scaling documents when the average correlation between word frequencies is below .6 to .7 resulted in poor performance. Additionally, increasing the number of unique words beyond 2,000 without also increasing document length resulted in the recovery of biased positions. Regarding the length of the included documents, we found that scaling corpora with fewer than approximately 1,800 words in the average text is infeasible using both algorithms. For Wordscores, however, corpora consisting of very short texts (below 400 words on average)

can be scaled if the reference documents provide good coverage of the virgin texts.

The analyses also provide some insight into how these problems can potentially be handled. First, when there are many unique words relative to the average document length, it will often make sense to trim away very infrequent words, as this removes noise in the Wordfish estimation and improves the representativeness of the reference texts for the Wordscores algorithm.

Second, if texts are simply incomparable, it can make sense to redefine the population of interest to a more coherent group of texts, if it is possible given the particular research question. If estimation is done over time, it would make sense to split the sample up and run the algorithm within more narrow time periods. If the positions of many different types of groups are to be recovered, one can focus on the most relevant one and only include that one in the estimation.

Finally, if the average document is very short, and there is no way to increase the amount of data available for estimation, Wordscores seems to be able to recover good, but noisy, positions – under the important condition that the reference texts are representative of the remaining corpus.

### **WHAT IS THE ROAD AHEAD?**

In this chapter, we have laid out the conceptual foundations for scaling policy positions from hand-coded texts as well as automated content analysis. We argued that a spatial model of how policy positions relate to language use is a necessary condition for scaling texts, and that the choice of statistical model should follow from this conceptualization. Based on this framework for understanding text scaling, we outlined the necessary and sufficient assumptions for estimating positions from political texts. This highlights how important conceptualization and the accompanying statistical

assumptions are – which, yet again, shows that automated scaling can never replace human judgment, only augment it (Grimmer and Stewart, 2013).

We then introduced commonly used techniques for scaling text through supervised as well as unsupervised machine learning (the Wordscores and Wordfish estimators, respectively). Based on the conceptual framework outlined at the beginning of the chapter, we discussed the assumptions embedded in these techniques, and how they could be broken. We illustrated the use of these three scaling models with diverse corpora of political text.

We then proceeded to investigate what happens when two important assumptions are broken. By simulating random changes in two corpora of real-world text, we illustrated the consequences of estimating positions from texts that (a) use their words too differently, and (b) are too short. The results suggested that the performance of Wordscores is affected less by changing these factors as long as the reference texts are representative, and yielded some tentative guidelines about how to construct a corpus. The key takeaway is that performance varies systematically and according to observable features of a corpus. While the thresholds uncovered in the specific corpora investigated here may not hold in any given other setting, the findings can still inform researchers when they construct a corpus. The results illustrate how we can use observable features of a corpus to judge its suitability for scaling – we do not always have to rely on abstract argumentation.

In concluding, we will briefly discuss two potential venues for future research into text scaling which we think may be fruitful: (a) potential ways of improving our scaling models and (b) how we can think about measurement error when we incorporate scaled positions in econometric models. One goal of this diversity was to illustrate how broadly the spatial model of politics can be construed.

### ***Improving Scaling Models: Dealing with Comparability and Conditional Independence***

Throughout this chapter, we have repeatedly discussed two assumptions about the textual context: (1) conditional independence of word (or n-gram) frequencies, and (2) similarity in the way authors ascribe meaning to words (i.e. stability of the vocabulary). We believe that important new work on word and text embeddedness (Mikolov et al., 2013; Ng, 2017) may offer ways to deal with this within existing scaling frameworks. We will briefly discuss two such possibilities.

#### ***Conditional independence of words***

While most scaling techniques perform relatively well, in spite of relying on the simplifying ‘bag of words’ assumption, there is little doubt that it can hurt algorithm performance – and sometimes severely. Throughout the text, we have emphasized that using the frequencies of single words – while the political science standard – is not the only feasible level of analysis. Any n-gram could conceivably be used. This, however, could induce more noise than looking at single words. One way to think about the utility of word embeddings for scaling techniques is that they offer a statistical model for learning about word context – and, among other things, ways to construct n-grams in a principled manner. The iconic word2vec (Mikolov et al., 2013; Ng, 2017) technique, for instance, either uses the context of any given word to predict it, or uses that word to predict its own context. Either provides a set of probabilities that words co-occur, which in turn gives a foundation for finding the optimal n-gram, which could make the conditional independence assumption more valid.

#### ***Grouping most similar documents***

A well-known fact – which we substantiated through simulations – is that dissimilarity in word use and the ways in which meaning is

ascribed to individual words can severely harm the performance of scalars. Embedding texts in their contexts might help in this. If we think about dissimilarity of documents as a confounding factor, learning the context of documents may offer a way of modeling it directly. Seeing as position scales will be severely confounded with the most prevalent topics when documents are dissimilar enough, one way of doing this would be by using topic modeling to learn about the topics in documents before scaling, and then conditioning on them during the estimation of document positions. Only comparing documents that concern similar topics during estimation could be one way of estimating the policy positions of highly dissimilar actors on the same latent scale.

### ***Improving the Use of Scales: Systematic Measurement Error in Political Positions***

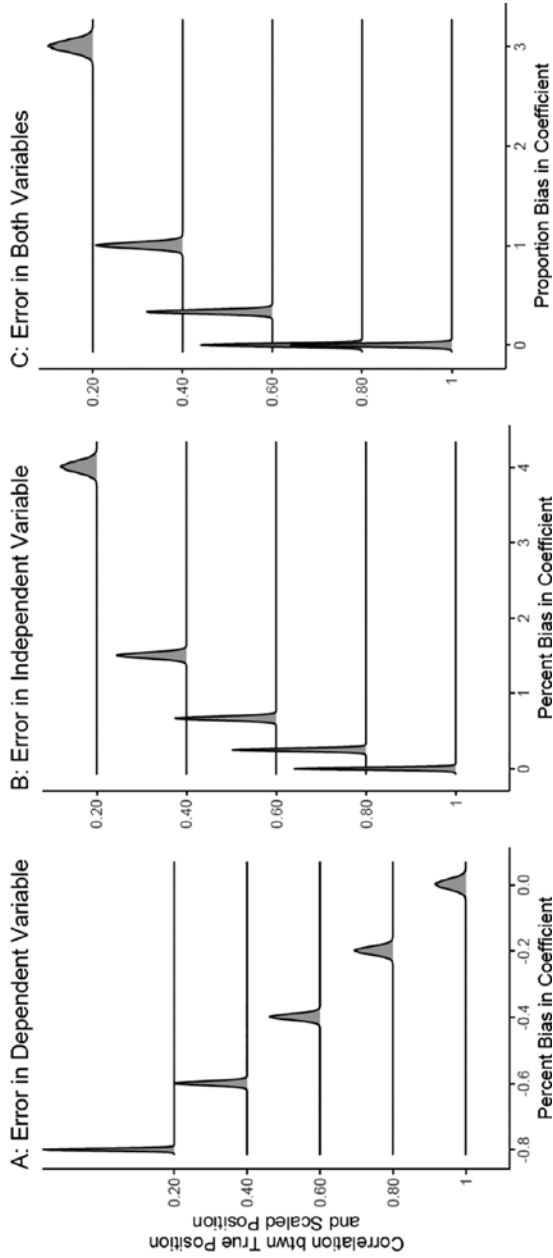
No matter how much we improve our models or how complex they become, they will always be erroneous approximations of real-world text generation. While it is relatively straightforward to deal with the special cases, when our models perform either dismally or in a superb manner, we need more research on how to handle the intermediate cases, where a model provides estimates that correlate with true positions but far from perfectly. This is especially pertinent when we use those position scales in econometric models seeking to estimate their correlates or use them to explain other political phenomena. The default solutions seem to be to either disregard the measurement error (or assume that it is random), or to disregard any results based on these imperfect scales – neither of which are satisfactory. Because position scales in any given case will be flawed, they will correlate imperfectly with the true positions of a set of actors. Thus, by definition, measurement error in the estimated scales cannot be random. But on the other hand, it

seems foolish to discard estimates that, while wrong, provide some useful information.

While a complete treatment of problems with measurement error is beyond the scope of this chapter, we will briefly illustrate how we can think about problems arising when including erroneous scaling estimates in econometric models. To do so, we conduct a number of Monte Carlo simulations. We use variations of a simple setup: we include a variable measured with some systematic error as either dependent, independent or both variables in a linear regression. We vary the measurement error in increments of 0.2, so the observed scales correlate with the true position estimates with  $r \in \{1; .8; .6; .4; .2\}$ . To keep things simple, we assume that all remaining Gauss-Markov assumptions are met. We run the three models in 10,000 random samples each including 1,000 observations. Figure 27.7 shows the distributions of bias arising from each scenario.

Because the setup does not necessarily generalize, the most important thing to note is the direction of bias: systematic error in the dependent variable biases the regression estimate downward, while the opposite holds when the independent variable is measured with error. There is reason to believe that the errors to some extent cancel out when both variables are measured erroneously. This conclusion, however, rests on the assumption that errors in variables on either side of the equation induce an equal amount of bias in the estimation. In this particular setup, that only seems to be the case when the independent variable correlates by more than .6 with the true concept. For lower correlations, bias induced by systematic measurement error on the right hand side is much larger in this scenario.

This provides an initial illustration of how measurement error in position scales may affect results downstream when they are included in econometric models. While the scenarios are general, the exercise shows that researchers should think hard about how measurement error in their scales may impact



**Figure 27.7 Error in scales and bias in econometric models**

Note: Each distribution shows the percentage difference between estimated coefficient and the true effect as measurement error varies. Each distribution is made up of 10,000 iterations of an ordinary least squares (OLS) regression, each with a sample size of 1,000 observations.



later stages of their research – not just ignore or discard them – and that more research into the topic might be needed.

## Notes

- 1 We thank Lucas Leemann as well as the editors Luigi Curini and Robert Franzese, who provided very valuable feedback on earlier versions of this chapter.
- 2 This is the so-called proximity based model of space. While a *directional* model obviously is possible as well, none of the currently implemented models use it (Armstrong et al., 2014).
- 3 While this is typically what we model, it is not the only conceivable form.
- 4 Additionally, in the statistical model for Wordscores proposed by Lowe (2008), where words differ in informativeness, text and word positions should be closely spaced relative to each word's discriminatory power (informativeness).
- 5 This is where the informativeness of the text corpus is particularly important: for estimation to be done, there has to be enough data for the curvature of the log likelihood to be approximately quadratic (Lowe and Benoit, 2013).
- 6 When using supervised techniques like Wordscores, the researcher to some extent defines the policy space herself through her choice of reference texts. This still entails an assumption about unidimensionality, which is obviously likely to be wrong, but if the reference texts are chosen well enough, the estimator is unlikely to estimate a policy space that is very different from the one the researcher is interested in.
- 7 E.g. Lowe and Benoit (2011) investigate what happens when distributional assumptions are broken.
- 8 It should be noted that the speed with which the error increases is in part driven by the fact that more words are deleted within each iteration in the beginning, because there simply is more content to delete.

## REFERENCES

- Armstrong, David, Ryan Bakker, Royce Carroll, Christopher Hare, Keith Poole and Howard Rosenthal. 2014. *Analyzing spatial models of choice and judgment with R*. CRC Press.
- Barberá, Pablo. 2015. 'Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data.' *Political Analysis* 23(1):76–91.
- Baturo, Alexander and Slava Mikhaylov. 2013. 'Life of Brian revisited: assessing informational and non-informational leadership tools.' *Political Science Research and Methods* 1(1):139–157.
- Benoit, Kenneth and Michael Laver. 2006. *Party policy in modern democracies*. Routledge.
- Benoit, Kenneth, Michael Laver and Slava Mikhaylov. 2009. 'Treating words as data with error: uncertainty in text statements of policy positions.' *American Journal of Political Science* 53(2):495–513.
- Benoit, Kenneth, Kohei Watanabe, Haiyan Wang, Paul Nulty, Adam Obeng, Stefan Müller, and Akitaka Matsuo (2018). 'quanteda: An R package for the quantitative analysis of textual data.' *Journal of Open Source Software* 3(30):774–778.
- Bernhagen, Patrick, Andreas Dür and David Marshall. 2014. 'Measuring lobbying success spatially.' *Interest Groups & Advocacy* 3(2):202–218.
- Bobbio, Norberto. 1996. *Left and right: the significance of a political distinction*. University of Chicago Press.
- Bond, Robert and Solomon Messing. 2015. 'Quantifying social media's political space: estimating ideology from publicly revealed preferences on Facebook.' *American Political Science Review* 109(1):62–78.
- Bonica, Adam. 2013. 'Ideology and interests in the political marketplace.' *American Journal of Political Science* 57(2):294–311.
- Bonica, Adam. 2014. 'Mapping the ideological marketplace.' *American Journal of Political Science* 58(2):367–386.
- Crosson, Jesse, Alexander Furnas and Geoffrey Lorenz. 2018. 'Estimating Interest Group Ideal Points with Public Position-Taking on Bills in Congress.'
- De Vries, Erik, Martijn Schoonvelde and Gijs Schumacher. 2018. 'No longer lost in translation: evidence that Google Translate works for comparative bag-of-words text applications.' *Political Analysis* 26(4):417–430.
- Denny, Matthew and Arthur Spirling. 2018. 'Text preprocessing for unsupervised learning: why it matters, when it misleads, and what to do about it.' *Political Analysis* 26(2):168–189.
- Downs, Anthony. 1957. 'An economic theory of political action in a democracy.' *Journal of Political Economy* 65(2):135–150.

- Dür, Andreas. 2008. 'Measuring interest group influence in the EU: a note on methodology.' *European Union Politics* 9(4):559–576.
- Egerod, Benjamin C.K. 2016. Poisson scaling of interest group positions from text in EU consultations. Presented at the *Amsterdam Text Analysis Conference*.
- Grimmer, Justin and Brandon Stewart. 2013. 'Text as data: the promise and pitfalls of automatic content analysis methods for political texts.' *Political Analysis* 21(3):267–297.
- Hotelling, Harold. 1990. Stability in competition. In: *The collected economics articles of Harold Hotelling*. Springer pp. 50–63.
- Jackman, Simon. 2001. 'Multidimensional analysis of roll call data via Bayesian simulation: identification, estimation, inference, and model checking.' *Political Analysis* 9(3):227–241.
- King, Gary. 1998. *Unifying political methodology: the likelihood theory of statistical inference*. University of Michigan Press.
- Klingemann, Hans-Dieter, Andrea Volkens, Michael D. McDonald, Ian Budge and Judith Bara. 2006. *Mapping policy preferences II: estimates for parties, electors, and governments in Eastern Europe, European Union, and OECD 1990–2003*. Vol. 2 Oxford University Press on Demand.
- Klüver, Heike. 2009. 'Measuring interest group influence using quantitative text analysis.' *European Union Politics* 10(4):535–549.
- Lauderdale, Benjamin and Alexander Herzog. 2016. 'Measuring political positions from legislative speech.' *Political Analysis* 24(3):374–394.
- Laver, Michael, Kenneth Benoit and John Garry. 2003. 'Extracting policy positions from political texts using words as data.' *American Political Science Review* 97(2):311–331.
- Lo, James, Sven-Oliver Proksch and Jonathan Slapin. 2016. 'Ideological clarity in multiparty competition: a new measure and test using election manifestos.' *British Journal of Political Science* 46(3):591–610.
- Lowe, Will. 2008. 'Understanding wordscores.' *Political Analysis* 16(4):356–371.
- Lowe, Will. 2015. 'Austin Vignette.' *Austin: Do things with words*. Available at <https://conjugateprior.github.io/austin/articles/austin.html>
- Lowe, Will and Kenneth Benoit. 2011. Estimating uncertainty in quantitative text analysis. Presented at the *Annual Conference of the Midwest Political Science Association*.
- Lowe, Will and Kenneth Benoit. 2013. 'Validating estimates of latent traits from textual data using human judgment as a benchmark.' *Political Analysis* 21(3):298–313.
- Martin, Andrew D. and Kevin Quinn. 2002. 'Dynamic ideal point estimation via Markov chain Monte Carlo for the US Supreme Court, 1953–1999.' *Political Analysis* 10(2):134–153.
- Martin, Gregory and Ali Yurukoglu. 2017. 'Bias in cable news: persuasion and polarization.' *American Economic Review* 107(9):2565–2599.
- Martin, Lanny and Georg Vanberg. 2008. 'A robust transformation procedure for interpreting political text.' *Political Analysis* 16(1):93–100.
- Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of Advances in neural information processing systems 2013*. pp. 3111–3119.
- Monroe, Burt and Ko Maeda. 2004. Talk's cheap: text-based estimation of rhetorical ideal-points. Presented at the *Annual Meeting of the Society for Political Methodology*. pp. 29–31.
- Ng, Patrick. 2017. 'dna2vec: Consistent vector representations of variable-length k-mers.' *arXiv preprint arXiv:1701.06279*.
- Perry, Patrick and Kenneth Benoit. 2017. 'Scaling text with the Class Affinity Model.' *arXiv preprint arXiv:1710.08963*.
- Poole, Keith and Howard Rosenthal. 1985. 'A spatial model for legislative roll call analysis.' *American Journal of Political Science*, 29(2): 357–384.
- Poole, Keith and Howard Rosenthal. 2000. *Congress: a political-economic history of roll call voting*. Oxford University Press on Demand.
- Slapin, Jonathan B. and Sven-Oliver Proksch. 2008. 'A scaling model for estimating time-series party positions from texts.' *American Journal of Political Science* 52(3):705–722.
- Slapin, Jonathan B. and Sven-Oliver Proksch. 2014. 'Words as data: content analysis in legislative studies. In: Shane Martin, Thomas Saalfeld and Kaare W. Strøm (eds.): *The Oxford handbook of legislative studies* pp. 126–144. Oxford University Press.
- Smithies, Arthur. 1941. 'Optimum location in spatial competition.' *Journal of Political Economy* 49(3):423–439.



# Classification and Clustering

Sarah B. Bouchat

Text serves as a critical source of political data – comprising everything from official statements to news coverage, social media responses and legal documents – that is also increasingly accessible via large-scale digitization and open-source analytical tools. Perhaps even more critically for political science, classifying and characterizing these complex data facilitates inferences about policies, norms, strategic communication and the ideological landscape. Much like typological research that has a long lineage in the social sciences, classification using text data allows us to make fundamental claims about the nature of discourse and communication in politics that generalize across space and over time.

This chapter examines both supervised and unsupervised learning techniques for classification and clustering as they apply to text data, detailing both the history and available methodologies developed in information retrieval, natural language processing and computational linguistics, as well as the

extant applications of these approaches to questions in political science.

The remainder of the chapter proceeds as follows. First, it describes the goals of text classification and its theoretical underpinnings, including what distinguishes supervised and unsupervised alternatives. The second section offers a brief history of the development of these methods, including their origins and uses in cognate disciplines. The following sections explain how to implement text classification using both supervised and unsupervised approaches, including Naïve Bayes and Latent Dirichlet Allocation.

These sections also provide examples of work within political science and international relations that applies these methodologies, using these works to highlight the significant opportunities that exist for text-as-data within political science, but also to examine the drawbacks and limitations of classification approaches. The final section expands upon these insights to describe the practical issues that researchers might

confront in using text classification methodologies, including what problems are most and least suitable for these types of methods and the research design considerations they require. The chapter concludes with suggestions for further development in classification methodologies that would benefit political scientists.

## WHAT IS TEXT CLASSIFICATION?

A variety of classification techniques exist within the machine learning literature that apply to a broad set of data types, where boundaries that discriminate between distinct classes within the data may be linear or non-linear. A regression-based understanding of this type of classification specifies a separate *discriminant function* for each class, into which data are then sorted according to which class maximizes this function. While students of statistical and machine learning approaches to data classification more generally will be familiar with logistic regression or linear discriminant analysis approaches, this chapter highlights methods that are particularly common for analyzing text data.

Broadly speaking, and as is true for other types of data, text classification aims to map a set of inputs (e.g., documents) to a predicted class as the output. The utility of these approaches for textual data is directly related to the high dimensionality of text: automated approaches allow for dimensionality reduction in ways that manual coding and analysis do not. As with other types of data, text classification using supervised learning requires a defined set of classes  $\mathbf{C}$  into which text can be classified:  $\mathbf{C} = c_1, c_2, c_3, \dots, c_j$ . Suppose that text is organized coherently into documents  $d_i \in \mathbf{D}$ . The classification learning procedure then wants to determine  $f(d_i) \in \mathbf{C}$ ; that is, it learns  $f$ , the *classification function*. In order to do this, a researcher requires a training set of documents selected from

the corpus  $\mathbf{D}$  that already have labels in  $\mathbf{C}$ . The remainder of this portion of the chapter explains how each of the previously mentioned classification approaches applies specifically to text data.

For research tasks in which the underlying set of categories is not known or assumed, unsupervised approaches using clustering are appropriate. These approaches seek to ‘classify’ documents or components of text into classes that share some characteristics, where ‘discovering’ these classes and the features on which they rely is part of the research process. Grimmer and Stewart (2013) provide an extensive overview of the varying types of classification approaches for unknown categories, including both fully automated and computer-assisted clustering techniques.

## THE EVOLUTION OF TEXT CLASSIFICATION METHODS

Prior to the advent – and increasing popularity – of automated text classification, manual approaches dominated, particularly for social and political science projects. Many research projects utilizing qualitative text analyses used the presence or absence of keywords to determine whether a given text matched a category of interest. This manual undertaking can now be automated via supervised machine learning, where dictionary methods measure relative frequencies of a set of words to classify documents into categories. Automation further facilitates even more complex analyses that take into consideration more features of text.

Supervised classification approaches in particular can be thought of as strategies to improve the efficiency and accuracy of tasks that have been undertaken in content analysis of political texts for decades. While automation indeed promises to diminish the significant time and human resource expenditure necessary to perform some of these analyses, classification approaches are not without

their own set of assumptions, challenges and modeling constraints.

One popular example of text classification is spam filtering in email. Spam filters learn, via initialized settings as well as user-tagged examples, what types of senders, words, links and other characteristics are most indicative that an email is spam. While generating an initial set of indicators for spam emails may seem simple and obvious (e.g., misspellings, mysterious senders, Nigerian princes), generating a list of characteristics that is reliable or exhaustive, and which is highly accurate across accounts, languages or time, suggests just one of the main challenges of supervised learning tasks: specifying categories and characteristics. Search engine indices, another application of text classification, also confront this challenge of applying a constrained set of categories to an increasingly large number of highly complex pages that change over time. As daunting as these tasks may appear in applications that many people utilize on a daily basis, the underlying challenge applies to classification tasks in social and political science as well.

Mosteller and Wallace (1963) demonstrate the applicability of this type of text classification to questions in political science. Mosteller and Wallace undertake an authorship attribution task to attempt to identify the authors of the anonymously written *Federalist Papers*, using a Bayesian approach (a predecessor to Naïve Bayes) to distinguish papers most likely written by Hamilton from those most likely authored by Madison. This approach also generalizes to other authorship attribution problems focused on differences in writing style, including identifying the author's gender or demographic characteristics. Sentiment analysis, discussed in depth elsewhere in this *Handbook*, is another common example of classification: identifying words that appear in text and counting and categorizing them according to their tone, mood or emotional content.

A key component of the successful implementation of classification in a case like

Mosteller and Wallace's, however, is the presence of text with known category labels. For Mosteller and Wallace, these were texts known to be authored by Hamilton and Madison, to which unattributed *Federalist Papers* could be compared. While automation may speed the sorting of documents into classes, generating labels for texts and applying these labels to an initial 'training' set of documents presents both theoretical concerns and practical limitations when creating hand-coded labeled corpora for analysis (Manning et al., 2009). In addition to resource-intensiveness, issues with inter-coder reliability in applying known class labels can threaten the consistency and quality of results in text classification projects.

The development of crowdsourcing platforms to distribute the task of labeling training data, or algorithms such as *ReadMe* (Hopkins and King, 2010) that can efficiently predict classes with fewer training observations, have facilitated more text classification projects in political science than resources would previously allow. Yet these gains in efficiency do not resolve the primarily theoretical problem of defining and delineating categories to structure supervised learning.

Unsupervised approaches, in turn, face different assumptions and raise additional challenges despite the appearance of being less model-reliant. For topic modeling projects, for example, while an unsupervised approach allows for the discovery of new classes and latent organization within and among texts, it remains the job of the researcher to interpret the topics identified in the modeling procedure and the words that characterize them. Making these automatically generated topics legible and ensuring their relationship to the initial research question is related to parameter choices made at the outset of the modeling process, but is not obviated by them. The next section discusses these design considerations and challenges for both supervised and unsupervised approaches in greater depth.

## DESIGN AND PROCEDURE

This section describes the design and implementation of text classification projects from the initial step of identifying documents and labeling training data through to the analysis of the test set and validation of the results. While the process used to collect and extract the text data will influence the ultimate quality of the analysis, the steps for collecting and processing text data are covered extensively elsewhere and therefore omitted from this discussion. Rather, this section begins with a discussion of how to define and design a text analysis project with classification in mind.

A key, often undiscussed, assumption that inheres to text classification projects is the coherence of the corpus under examination. Supervised approaches may not adequately distinguish among texts that fit the underlying assumptions of the categorization and those that do not, while unsupervised approaches will merely ‘discover’ structural qualities of the text (e.g., shifts across time or subject matter) that might have been apparent prior to the analysis or that may be of less interest. That is, using automated text analysis procedures does not absolve the researcher of responsibility for ensuring that the texts under examination are related to the research question and have theoretical consistency. For example, texts may be selected for inclusion in a corpus on the basis of keyword searches, themselves a form of classification that can produce errors, omit relevant documents and include irrelevant ones (Hillard et al., 2008). Even having identified and justified the integrity of a particular corpus, the unit of analysis will also influence the nature and quality of classification results. Throughout this chapter, the unit of analysis for classification is referred to as the ‘document’, but classification can be conducted at lower units of analysis as well (e.g., sections, paragraphs, sentences, words).

While not discussed in depth here, a prerequisite for classifying text data is formatting the data to be machine readable and structured according to the features most

appropriate for analysis, where terms (words) are the most commonly selected feature of interest. Most applications of text classification use a *bag-of-words* approach, meaning that words are treated as units without respect to their order. Each document, then, appears in the dataset as a vector of counts for each word present, eventually comprising a *document–term matrix*. Words may also be weighted according to how rare or frequent they are in the corpus via a *tf–idf* (term frequency–inverse document frequency) matrix (Manning et al., 2008).

Once the researcher has constructed their corpus and divided it into appropriate units of analysis, procedures diverge depending on whether the ultimate objective is a supervised or unsupervised task. For a supervised task, data should be divided into training and test sets. The classification algorithm will eventually ‘learn’ about appropriate categories and classes via the training data, and then apply these insights to classify the text data. Practically speaking, separating data into training and test sets is typically achieved via random partitions. At one extreme, where no training data are available, or the dataset does not allow for separation into training and test sets, relying on a priori manual coding rules is necessary; while the dictionary-based classification can be automated, the project is greatly simplified (e.g., only documents containing the word ‘Republican’ will be classed as ‘conservative’).

How much data are allocated to each set will depend in part on the project objective, the size and diversity of the corpus and the complexity of the problem. If a category does not occur, or occurs extremely rarely, in the training set, there is insufficient opportunity to ‘learn’ about this category and its properties, which will in turn interfere with the process of classifying test-set documents into this category correctly. When attempting to detect small changes or rare categories, therefore, increasing the probability that they are observed in the training set often means increasing the size of the training set relative

to the test set. The feasibility of achieving an optimal training/test split for a given research question will in part depend on the size of the corpus; that is, some research questions emphasizing particularly complex or difficult-to-detect classes, which in turn require larger training sets, may be best served by selecting a large corpus at the outset and allowing flexibility in achieving the appropriate training/test split. Hopkins and King (2010) provide additional guidance on selecting the size of training and test sets.

## SUPERVISED METHODS

For supervised problems, the researcher is aiming to classify documents into a set of known or assumed categories based upon rules or information that can be learned from the training set. This requires *labels* in the training set from which to infer categories in the test set. Once labels are applied to the training set and a supervised learning method generalizes from this training set to the test set, only validation of the model remains. While data in some sorts of tasks, such as author-attribution problems, are conveniently pre-labeled, most often this step of the research requires human effort, whether via experts, trained research assistants or crowdsourcing on platforms like Amazon's Mechanical Turk.

Manually generating this initial set of labels can prove arduous and time-consuming, but is also fraught with concerns about consistency and accuracy. King et al. (2017) highlight this problem with respect to human keyword selection: when attempting to identify and recall keywords for search and collection tasks, humans are extremely unreliable and inconsistent. That is, while labeling training data often requires the use of human coders to sort texts into desired categories, human coding lacks consistency and reliability both within and across individuals, above and beyond the time and expense required to complete the task.

Conditional on having identified a set of classes into which documents will fit, and defined labels for a training set of those documents, several possible supervised learning approaches exist to conduct text classification, including Naïve Bayes, Support Vector Machines (SVMs), k-Nearest Neighbors and logistic regression approaches.

Once texts are correctly processed and formatted for analysis, several classification techniques are available. Naïve Bayes is an appropriate tool for classification of high-dimensional feature spaces, of which text is a classic example (Hastie et al., 2009: 210). Naïve Bayes applies Bayes' Rule to a given document  $d_i$  for a given class  $c$ :  $P(c|d_i) = \frac{P(c) \prod_j P(d_{ij}|c)}{\sum_{c'} P(c') \prod_j P(d_{ij}|c')}$ . That is, Naïve Bayes determines the class into which a given document most likely fits given the features of the document. Notably, Naïve Bayes treats these features as though order does not matter (a typical bag-of-words condition) and as though they are independent given the class (conditional independence). The latter assumption about feature independence conditioned on class is what makes Naïve Bayes 'naïve': this is a *strong* assumption that is necessarily inaccurate for text data, where the existence of some features will highly correlate with the existence of others. That is, using words as features, documents are predicted to belong to a class according to whether some words are present while other words are not, but words often co-occur with a regular frequency, which can complicate the classification process.

To provide an intuition for Naïve Bayes, suppose you want to classify newspaper articles according to whether they cover a 'political' subject or not. You encounter the following example phrases in your training set:

<i>Phrase</i>	<i>Class</i>
'The end of election season'	Political
'A new storefront'	Not political
'Closing the polls'	Political
'Parades drew crowds'	Not political
'A close contest'	?

Using Naïve Bayes, you want to estimate  $\Pr(\text{Political} \mid \text{'A close contest'})$ , using features such as relative word frequency (i.e., how often

does that phrase appear in articles known to be political versus not?). Constructing the desired probability via Bayes' Rule, you would have:

$$\Pr(\text{Political} \mid \text{'A Close Contest'}) = \frac{\Pr(\text{A close contest} \mid \text{Political}) \times \Pr(\text{Political})}{\Pr(\text{A close contest})}$$

Note that the divisor serves as a scaling constant across both classes of interest. This exact phrase may not appear in the training set, however. Phrases are larger components of text than individual words, and are therefore less probable than the words that

comprise them. To overcome this challenge, the bag-of-words assumption then is useful – rather than calculating the probability of the phrase, we can calculate the joint probability of these component words due to their assumed independence:

$$\Pr(\text{'A close contest'}) = \Pr(a) \times \Pr(\text{close}) \times \Pr(\text{contest})$$

Useful though it may be, this assumption is necessarily strong: words tend to co-occur regularly, rather than appearing with independent frequency. Furthermore, the model incorporates Laplace smoothing of final estimates to compensate for the non-appearance of certain terms in the training data.

Because it primarily emphasizes counting features, Naïve Bayes has persisted as a fast, reliable and less memory-intensive approach that works well even with relatively little training data (although it suffers from bias, as per Ng and Jordan, 2002), and significantly better with a very large amount of data. Likewise, it works well with many, equally weighted features, while irrelevant features will cancel out, and does not tend toward overfitting on smaller datasets as other approaches do.

For medium-sized datasets, where Naïve Bayes suffers most from high bias, Support Vector Machines (SVMs) provide a viable alternative. SVMs find the optimal separating hyperplane between data points by maximizing distance (the 'margin') between a given hyperplane and the points nearest to the decision boundary that would be most challenging to classify; that is, it is one example of a *maximal margin classifier*. Because full separability among latent classes may not exist, this margin is 'soft', which gives rise to the support vector. SVMs

generalize to sets of classes greater than two as well, either via using a multinomial loss function or by reducing the problem to a two-class solution (comparing one-to-one or one-to-all).

k-Nearest Neighbors (KNN) approaches similarly use observations within the training set and locate other observations in their 'neighborhood' to predict outcome values in a regression framework. That is, the regression identifies the  $k$  training observations that exist closest (in Euclidean distance) to a given point of interest  $x_0$  (a 'neighborhood'  $N_0$ ), and then estimates  $f(x_0)$  using the average of those training responses:

$$f(x_0) = \frac{1}{|N_0|} \sum_{x_i \in N_0} y_i$$

Implementing kNN requires selecting an optimal value of  $k$ , where that optimal value depends on a researcher's assessment of the bias–variance tradeoff. Small values of  $k$  will generate low bias, high variance responses; larger values generate 'smoother' outcomes (high bias, low variance).

## UNSUPERVISED METHODS

Supervised classification methods assume defined types or categories into which



documents fit. Within the realm of computer science or information retrieval, supervised approaches are therefore definitionally ‘classification’, since the typology into which documents are sorted already exists, and the procedure for correctly identifying documents that belong to these prescribed categories is the central focus. In political and social science applications, however, ‘classification’ is a broader theoretical project, part of the overarching objective to generalize and infer information from the data at hand. To that end, even projects where categories and major characteristics are unknown can be thought of as ‘classification’ projects.

For this type of goal, when the research design seeks to discover the set of categories among the texts in a corpus, unsupervised methods such as clustering and topic modeling are appropriate. Unsupervised approaches can both generate categories from the features of text and assign (classify) texts according to those categories. While a variety of techniques and variations or refinements of general algorithms exist to conduct unsupervised learning about text categories, this section primarily focuses on two commonly used options: k-Means and Latent Dirichlet Allocation (LDA).

For unsupervised problems, no pre-set categories exist into which documents must be classified. Rather, ‘classification’ occurs via the discovery of underlying characteristics and categories within text. With high-dimensional and complex data like text, identifying all possible partitions of the text is neither feasible nor useful for answering social science questions. Rather, unsupervised approaches ideally allow researchers to uncover characteristics and categories that have theoretical significance and perhaps substantive importance.

Directly generalizing from the supervised case in which a researcher seeks to classify documents according to their known set of categories (e.g., according to partisan identification or authorship), an analogous

unsupervised task supposes that documents or sets of text belong to a single category or class, but does not impose a label on that class a priori. For the k-Means algorithm, this assumption means that the model seeks to minimize the difference between documents that define a cluster, where that difference is measured as their squared Euclidean distance in vector space. Once the researcher selects a number of clusters  $k$ , the k-Means algorithm produces a set of partitions for the documents in the corpus that minimize the distance between documents within each cluster from the cluster centroid. That is, it seeks to minimize within-cluster variation across all possible clusters.

Yet because k-Means relies on an approximation for optimization, it is subject to convergence problems and dependence on initialization values. For social science questions, furthermore, k-Means has additional pitfalls in its ability to adapt to the inclusion of new data or to appropriately evaluate true ‘outlier’ observations: new data may disrupt the pattern giving rise to the original partition, and determining which classification to ‘prefer’ is a matter of theory rather than one of model output, while ‘outlier’ observations are likewise coerced into inclusion in clusters as part of the distance minimization procedure. In addition, the selection of  $k$  is based on a metric of diminishing marginal returns for additional clustering – once the total within-cluster sum of squared distance decreases past a certain threshold, a larger  $k$  no longer decreases within-cluster variation. While practical, this parameterization may seem deeply atheoretical for certain social science projects.

Likewise, k-Means operates from an assumption that each document must belong to a category and that categories do not overlap. Mixed membership models such as *topic models* instead emphasize categories at the level of the word that can then generalize to the level of the document. The most common type of topic modeling, Latent Dirichlet Allocation, begins from the assumption that

each document is a mixture of topics; that is, topics are multinomial random variables defining a mixture model from which words are sampled (Blei et al., 2003). Documents, then, are a probability distribution over topics (Blei, 2012). In this sense, a whole document may be ‘classified’ into a given topic, but more accurately portions of documents are classified into topics across the entire corpus.

Colloquially, the notion that unsupervised methods are superior, in that they are not ‘model dependent’ like supervised approaches, is prevalent. Yet, far from achieving superiority or more ‘objectivity’, unsupervised approaches present their own set of challenges, namely in validation and researcher discretion. In addition to topic models producing unstable results across iterations (Wilkerson and Casas, 2017), validation of unsupervised models can prove difficult since labels and categories were not chosen prior to analysis. Several possible methods of validation are possible. Human coders can identify words that distinguish topics and conduct manual labeling to check the output of the automated model, but this is often intractable or counter to the purpose of an unsupervised method. Experimental validation can vary word or topic inclusion in the model to assess variation in results, or metrics such as exclusivity (words having a high probability in one topic have low probabilities in others), entropy (difference between topics) and cohesion (the affinity of particular words with a given topic versus others, such as with pointwise mutual information) may be used. Grimmer and Stewart (2013: 287) distinguish between semantic validity (coherent representation of concepts within topics) and predictive validity (validity in comparison to external circumstances or shocks). Which approach is most suitable for validation is not clear *ex ante*, and current validation procedures are often designed as bespoke solutions for a particular research problem rather than as tried and tested methods that generalize for unsupervised approaches.

## APPLICATIONS OF SUPERVISED METHODS

Supervised methods have been used for a range of political science questions and applications, for projects ranging from political economy to international relations. Within these varying applications, furthermore, authors select the type of supervised approach most appropriate to the topic. O’Halloran et al. (2016), for example, utilize Naïve Bayes to evaluate financial regulatory laws from the United States, assessing the extent to which legislation reflects ‘agency discretion’ since 1950. For their research, Naïve Bayes performs better than focusing only on manually coded text features or a set of computer-generated features when attempting to classify documents according to a ‘score’ for the amount of delegation they prescribe (O’Halloran et al., 2016: 106). This research on financial regulation is just one instance in which automated classification demonstrates improvements over a research design relying solely on human analysis.

D’Orazio et al. further note that while researchers may want to gather information on a specific topic, *how* exactly to identify what information is relevant and to organize these data for analysis is not clear (D’Orazio et al., 2014). While the authors address document classification in political science broadly as a technique with a variety of applications, they particularly emphasize the value of SVMs as a linear classification approach that does not require as much labeled training data but nevertheless handles sparse text data well. The paper introduces a two-step SVM approach that facilitates classifying 1.74 million news documents from the Correlates of War Militarized Interstate Dispute (MID) data. While D’Orazio et al. highlight the challenges that other types of classification (e.g., Naïve Bayes or k-Nearest Neighbors) would pose for their data, other applications to questions of political interest seek to evaluate appropriateness or combine approaches for greater accuracy and

flexibility. Mickevicius et al. (2015), for example, test both SVM and kNN in classifying the topics of Lithuanian parliamentary votes to evaluate which approach is most efficient, concluding that SVM is most accurate in their desired application. Similarly, Collingwood and Wilkerson (2011) assess how much labeled training data is necessary for a variety of approaches, including SVM and Naïve Bayes, noting that particularly for SVM, machine classification is on par with trained human classification even with only 15% of data labeled.

Another classic example of classification in political science pertains to party or ideological classification, often from Congressional speech or legislation (e.g., Yu et al., 2008). Diermeier et al. (2011) use SVMs to identify words most indicative of conservative versus liberal policy positions using legislative speeches from the US Senate. That more ‘cultural’ terms, rather than economic ones, provide the most leverage in distinguishing between political ideologies runs counter to expectations, and demonstrates the value of directly engaging with and classifying textual references in an automated fashion. Human coders with prior training in political science theory might have been primed to think economic language was most distinct or important and classified accordingly, while even an automated approach predicated on knowing the ideological position of some senators in the sample was able to uncover ‘culture’ words as a useful measure of distinction. This type of task, predicting party affiliation or ideological leaning based on speech, is particularly prevalent among political science applications. Evans et al. (2007) likewise utilize several modeling techniques, including Naïve Bayes, to classify amicus curiae briefs from the affirmative action cases in *Bakke* (1978) and *Bollinger* (2003). Classification research seeking to identify ideological dimensions now also leverages new data sources, such as Facebook pages and posts (see, e.g., Chiu and Hsu, 2018), and can extend beyond these typical dimensions

like partisanship to apply to other areas, like re-constructing China’s censorship architecture via observed banned terms and websites (King et al., 2013).

As these examples illustrate, the potential applications for supervised text classification in political science are many and varied, but authors have the responsibility to adjudicate between existing classification algorithms and approaches to determine which is most suitable given the size and nature of their text data. In addition, while manual label generation for supervised methods remains costly, recent efforts to automate procedures for providing labels for text documents could significantly improve the uptake and efficiency of supervised approaches. Hopkins et al. (2010) introduced the *ReadMe* software, which takes text documents and a classification scheme selected by the user as inputs to produce a proportion of documents aligning with those categories. Likewise, *RTextTools* combines all preprocessing and classification steps, as well as assessments of algorithm accuracy and ensemble results, into a single package (Jurka et al., 2013). Miller et al. (n.p., 2018), for example, draw on a variety of active learning approaches to overcome the issue of sampling unlabeled data for the purposes of text classification. Creative combinations of approaches and applications to new data – see, e.g., Ku and Leroy (2014) on classifying crime reports using a combined ‘decision support system’ with aspects of Naïve Bayes and logistic regression, or Sapiro-Gheiler (n.p., 2018) comparing results across Naïve Bayes, SVM, lasso and decision trees when evaluating Congressional speech data – suggest that supervised text classification has many promising future directions.

## APPLICATIONS OF UNSUPERVISED METHODS

The increasing popularity of text-as-data methods in political science has led to an

explosion of research, particularly in topic modeling. Wilkerson and Casas (2017), for example, evaluate the topics that arose in 10,000 short floor speeches within the US Congress in 2013–14. LDA enables the authors to evaluate which topics emerge most frequently, and whether partisan affiliation impacts which topics are spoken about. Mueller and Rauh (2018) also use LDA to assess topics within newspaper text, which then serve as a basis for predicting conflict in a panel regression setting. The authors leverage evolution and change in topics within text, as well as their relative stability over time given the nature of newspaper coverage, to tackle the notoriously difficult task of predicting variation in conflict onset.

In a classic example of topic modeling for political texts, Quinn et al. (2010) emphasize temporal variation in topic analysis for political texts, specifically political speeches that have a single (but unknown) topic. Again, the distinctions which the authors highlight in their data (e.g., that speeches focus on a particular topic but that distinguishing and identifying topics presents a challenge) relative to computer science applications of dynamic topics indicates that political science has much to contribute in the evolution of text classification methods for questions and subjects that might otherwise be overlooked.

## COMPLICATIONS AND EXTENSIONS

The starkest limitation of supervised classification methods is the requirement that categories be pre-defined. This theoretical limitation then directly relates to the practical challenge of acquiring labeled data on which to train a classification algorithm. The arduousness of labeling training data for supervised approaches should not, however, be interpreted to suggest that unsupervised methods have greater utility or more efficiency or are more ‘objective’ by dint of omitting larger scale manual coding. Rather,

supervised and unsupervised approaches may be used in tandem for similar projects and research questions, but should also each be evaluated on their merits for a particular application according to their strengths and pitfalls. The advent of automated methods for text classification opens up a vast array of possibilities for research questions in every subfield of political science that seek to utilize increasingly large and complex text datasets.

As Hopkins and King (2010) note, however, methods adapted from computer science do not necessarily satisfy the research aims of social science, and should be adopted and adapted with care. Rather than an emphasis on classifying particular documents, they argue, social scientists aim to generalize: a task to which supervised classification methods are not particularly well suited. The development of text classification, categorization and clustering techniques by social scientists is therefore critical to the existence of models and approaches that better account for the objective of inference, rather than prediction.

For many social science questions, for example, metadata about the source, speaker, date or other characteristics of the documents in a corpus can provide useful insights and condition the presence and prevalence of topics or categories in text. This observation prompted the introduction of structural topic models (STM) by Roberts et al. (2016), where topics are conditioned on document-level covariates. Roberts et al. (2014) apply this method to open-ended survey responses, bridging a gap in literatures between machine learning methods for text analysis and causal identification – a goal of particular interest to social and political scientists. As with topic modeling using LDA, STM requires the researcher to select an appropriate number of topics for their corpus, and while the output has greater structure than with LDA, interpretation of topic prevalence and content is still at the researcher’s discretion. While this is only one example, STM represents the

promise and possibility of not only applying text classification techniques from other fields to the questions and problems that are most pressing in political science, but also innovating on these tools to better serve social science research aims.

## REFERENCES

- Blei, David M., Andrew Y. Ng and Michael I. Jordan. 2003. 'Latent Dirichlet Allocation.' *Journal of Machine Learning Research* 3(4–5): 993–1022.
- Blei, David M. 2012. 'Probabilistic Topic Models.' *Communications of the ACM* 55(4): 77–84.
- Chiu, Shu-I and Kuo-Wei Hsu. 2018. 'Predicting Political Tendency of Posts on Facebook.' *Proceedings of the 2018 7th International Conference on Software and Computer Applications* 110–114.
- Collingwood, Loren and John Wilkerson. 2011. 'Tradeoffs in Accuracy and Efficiency in Supervised Learning Methods.' *Journal of Information Technology and Politics* 4: 1–28.
- D'Orazio, Vito, Steven T. Landis, Glenn Palmer and Philip Schrod. 2014. 'Separating the Wheat from the Chaff: Applications of Automated Document Classification Using Support Vector Machines.' *Political Analysis* 22(2): 224–242.
- Diermeier, Daniel, Jean-François Godbout, Bei Yu and Stefan Kaufmann. 2011. 'Language and Ideology in Congress.' *British Journal of Political Science* 42(1): 31–55.
- Evans, Michael, Wayne McIntosh, Jimmy Lin and Cynthia Cates. 2007. 'Recounting the Courts? Applying Automated Content Analysis to Enhance Empirical Legal Research.' *Journal of Empirical Legal Studies* 4(4): 1007–1039.
- Grimmer, Justin and Brandon Stewart. 2013. 'Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts.' *Political Analysis* 21(3): 267–297.
- Hastie, Trevor, Robert Tibshirani and J. H. Friedman. 2009. *The Elements of Statistical Learning* (2nd ed.). New York: Springer.
- Hillard, Dustin, Stephen Purpura and John Wilkerson. 2008. 'Computer-Assisted Topic Classification for Mixed-Methods Social Science Research.' *Journal of Information Technology and Politics* 4(4): 31–46.
- Hopkins, Daniel and Gary King. 2010. 'Extracting Systematic Social Science Meaning from Text.' *American Journal of Political Science* 54(1): 229–247.
- Hopkins, Daniel, Gary King, Matthew Knowles and Steven Melendez. 2010. *ReadMe: Software for Automated Content Analysis*. <https://gking.harvard.edu/readme>.
- Jurka, Timothy, Loren Collingwood, Amber Boydston, Emiliano Grossman and Wouter van Atteveldt. 2013. 'RTextTools: A Supervised Learning Package for Text Classification.' *The R Journal* 5(1): 6–12.
- King, Gary, Patrick Lam and Margaret E. Roberts. 2017. 'Computer-Assisted Keyword and Document Set Discovery from Unstructured Text.' *American Journal of Political Science* 61(4): 971–988.
- King, Gary, Jennifer Pan and Margaret E. Roberts. 2013. 'How Censorship in China Allows Government Criticism but Silences Collective Expression.' *American Political Science Review* 107(2): 1–18.
- Ku, C. H. and G. Leroy. 2014. 'A Decision Support System: Automated Crime Report Analysis and Classification for E-government.' *Government Information Quarterly* 31(4): 534–544.
- Manning, Christopher D., Prabhakar Raghavan and Hinrich Schütze. 2009. *An Introduction to Information Retrieval*. Cambridge: Cambridge University Press.
- Mickevicius, Vytautas, Tomas Krilavicius, Vaidas Morkevicius and Austra Mackute-Varoneckiene. 2015. 'Automatic Thematic Classification of the Titles of the Seimas Votes.' *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)* 225–231.
- Miller, Blake, Fridolin Linder and Walter R. Mebane, Jr. 2018. 'Active Learning Approaches for Labeling Text: Review and Assessment of the Performance of Active Learning Approaches.' Working paper, available at <http://www-personal.umich.edu/~wmebane/active-learning-approaches-4-18-2018.pdf>
- Mosteller, Frederick, and David L. Wallace. 1963. 'Inference in an Authorship Problem.'

- Journal of the American Statistical Association* 58(302): 275–309.
- Mueller, Hannes, and Christopher Rauh. 2018. 'Reading Between the Lines: Prediction of Political Violence Using Newspaper Text.' *American Political Science Review* 112 (2): 358–375.
- Ng, Andrew Y., and Michael I. Jordan. 2002. 'On Discriminative vs. Generative Classifiers: A Comparison of logistic regression and naïve Bayes.' *Advances in Neural Information Processing Systems* 841–848.
- O'Halloran, Sharyn, Sameer Maskey, Geraldine McAllister, David K. Park and Kaiping Chen. 2016. 'Data Science and Political Economy: Application to Financial Regulatory Structure.' *The Russell Sage Foundation Journal of the Social Sciences* 2(7): 87–109.
- Quinn, Kevin M., Burt L. Monroe, Michael Colaresi, Michael H. Crespin, and Dragomir R. Radev. 2010. 'How to Analyze Political Attention with Minimal Assumptions and Costs.' *American Journal of Political Science* 54(1): 209–228.
- Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson and David G. Rand. 2014. 'Structural Topic Models for Open-Ended Survey Responses.' *American Journal of Political Science* 58(4): 1064–1082.
- Roberts, Margaret E., Brandon M. Stewart and Edoardo M. Airoldi. 2016. 'A Model of Text for Experimentation in the Social Sciences.' *Journal of the American Statistical Association* 111(515): 988–1003.
- Sapiro-Gheiler, Eitan. 2018. "'Read My Lips": Using Automatic Text Analysis to Classify Politicians by Party and Ideology.' *arXiv preprint arXiv:1809.00741*
- Wilkerson, John and Andreu Casas. 2017. 'Large-scale Computerized Text Analysis in Political Science: Opportunities and Challenges.' *Annual Review of Political Science* 20: 529–544.
- Yu, Bei, Stefan Kaufmann and Daniel Diermeier. 2008. 'Classifying Party Affiliation from Political Speech.' *Journal of Information Technology, and Politics* 5(1): 33–48.



# Sentiment Analysis and Social Media

Luigi Curini and Robert A. Fahey

## **'HOW DOES IT MAKE YOU FEEL?'**

Sentiment analysis, sometimes also called tone analysis (Grimmer and Stewart, 2013) or opinion mining (Dave et al., 2003), is a set of methods designed to answer the same deceptively simple question – how do people feel about a given subject? More precisely, given a specific topic (which might be a product in the case of a marketing analysis, or a politician, party or policy in the case of a political analysis), do the authors of a certain set of texts express positive or negative feelings with regard to that topic?

This is the kind of task which humans generally find intuitive and simple, but which transpires to be exceptionally complex and difficult for computers to perform. A class instructor reading over their end of term feedback forms will quickly form a sense of whether students enjoyed the course or not. A computer performing text analysis on the same data, however, faces a number of daunting problems. A text analysis algorithm will

lack context, so students' references to specific topics or events during the semester will be rendered meaningless. It may find some of the casual or idiomatic language students use difficult to parse or recognise (in this, at least, many instructors may sympathise). Moreover, given the inherently stochastic nature of text and working from such a limited set of data – just a few dozen pieces of text – it may be unable to establish any consistent framework for what a 'positive' or 'negative' piece of feedback actually looks like.

In spite of this difficulty, sentiment analysis has become an essential tool in many fields across the social sciences. Like many other text analysis techniques, its great strength lies in its ability to handle large volumes of data. While a human being may make a more reliable and nuanced judgement of sentiment on a single piece of text, or a small number of texts, it is impossible for a human to read and analyse the volume of texts generated by, for example, product reviews on a major website, or the customer feedback form for

a popular product. The problem of volume is compounded further when social media is the focus of study; a major event can generate millions of posts per minute on a site like Twitter and even a relatively obscure topic may leave researchers with hundreds of thousands, if not millions, of posts to analyse. Effective application of sentiment analysis to social media has created remarkable new possibilities for political and social researchers – for example, real-time sentiment analysis of Twitter data has allowed researchers to see how audiences are reacting moment-to-moment to election broadcasts and candidate debates (Wang et al., 2012; Smailović et al., 2015).

This chapter will introduce a range of different approaches which are used for sentiment analysis, broadly categorising them according to the form of methodology employed and discussing the benefits and limitations of each approach. There is no single ‘best in breed’ approach to sentiment analysis and a researcher’s choice of technique will depend heavily on the kind of texts they wish to analyse and the resources – in terms of both time and technical skill – at their disposal.

## APPROACHES TO SENTIMENT ANALYSIS

Sentiment analysis approaches can broadly be divided into two major categories – dictionary techniques and supervised learning techniques. Dictionary techniques, as the name suggests, involve constructing a dictionary of sentiment-scored terms and applying it to a text using an algorithm which uses the dictionary scores to calculate the sentiment of the overall text. These approaches range from relatively naïve sentiment scoring of individual words through to more complex techniques, for example taking into account the surrounding context of a word in determining its scoring and weight. Supervised

learning techniques, on the other hand, rely on the creation of a labelled set of data – usually by getting human coders to classify the sentiments of a sub-sample of the data to be analysed. This human-coded data is then used to train an algorithm which will classify the remainder of the (unlabelled) data. While many supervised learning approaches to sentiment analysis, described below as ‘classification methods’, are effectively special cases of the broader text classification methodologies outlined in Chapter 28 of this *Handbook*, another group of approaches – which we refer to herein as ‘aggregate methods’ – do not try to assign a specific classification to every piece of text in a corpus, instead aiming to probabilistically estimate the distribution of sentiment across the corpus as a whole. These aggregate methods are especially well suited to corpora with very large numbers of very short texts – for example, Twitter posts – where individual messages might not contain enough information for standard classification approaches to make a reliable determination about sentiment.

### **Dictionary Approaches**

Dictionary, or lexicon-based, approaches to sentiment analysis rely, at the most basic level, on identifying specific words as being positive or negative and calculating the sentiment of a text based on those scores.<sup>1</sup> The most challenging aspect of this approach is compiling the sentiment dictionary itself; many different ways of tackling this problem have been developed, such as crowdsourcing lexicon entries through services such as Mechanical Turk (Haselmayer and Jenny, 2017); using a thesaurus to iteratively ‘snowball’ out from an initial set of human-chosen positive and negative seed words; estimating positive and negative weights for new words based on the frequency of their co-occurrence with known words in a large text corpus; or using pointwise information to calculate the probability of given words appearing in



positively or negatively coded documents. This process is generally complex and resource-intensive, and generally beyond the scope of a political science research project; while there may be circumstances in which a political science researcher decides to compile their own sentiment dictionary (either selecting terms by hand, an example of which can be seen in Rooduijn and Pauwels' (2011) work on identifying populism in text, or by refining existing lists of words to make the sentiment scoring more relevant to a specific field, such as in Loughran and McDonald (2011)), it is much more common to use an existing dictionary that has been compiled and tested by researchers in the fields of natural language processing or computational linguistics.

The availability of pre-compiled sentiment dictionaries – such as the MPQA Subjectivity Lexicon (Wilson et al., 2005), the Hu & Liu Opinion Lexicon (Hu and Liu, 2004), SentiWordNet (Baccianella et al., 2010), AFINN (Nielsen, 2011) or EmoLex (Mohammad and Turney, 2013) – makes dictionary-based sentiment analysis an attractive option for researchers simply due to its simplicity. In fact, popular text analysis software packages such as R's *tidytext* (Silge and Robinson, 2017) or Python's NLTK (Bird et al., 2009) come with a variety of sentiment dictionaries pre-installed, making it very easy for researchers to try out dictionary approaches on their text data. The major advantages of dictionary-based sentiment analysis approaches lie in their speed and simplicity, allowing researchers to rapidly assess the sentiment expressed in very large datasets with minimal use of resources. While this makes these approaches attractive for certain applications, especially where real-time analysis is required, significant concerns over the accuracy of these methods limit their potential in political science research.

The major weakness of these pre-compiled dictionaries is that they are by nature very generalised and lack the domain-specific

features which might be required for sentiment analysis of a topic that has specific associated vocabulary and language features. On a broad level, this means that some dictionaries are not appropriate for classifying certain types of text – a dictionary trained using a corpus of formal language (such as the text content of Wikipedia or a large corpus of newspaper articles) would perform poorly in classifying the sentiment in a casual, slang-heavy corpus of social media posts, and vice versa. To address this problem, domain-specific pre-compiled sentiment dictionaries have been created: AFINN and the VADER Sentiment Lexicon (Gilbert and Hutto, 2014) are designed for the kind of text commonly found in social media posts, while other dictionaries encompass vocabulary related to a specific topic (for example, the Loughran & McDonald sentiment lexicon for financial texts: Loughran and McDonald, 2011). These dictionaries go some way towards making lexicon-based approaches viable for a wider range of applications, but fundamental problems remain. Dictionaries are static, while language is ever-evolving – especially in the political domain, where words and phrases can take on new meanings or implications very rapidly during the course of an electoral campaign or political news cycle.

A further challenge for dictionary approaches relates to how the scoring of the lexical dictionary is applied to the text. In the simplest possible model, the positive and negative sentiment scores for each word in the text are added together to yield a final score. While this approach is intuitive, its drawbacks are easy to see. Consider the following sentence from a review of a movie:

*The movie has a fantastic cast, an interesting concept and amazing special effects – but it is utterly boring.*

A naïve calculation based on a sentiment dictionary would almost certainly find that this sentence was positive due to the presence of a large number of positive words ('fantastic', 'interesting', 'amazing', 'special') and

only one negative word ('boring') – yet to a human reader it's very obvious that the overall tone of the sentence is in fact negative. In order to improve on the accuracy of this kind of approach, algorithms have been developed which take into account additional features of the text. Where the naïve dictionary approach is a pure 'Bag of Words' approach, which is common to many text analysis methodologies and treats a document purely as a collection of words, ignoring the order of those words or the relationships between them, more advanced algorithms draw instead on the field of Natural Language Processing (NLP). A simple but effective example is the VADER algorithm, which incorporates a number of features that improve the accuracy of dictionary-based classification – for example, it understands various forms of negation (so 'not good' and 'this wasn't good' would both be treated as negative despite 'good' being a positive word), contrastive conjunctions such as 'but' or 'however', words which alter the intensity of a sentiment (like 'utterly' in the example sentence above) and the presence of emoji in social media posts. These improvements make VADER and similar algorithms more effective than naïve dictionary-based sentiment scoring without increasing the complexity for researchers using the tool. Even more advanced ways of applying sentiment dictionaries continue to be developed by NLP researchers, but often introduce a degree of technical complexity or resource-intensive computation which puts them beyond the reach of most researchers seeking to implement a practical sentiment analysis: the Stanford Sentiment Treebank (Socher et al., 2013), for example, achieved an improvement in classification accuracy over other approaches by using an artificial neural network to process sentence structure, but this introduces a computational and technical overhead which is likely to be impractical for most researchers.

The domain-specificity of sentiment dictionaries means that a given dictionary can

only be used effectively to classify sentiments in text that follows broadly the same linguistic styles and standards as the corpus used to create that dictionary. As a consequence, while a large number of sentiment dictionaries exist for English, and dictionaries have also been created for other major languages, there are a significant number of languages for which few, if any, high quality sentiment dictionaries exist (Mohammad, 2016). This significantly restricts the potential for using sentiment dictionary approaches in analysis of text in minority languages, or in cross-linguistic analyses; while attempts to apply English-language sentiment dictionaries to other languages using machine translation tools have been made, there is little evidence to support the effectiveness of such an approach. Nonetheless, sentiment dictionaries remain a useful set of techniques for use in exploratory analysis of data, or for providing supporting evidence in research where sentiment analysis is not the primary analytic tool being employed.

### ***Supervised Learning Approaches***

Unlike dictionary-based approaches, which make *a priori* assumptions about the positive or negative meanings of specific words in a lexicon, supervised learning approaches begin with a blank slate each time, making no assumptions about the sentiment or meaning of any word in the corpus. Instead, a subset of the corpus is classified by human coders and this labelled set is used to train a machine learning algorithm, which will then estimate the sentiment of the rest of the corpus based on the patterns and correspondences found in the labelled set. One set of algorithms, the classification algorithms, attempt to accurately classify the sentiment of each individual unit of text in the corpus (in this regard resembling the dictionary approach described above); another, the aggregate algorithms (sometimes called proportional classification algorithms:

Wiedemann, 2018), focus instead on accurately estimating the distribution of sentiment across the entire corpus.

The drawback of such approaches is immediately obvious – you cannot proceed without first creating labelled data. While a pre-compiled sentiment dictionary allows you to instantly start exploring the sentiments in a corpus of text data, supervised learning approaches require that you first classify a sample of the data by hand. This brings with it the usual problems associated with human content coding – the need to train coders, to verify the consistency of their work using inter-coder reliability measures and to decide on an appropriate sampling approach and sample size to classify.

Of these, the question of how much data is needed to train an algorithm is a common sticking point for researchers. The most common response is ‘as much as possible’; Hopkins and King (2010) offer 500 as a rule of thumb. While algorithms have been shown to have a learning curve which eventually flattens out, leading to diminishing returns from the creation of additional labelled data (Figuroa et al., 2012), the complexity of text classification and the resource constraints faced by most researchers make it unlikely that they will reach this point, so as a general rule more labelled data is better. In more practical terms, we can think of the volume of labelled data required as a function of the number of inputs (i.e. the size of the vocabulary being used in the texts) and the number of outputs (the number of categories to classify). A text corpus which addresses a wide range of different issues will require more labelled data than a more narrowly focused corpus; for example, a set of social media posts addressing a specific policy initiative will have a smaller lexicon of relevant vocabulary than a set of posts about a candidate in a major election, as the latter will likely include posts discussing a range of different issues and policies. As such, more labelled data would be required for the latter data set, even if the actual sizes of the two corpora

were the same. The choice of algorithm is also relevant to this question; as a general rule, aggregate algorithms require less labelled data than classification algorithms.

The other main challenge to keep in mind when dealing with building an appropriate and reliable set of labelled data relates to the sampling of documents. Classification methods implicitly assume that the labelled set is a random sample from the population of documents to be coded, in terms of the opinions expressed (but see below the discussion about aggregate algorithms) as well as, crucially, the language employed. This is because supervised learning methods use the relationships between categories and features in the labelled set to classify the remaining documents. Such algorithms are flexible enough to classify any kind of data that can be broken down into ‘tokens’ (features), but simply cannot classify the sentiment of any feature which was not encountered in the labelled data. As such, if an important word or other such feature (like an emoji or a hashtag) was not included in the texts classified by human coders, the trained algorithm will miss its relevance to the sentiment of the texts it classifies. This presents particular difficulty when all the data are not available at the time of coding of the labelled set, for example in situations where all the data have not yet been digitised or, as is often the case with sentiment analysis on social media, where data will continue to be produced in the future. This causes situations where the difference in the language between the labelled set and the unlabelled set grows over time, such as when new words and phrases, or new meanings of existing words and phrases, appear in the latter set but not in the former set. For example, a politician may introduce a new slogan during a speech which was not part of the common lexicon at the outset of a campaign period, or a common word may take on a new meaning through association with a scandal or a major policy. Alternatively, the inverse may occur, and words, phrases and their meanings may exist in the labelled set

but not in the unlabelled data. This is most common in long-term analyses; for example, in the years after the fall of the Berlin Wall, terminology related to the Cold War largely disappeared from political rhetoric. Some of these challenges may be partially addressed with the use of distributed word embeddings, a technique which has become popular in natural language processing and computational linguistics (Goldberg, 2016; Pennington et al., 2014; see also Chapters 26 and 55 of this *Handbook*). Word embeddings represent words in a continuous vector space in which words with similar meanings are mapped closer to each other. Using a set of word embeddings appropriate to the texts being analysed can allow new words encountered in the unlabelled texts to be classified according to their similarity to words that were present in the labelled texts (Rudkowsky et al., 2018). This does not, however, account for the arising of new meanings for words, such as the naming of new scandals – word embeddings produced from early 1970s texts would correctly identify ‘Watergate’ as being close in meaning to the word ‘hotel’, but would need to be regenerated from new texts to identify its new meaning closer to ‘scandal’ and its consequent importance to political sentiment.

Once the sample has been classified, the subsequent process is broadly the same regardless of which kind of algorithm is used. As a result, researchers often train multiple algorithms and test to see which approach, and which algorithm, is more effective for their specific case. This testing is performed by repeatedly splitting the labelled sample data into ‘training’ and ‘test’ sets, training the algorithm on the former and then testing it on the held-out data in the latter – a process known as cross-validation. The algorithm which performs most effectively and consistently in these tests is then used to classify the entirely unseen texts in unlabelled data set.

For the purposes of explaining classification and aggregate algorithms, let us assume that we have a labelled corpus (classified

by human coders) of  $N$  distinct texts, sampled from the overall corpus (which we label  $C$ ) we wish to classify. Let us denote by  $D = \{D0, D1, D2, \dots, DM\}$  the set of  $M + 1$  possible *categories* (that is, sentiments) expressed in the texts.  $D0$  denotes the likely most prevalent category in the data – generally, ‘off-topic’ texts or those which express opinions/sentiments not relevant with respect to the analysis, that is, the *noise* in this framework. Noise is commonly present in any corpus of texts crawled from social media and the internet in general and should be taken into account when we evaluate a classification method (a point to which we will return below).

Suppose that after pre-processing the text (removing irrelevant words and features, stemming or lemmatising as appropriate, and so on) we are left with  $L$  features (word stems, or tokens). The document-term matrix  $S$  representing our corpus then has  $N$  rows and  $L$  columns, with each document  $j$  being represented by a vector  $S_j$  of length  $L$ . The value of each element of the vector  $S_j$  may either be the frequency with which that feature appeared in document  $j$ , or a binary value – 1 if the feature appeared at all, 0 if it did not. This latter representation is commonly used for short documents such as social media posts, since the frequency of a given word is most likely to be 0 or 1 anyway; for longer documents the range of word frequencies is much greater and of more importance. If the lengths of the documents being studied differ significantly (which is usually not the case for social media posts) it is also necessary to account for this in their vector space representation, for example by weighting the vocabulary frequency inversely to the length of the document. Note that the overall document-term matrix ( $S$ ) is often quite a large data set but very sparse, in that it contains a large number of zero values since each document only contains a small subset of the overall lexicon of the corpus  $L$ .

Regardless of the subsequent approach taken to sentiment analysis, this document-term

matrix  $S$  will be the basic data set used. To ensure the robustness of the entire estimation strategy, cross-validation is performed. This process entails dividing the labelled data into a number of equal-length, randomised segments (usually between 5 and 10) and using each segment in turn to train the algorithm being tested. Each of these trained algorithms is then tested against the remainder of the labelled data, which is called the ‘held-out’ data – so for example, if the labelled data is divided into five segments, the algorithm being evaluated would be trained five times, each time on 20% of the data, and tested each time against the remaining, or held-out, 80% of the data. If an algorithm performs well on each step of cross-validation (for example in terms of its accuracy, that is, the proportion of correctly classified documents)<sup>2</sup> it can be expected to perform reliably on entirely unseen data as well (that is,  $C-N$ ); conversely, an algorithm which fails on a given cross-validation step may be poorly suited to a specific permutation of features that arises in the data.

### *Classification algorithms*

There is a very wide variety of different classification algorithms, ranging from those which will be familiar to most researchers with a background in statistics, such as Naïve Bayes, through to complex and computationally intensive algorithms such as Neural Networks and Random Forests (see Dreiseitl and Ohno-Machado, 2002 for details of many of these models). The differences between classification algorithms are the subject of very extensive literature in statistics and computer science which is beyond the scope of this chapter – an in-depth overview of the field as a whole is provided in Aggarwal (2014). For the purposes of the political science researcher attempting to perform sentiment analysis, the key thing to note is that the training process for any one of these models takes the same input (a training subset of the term-document matrix  $S$  described above) and gives the same output,

a model (or function) which predicts the category  $D_j$  to which a given document  $j$  belongs given that its features are represented by the vector  $S_j$ . This model can be represented as  $P(D|S)$  – for a given document  $j$  and set of categories  $M$ , it will find the value of  $m$ , that is, the classification, which maximises the model  $P(D_{m=0,1,\dots,M} | S_j)$ . Expressing this as a matrix model, the classification algorithm is  $P(D) = P(D|S) P(S)$  – where  $P(D)$  is a vector of length  $M+1$ ,  $P(D|S)$  is a matrix of conditional probabilities and  $P(S)$  is a vector representing the distribution of text vectors across the corpus of texts. This means that regardless of the broad differences in how the algorithms function internally, it is possible to train them using the same data, directly compare their results, and select the one which is most effective for your specific data and usage case. In fact, popular machine learning packages such as Python’s *scikit-learn* (Pedregosa et al., 2011) provide pipeline structures which are designed to automate this entire process of training, testing and comparing algorithms.

Once trained, classification algorithms can be used to predict the sentiment classification of an entirely unseen text. This approach has some significant advantages over the dictionary-based approach. It is by nature domain-specific; the algorithms base their predictions on words or patterns which were relevant to the classification categories in the training set and can often identify the sentiment of a phrase which would be entirely missed by a dictionary approach. A good example is political catchphrases – in the 2016 US presidential election, the phrase ‘lock her up’ was strongly associated with negative sentiment towards Hillary Clinton, while ‘build the wall’ was associated with positive sentiment towards Donald Trump. An effective classification algorithm trained on tweets related to the election would correctly identify the relevance of those terms, while a sentiment dictionary approach would treat both phrases plausibly as being neutral. Classification algorithms are also language-independent;

while some languages require more complex pre-processing steps before the algorithms can be applied, the algorithms themselves do not know or care which language the words (or ‘tokens’) are in, making this approach perfectly effective even in ‘resource-scarce’ languages which lack high-quality sentiment dictionaries.<sup>3</sup>

Nevertheless, classification algorithms are by no means perfect. This is in part due to the properties of the sentiment analysis task itself: not only is language intrinsically stochastic, but there is also a subjective element to the sentiment classification of many edge cases and disagreement even between human coders is normal. Moreover, classification algorithms generally rely on the Bag of Words approach and are unaware of sentence structure; this provides for fast and effective classification but fails on certain edge cases (for example, the Bag of Words representation of the sentence ‘I’m voting for Clinton because Trump is the most corrupt candidate ever’ is exactly the same as the representation for ‘I’m voting for Trump because Clinton is the most corrupt candidate ever’, so a classifier would be unable to distinguish between these polar opposite sentiments). As such, no classifier should be expected to achieve or even approach 100% accuracy in testing against held-out data – in fact, approaching 100% accuracy is likely a sign that the algorithm has been overfitted to the training data and will not be effective at classifying unseen data.

A further problem with classification algorithms can arise when a single category is predominant in the data. This means that during the training process the algorithm ‘learns’ that the vast majority of feature vector configurations are associated with this dominant category  $D_0$ ; only a small subset of specific vectors express any other category  $D_j$ . As a result, for a large proportion of input vectors  $S_j$  the output of  $P(D_i|S_j)$  will be zero, or extremely low, for all categories other than the dominant one ( $i=0$ ). The trained model will therefore tend to overestimate this

category vastly over-represented compared to others when classifying unseen texts. This effect is especially strong when the dominant category is a ‘Off-Topic’ category (a catch-all featuring all documents which did not include a topic or sentiment of interest to the study) as is commonly the case with social media data. This category will contain all of the ‘noise’ that is present in data collected from social media; only a specific subset of input vectors will correspond to the other categories being trained, with every other possible vector (a much larger domain) being implicitly ‘noise’. Depending on the objectives of the researcher, this over-estimation of a dominant category can sometimes be unproblematic – but where the objective is to find the overall distribution of sentiment within a data set, this systematic overestimation can result in significant bias. In the next section we will discuss a new class of algorithms which have been developed also to deal with this bias.

### *Aggregate algorithms*

While classification algorithms aim to predict the sentiment of every individual piece of text in a data set, it is often the case that a researcher is looking for the distribution of sentiment within a corpus of texts and does not actually require an item-level classification. As a result, researchers end up aggregating the item-level classifications, which means that some of the trade-offs that are made by classifiers in the name of providing a single classification for every item – such as the issue of over-estimation of a dominant category, as discussed above – are actually unnecessary and may be needlessly reducing the accuracy of the prediction on the overall distribution. For example, a method that classifies 60% of documents correctly into one of eight categories might be judged successful and useful for classification. However, because the individual category percentages still might be off by as much as 40 percentage points, the same classifier may be useless for some social science purposes

(if individual-level errors do not cancel each other out). Recognising this problem has led to the development of a new set of algorithms which treat the text corpus in an aggregate manner from the outset – they do not provide classifications for individual texts at any point in the process, instead producing results showing the distribution of sentiments within the overall corpus. The first of these algorithms to be released was ReadMe (Hopkins and King, 2010), which has since been joined by iSA (Ceron et al., 2016); while these algorithms have different strengths, they use similar inputs and produce similar outputs, making it possible to train and test them against one another in a similar manner to the algorithm selection process used for classification algorithms.

These aggregate approaches are less flexible than classification approaches, since they cannot be applied to any research that actually requires the classification of individual texts at any point – but in return they offer two major advantages. First, because they are not constrained by the need to produce a classification for every text, they can treat classifications as probabilities rather than binary predictions – so instead of decisively coming down on the side of a text being either ‘positive’ or ‘negative’ (or whatever categories are being used) when it actually shows some features of multiple categories, the algorithm can in essence treat it as being partially positive and partially negative when calculating its impact on the sentiment distribution of the entire corpus. This kind of approach, moreover, can deal well with situations in which the relative frequency of categories in the training set is very unevenly distributed – such as in the above mentioned case of a category containing off-topic ‘noise’ being statistically dominant within the data set – or where the distribution in the training set is quite different from the frequency in the unseen data set.

Expressed in statistical terms, aggregate algorithms follow from the insight of Hopkins and King (2010) that the estimation process used by classification algorithms and

shown in the prior section –  $P(D) = P(D|S)P(S)$  – can be reversed; they propose a solution as follows:

$$P(D) = [P(S|D)^T P(S|D)]^{-1} P(S|D)^T P(S)$$

In essence, this formula uses an inverse matrix to allow the estimation not of  $P(D|S)$  (the probability of category  $D$  given vector  $S$ ), but of  $P(S|D)$  – the probability of vector  $S$  given category  $D$ . Given sufficient coded texts for each category  $D$ , it is actually possible to estimate this significantly more accurately than  $P(D|S)$  – the trade-off being that we must estimate it across the entire text corpus, not on a per-document basis.

The secondary benefit of these algorithms is related to the required properties of the labelled set of documents. First, it is often possible to achieve good results with a smaller set of labelled data than is required to train an accurate classification algorithm. Moreover, aggregate algorithms do not require that the labelled set should be a representative sample of the full population of texts; the only demand for selection of labelled data is that the language used in the labelled documents to express some given concept must be the same as in the whole population of texts, and that the labelled data must include sufficient examples of each individual category to be classified as to permit generalisation.

While aggregate algorithms are not perfectly suited to all kinds of analysis, their specific strengths make them particularly attractive not only for analyses that rely entirely on aggregate data, but also for real-time applications (where it is likely not possible to generate large amounts of training data on the fly) and for projects that lack the resources to undertake human coding of a large sample of the data.

## SENTIMENTS ON SOCIAL MEDIA

One of the most common applications of sentiment analysis in recent years has been in

uncovering the sentiments being expressed about topics on social media. While the field of sentiment analysis itself far predates social media (or indeed widespread internet usage), the volumes of text data produced on social media are often orders of magnitude greater than any prior data source in the political or social science fields. Combined with the unprecedented level of access to the preferences and views of a large section of the public represented by social media posts, this has made the ability to programmatically calculate sentiment across large datasets extremely important.<sup>4</sup>

Using social media data for sentiment analysis introduces a number of opportunities and challenges for researchers (some of them already discussed in the previous pages). In essence, social media data is generally made up of a large number of short texts, which makes it ideal for this kind of analysis. When applying sentiment analysis to longer texts such as newspaper articles or political speeches, it is necessary to decide what the actual unit of analysis will be – an entire document, a paragraph or an individual sentence. While some social media platforms do permit the posting of long texts, posts are generally short enough to settle on the individual post as a unit of analysis – especially on Twitter, a platform whose open nature makes it especially useful to researchers and whose strict character limit makes it easy to treat every post as a distinct unit for analysis.

### ***The Pre-Processing Stage***

Social media posts are often written and presented in a very different style to formal texts such as newspaper articles or speeches, however. Slang, abbreviations and shorthand are common (especially on Twitter, whose tight character limits encourage brevity). Along with platform-specific text features like @usernames and #hashtags, posts may also include images, web links and emoji (pictographs). In some regions, users commonly

construct faces from punctuation marks and other characters to express emotions using ‘smileys’ or ‘kaomoji’. These text features create special challenges for researchers, not least because many kinds of text analysis software can’t handle them – it’s not uncommon for punctuation marks to be removed in pre-processing (which therefore removes smileys and kaomoji) and for emoji either to be ignored or to result in software errors. Handling of web links can also create problems; software often splits links up into multiple tokens (e.g. changing ‘http://www.google.com’ into ‘http’, ‘www’, ‘google’ and ‘com’) and the high incidence of tokens like ‘http’ and ‘www’ (which appear in every URL) can result in feeding bad data to the sentiment analysis process. It is therefore important for researchers to pre-process social media data (texts) in a way that is sensitive to these specific text features; in particular, approaches which dispose of emoji or smileys should be avoided, since these features often have a very direct impact on the sentimental content of a text. A single emoji at the end of a sentence can entirely change the sentimental tone of that sentence, so any sentiment analysis approach which ignores emoji is excluding important data from the outset.<sup>5</sup>

### ***The Data Access Riddle***

A further problem with social media data relates to the relevance and reliability of the data itself. Gathering representative data from social media given the limitations placed by factors such as API limits (programmatic limitations on the amount of data that can be downloaded in a given time period) and account privacy settings is a major challenge that is beyond the scope of this chapter, but researchers wishing to perform sentiment analysis on social media data need to make themselves aware of these issues and develop a strategy to tackle them. More specifically related to sentiment



analysis is the problem of keyword specificity – the keywords used to gather and filter social media data must be chosen very carefully in order to ensure that the data is actually relevant to the topic being studied. There are several problems which can emerge in this regard, the simplest of which is the use of a keyword which actually has different meanings in different cultural or geographic contexts. For example, in a project on the Kenyan elections which one of the authors participated in, the name of a presidential candidate, ‘Uhuru’, was used as a search keyword on Twitter; it only emerged in later analysis that this word is also used in a completely different way by some right-wing groups in the United States, resulting in a large amount of irrelevant data being downloaded. Without removing these irrelevant results from the data by some means, sentiment analysis would have given skewed results due to calculating sentiment for posts that actually had nothing to do with the election we were studying. This problem can be tackled in several different ways: limiting the posts gathered to a specific country or language can often be effective, but if this is impossible, adding the already discussed ‘off-topic’ classification to a supervised learning approach can allow human coders to effectively train the algorithm to recognise these irrelevant posts. A more complex problem arises when posts include multiple keywords which are being studied; for example, a post might mention the names of both candidates in an election, making it difficult to judge which candidate the sentiment classification actually relates to. One approach which can be effective in solving this problem is training a supervised learning algorithm to identify not just overall sentiment, but sentiment related to specific topics – so for example, in analysing a corpus of tweets related to the 2016 election, an algorithm might be trained to identify ‘Trump-Positive’, ‘Trump-Negative’, ‘Clinton-Positive’, ‘Clinton-Negative’ and ‘Off-Topic’ classifications (‘Neutral’ might

also be an option) instead of a simple polar ‘Positive’ or ‘Negative’ classification.

### ***Non-European Languages***

Related to the challenges and opportunities with respect to social media data is a further set of challenges which arise from working with data in non-European languages. In particular, different languages often require different pre-processing before being converted into a Bag of Words or a document–term matrix, a tabular view with the corpus’ vocabulary on one axis and the documents themselves on another, allowing each document to be represented as a sparse vector of vocabulary frequencies. East Asian languages such as Chinese and Japanese, for example, do not use spaces to separate their words and therefore require a more complex ‘tokenisation’ step for dividing up sentences into their component vocabulary. There is also a variety of different approaches to simplifying the vocabulary whose usage differs depending on the language being analysed: ‘stemming’ (which converts all conjugated words into their dictionary form) and ‘lemmatisation’ (which reduces words back to a more basic semantic meaning, so for example the words ‘preside’, ‘president’ and ‘presidential’ might all be converted into ‘presid’) require different approaches in different languages, and may not be applicable to some languages at all.

A further complication arises from the lack of high-quality resources for sentiment analysis in some languages. As already noted, this is especially problematic for dictionary- or lexicon-based approaches; while there are a wealth of different sentiment dictionaries and algorithms available for English, some languages may have none at all, or may only have poor-quality or incomplete dictionaries. In these cases, supervised learning approaches are essential; the great advantage of supervised learning (both classifier and aggregate approaches) is that the algorithms

are entirely language-agnostic. In fact, many of these algorithms are also used to classify data that isn't even text, such as sounds or images – as long as the features encountered in the unlabelled data to be classified also appeared in the labelled documents, it doesn't matter at all what those features actually are. Supervised learning algorithms will classify a document–term matrix of Japanese words just as easily as one made up of English words and will not care if some of those 'words' are actually other text features like emoji, smileys or hashtags.

## A WORKED EXAMPLE

We will conclude with a practical example of the different techniques we have outlined above, showing the differences between the methods and their outcomes and pointing out potential pitfalls researchers may encounter with different sets of data. For the purposes of this example, we will use the Stanford Large Movie Review Dataset (Maas et al., 2011) – a data set which includes 50,000 movie reviews drawn from the IMDb website and classified as 'positive' and 'negative' according to the score the user gave the film, with reviews scored 1 to 4 considered 'negative' and those scored 7 to 10 considered 'positive'. Since these are user-submitted reviews, they include many of the same features we would expect from social media texts – misspellings, slang and grammatical mistakes are commonplace, which can pose problems for some sentiment analysis approaches. For the purposes of this example we randomly sampled 5,000 reviews to treat as labelled data, with the remaining 45,000 being treated fictionally as unlabelled data and used to test the accuracy of the various approaches.

For this worked example, we will use Python with a number of popular and readily available scientific packages. A full version of the script, with additional information and

details of the packages and techniques used, is available in the online Appendix.<sup>6</sup>

## *Loading and Pre-Processing*

We first load the data into the script – in this case, the data set is distributed as a large folder of text files, one review per file, but it is also common to need to import data from CSV files, Excel files or databases. Regardless of the source, it is often easiest to convert the data into a list of text strings. For training purposes, we also need the classifications of the test data, which should also be stored in a list (in the same order as the documents to which the classifications belong).

The pre-processing which needs to be carried out on the text differs according to the text source and the language being analysed, but this case is typical – we remove HTML tags from the text, convert it all to lowercase and strip out all punctuation. For longer texts, it may be preferable to divide the text into sentences so that a sentence-by-sentence sentiment analysis can be carried out; for these short texts, we ignore sentence boundaries and other punctuation.

## *Dictionary/Lexicon Sentiment Analysis*

To demonstrate the performance of dictionary analysis approaches on these texts, we can use two of the built-in dictionaries in the NLTK (Natural Language ToolKit) package for Python – the Liu & Hu Sentiment Lexicon and the VADER Sentiment Scoring tool. The Liu & Hu lexicon is the simplest approach – for this we simply divide up each review into its component words (a process known as 'tokenisation') and check each word to see if it is listed as having a positive or negative polarity in the lexicon. Documents with more positive than negative words are treated as 'positive' and vice versa; documents

whose negative and positive words cancel one another out (or which have no such words) are ‘neutral’. Since this approach requires no ‘training’, we could employ the dictionary to directly label all the 50,000 reviews in our corpus. However, to allow a direct comparison of dictionary sentiment analysis with the supervised approaches employed below, we have instead estimated the scores for the 45,000 ‘unlabelled’ reviews. Despite the simplicity of the approach, it managed to correctly score 69.6% of the reviews in this corpus – detailed results can be seen in Table 29.1.

The second approach we use is the VADER Sentiment Analysis algorithm. This is a more context-sensitive approach to sentiment scoring and gives a ‘polarity score’ for each document – a score scaled between  $-1$  (completely negative) and  $+1$  (completely positive). This can be useful in some situations and VADER can often correctly identify the sentiment of edge cases which more simple lexicon approaches miss. However, on this set of test documents VADER yields largely similar results to the simpler Liu & Hu lexicon, with 69.6% of reviews correctly classified – largely due to a very high rate of negative reviews being incorrectly classified as positive. The algorithm was much more successful at classifying positive reviews, achieving 85.2% accuracy on these.

Note that both of these approaches are only applicable to the English language – for any other language, a different lexicon would need to be used (or created).

## Preparing a Document–Term Matrix

Most approaches to sentiment analysis rely on the previously mentioned Bag of Words approach, which represents each document as a collection of words – ignoring their position and relationship to other words. While this loses significant amounts of information, it has proved very effective for a range of text mining applications. For a simple approach like the Liu & Hu lexicon analysis shown above, it’s sufficient simply to divide the sentence up into its component words, but more complex approaches such as classification algorithms require a development of a *vector space model* – which converts each document into a vector recording the presence or frequency of each vocabulary word. This representation of a corpus is known as a *document–term matrix*, with each row being a document, and each column a vocabulary word. This matrix can grow extremely large since the range of vocabulary in a corpus is often in the order of thousands of words. Two key approaches are used to manage this; first, the vocabulary list is trimmed by excluding very common and uncommon words (which are of little use for distinguishing documents from one another) and by stemming or lemmatisation – processes which reduce words down to shorter ‘stems’ representing their core meaning. It is often possible to reduce a corpus’ vocabulary list to a few hundred words in this way. Second, the document–term matrix is stored as a ‘sparse matrix’ – an approach allowing

**Table 29.1 Results for Liu & Hu sentiment lexicon**

	<i>Actual-Negative</i>	<i>Actual-Positive</i>
<i>Predicted-Negative</i>	14597	4475
<i>Predicted-Neutral</i>	1662	1300
<i>Predicted-Positive</i>	6234	16732

**Table 29.2 Results for VADER sentiment analysis**

	<i>Actual-Negative</i>	<i>Actual-Positive</i>
<i>Predicted-Negative</i>	12142	3329
<i>Predicted-Neutral</i>	20	9
<i>Predicted-Positive</i>	10331	19169

computers to store large matrices where most values are zero in a fraction of the space normally required.

To construct a document–term matrix, we need to specify a tokeniser (a software function which splits sentences into component words, then applies stemming, lemmatisation and other such processes). In English and other European languages this step is relatively simple due to the spaces between words; in languages such as Japanese or Chinese, an additional piece of software is required to detect the word boundaries and carry out tokenisation. For Chinese, the Stanford Segmenter is a popular tool, while analysing Japanese is usually carried out using MeCab or Janome. Other languages, such as Korean, also require this step. A mistake in this tokenisation process can result in ‘junk data’ in the document–term matrix, so it’s important to pick appropriate tools for the language you are working with.

Next, the tokeniser is used to train a ‘vectoriser’ which will analyse the corpus to find which vocabulary terms are relevant; after this ‘fitting’ process, it can be used to generate a document–term matrix for the entire corpus, or a vector for a single unseen piece of text which will be compatible with the vectors in the document–term matrix (that is, each position on the vector will represent the same vocabulary term as it would in a row of the matrix).

## Classification Algorithms

The process of training and testing a classification model involves ‘fitting’ the model to a set of training data and then scoring its performance at predicting the classifications of a set of test data. This process is carried out using multiple different models in order to select the most effective one for a specific set of data. Using the movie review data set, we tested a selection of classification models using a ‘labelled’ set of 5,000 randomly selected reviews. To ensure the robustness of this process, *cross-validation* is used, wherein the data is divided into multiple ‘folds’ and the model is repeatedly trained and tested on different folds of data. In this example, the cross-validation process uses three folds (five to ten folds are also commonly used; the more labelled data you have, the more folds you can effectively use in this process), so the 5,000 labelled set is randomly divided into three subsets. We reserve one subset (the test set) and train the model on all other subsets (the training set). We then test the model on the reserved subset and record the prediction error, before repeating this process until each of the three subsets has served as the test set. We finally compute the average of the three recorded errors. This is called the cross-validation error and serves as a performance metric for the model. The results for the models tested can be seen in Table 29.3.

**Table 29.3 Initial results for classification algorithms**

<i>Model Name</i>	<i>Average Accuracy</i>	<i>Cross-validation Error</i>
Multinomial Naïve Bayes	0.822	+/- 0.030
Bernoulli Naïve Bayes	0.808	+/- 0.015
Support Vector Classifier	0.813	+/- 0.014
Linear Support Vector Classifier	0.501	+/- 0.000
Stochastic Gradient Descent	0.800	+/- 0.028
Random Forest (10 trees)	0.730	+/- 0.019
Neural Network (Multi-Layer Perceptron)	0.811	+/- 0.021

There is no one-size-fits-all classification algorithm; in this specific instance, the linear support vector classifier performed worst, with the random forest also achieving only marginally better scores than the dictionary-based approaches outlined above. For other data, however, these algorithms may outperform the others; it is important to test a variety of algorithms to see which is best suited to the specific combination of data and classifications being used. In this instance, the Multinomial Naïve Bayes classifier proved the most effective, with 82.2% accuracy on the held-out data set.

This is not the end of the process of training an effective classification model, however. Most algorithms also have a range of ‘hyper-parameters’ – assumptions and modifiers which can be set to different values prior to training – that can significantly impact performance. Finding the right set of hyper-parameters for a certain task is also largely a case of trial and error. However, several different packages, both in Python and R, provide ways to automate this task; this is known as a ‘grid search’, allowing researchers to exhaustively search through every combination of a set of hyper-parameters to find the best performing model. This process can take a lot of time – often in the order of several hours for algorithms with complex sets of parameters – but often yields better performance than the default parameter set. For example, in our case we were able to locate a set of parameters for the support vector classifier model which boosted its performance to 82.3% – a full percentage point higher than the algorithm’s performance in our initial test. This model would likely be considered sufficiently accurate and reliable for use in a research project. In the present example, it is possible to test that accuracy by scoring the trained algorithm’s performance on the ‘unlabelled’ set of 45,000 reviews; for the final trained algorithm (the support vector classifier with the best hyper-parameters chosen through the grid search procedure), those results are shown in Table 29.4. The

**Table 29.4 Results for final classification algorithm**

	<i>Actual-Negative</i>	<i>Actual-Positive</i>
<i>Predicted-Negative</i>	18261	3464
<i>Predicted-Positive</i>	4232	19043

algorithm achieved 82.9% accuracy overall, 84.6% on the positive reviews and 81.2% on the negative reviews (more detailed scoring, including the precision, recall and f1-score metrics mentioned above, can be found in the online Appendix). Note that in a real research situation you would have to trust the accuracy of your trained algorithm at this point, since the rest of your data really would be unlabelled.

One additional possibility which we have not explored in this worked example is combining different algorithms to achieve better accuracy than a single algorithm would manage on its own. This is called an ‘ensemble’ approach and comes in two basic forms. The first, ‘model averaging’, involves training multiple different algorithms and then combining the results of those which showed acceptable performance on the test set, either by blending their predictions or by running them independently and allowing them to ‘vote’ on the classification of the text. The second, ‘model stacking’, was introduced by Wolpert (1992) and involves feeding the predictions from an ensemble of algorithms into another algorithm as parameters for a final prediction. Further details of these two approaches can be found in Sardinha (2017).

### **Aggregate Algorithms**

Although aggregate algorithms function very differently from classification algorithms in many regards, they take exactly the same input – the document–term matrix – so we can re-use the matrix from the previous step.

For this example we are using the Python version of iSA, PyiSA – at the time of writing there is no Python implementation of the other major aggregate algorithm, ReadMe, but this example could be recreated in R to test ReadMe’s performance.

The procedure followed by iSA is very similar to the one used by the classification models shown above. First, we use the labelled documents to train the algorithm. Then the algorithm will predict the classification of the unlabelled documents. Unlike the classification models, however, iSA returns an estimate of the distribution of each classification within the overall corpus and also estimates its own standard errors. If you wished, it would also be possible to run a cross-validation of the iSA algorithm by checking the consistency of results from different sub-folds of the labelled data – we have not done so in this instance, but the procedure would be functionally the same as for a classification algorithm.

In our test on the movie reviews, iSA estimated that 51.4% of the reviews were negative (std. error 0.01) and 48.6% were positive (std. error 0.01) – very close to the actual 50:50 distribution within the overall corpus. While the classification algorithms also came close to estimating the correct distribution, this was largely due to the balance of mis-classifications (positives labelled as negative, and vice versa) being evenly distributed. If the sentiments were not evenly divided in the corpus, or if there were additional categories (e.g. a ‘neutral’ or ‘off-topic’ category), as already noted above, this could have produced significant skew in the results of the classification algorithms, which the aggregate algorithms would have handled more robustly. Furthermore, the aggregate algorithms are better able to cope with smaller amounts of labelled training data, while the performance of classification algorithms can drop off quickly when the amount of training data is reduced. For example, if we reduce the amount of labelled data in the above example to 1,000

reviews (a reasonably realistic number for many research projects), the accuracy of the best classification algorithm we could train fell to 77.5% – a drop of more than 6% – while iSA’s estimation remains extremely accurate; in fact, its accuracy improved despite the smaller amount of training data, in this case estimating the split in the corpus at 50.1% to 49.9%.

## Notes

- 1 Sentiment analysis is just one type of analysis a dictionary method can perform. The general concept of dictionaries makes them relatively easy and cheap to apply across a variety of problems, including the identification of words that separate categories (for example policy categories) and measuring the frequency of those words in different texts. See Laver and Garry (2000) for an approach in this regard.
- 2 To assess model performance, beyond accuracy, other commonly used statistics are recall (a measure of what proportion of actual instances of a given category the algorithm correctly identified) and precision (a measure of how many of the times the algorithm identified a category were actually correct, as against how many times were false positives). The f1 score is another commonly used measurement that combines both recall and precision on a per-category basis.
- 3 For a supervised learning approach applied to Arabic language, see for example Ceron et al. (2019) or Cunliffe and Curini (2018); an example in the Japanese language can be found in Fahey et al. (2018).
- 4 For a recent review of works within the social science field that apply such approach, see Ceron et al. (2017).
- 5 The pre-processing stage of the analysis (that is, the series of decisions a researcher makes on which types of word or features of a text to include or exclude from analysis and how to treat those text features, for example whether to apply stemming and lemmatisation) can have large consequences for the quality of the results of any automated text analysis. See Haselmayer and Jenny (2017) for a study that shows how pre-processing decisions impact on sentiment analysis.
- 6 See: [http://robfahey.co.uk/sentiment\\_example.zip](http://robfahey.co.uk/sentiment_example.zip) or [http://www.luigicurini.com/uploads/6/7/9/8/67985527/appendix\\_handbook.rar](http://www.luigicurini.com/uploads/6/7/9/8/67985527/appendix_handbook.rar)

## REFERENCES

- Aggarwal, Charu C., ed. 2014. *Data Classification: Algorithms and Applications*. CRC Press.
- Baccianella, Stefano, Andrea Esuli and Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC '10)*, pp. 2200–2204.
- Bird, Steven, Ewan Klein and Edward Loper. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media.
- Ceron, Andrea, Luigi Curini and Stefano Maria Iacus. 2016. ISA: A Fast, Scalable and Accurate Algorithm for Sentiment Analysis of Social Media Content. *Information Sciences* 367–368: 105–124.
- Ceron, Andrea, Luigi Curini and Stefano Maria Iacus. 2017. *Politics and Big Data: Nowcasting and Forecasting Elections with Social Media*. Routledge.
- Ceron, Andrea, Luigi Curini and Stefano Maria Iacus. 2019. ISIS at Its Apogee: The Arabic Discourse on Twitter and What We Can Learn from That about ISIS Support and Foreign Fighters. *Sage Open*. <https://doi.org/10.1177/2158244018789229>
- Cunliffe, Emma and Luigi Curini. 2018. ISIS and Heritage Destruction: A Sentiment Analysis. *Antiquity* 92(364): 1094–1111.
- Dave, Kushal, Steve Lawrence and David M. Pennock. 2003. Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews. In *Proceedings of the 12th International Conference on World Wide Web*, pp. 519–528. ACM.
- Dreiseitl, Stephan and Lucila Ohno-Machado. 2002. Logistic Regression and Artificial Neural Network Classification Models: A Methodology Review. *Journal of Biomedical Informatics* 35(5–6): 352–359.
- Fahey, Robert A., Tetsuya Matsubayashi and Michiko Ueda. 2018. Tracking the Werther Effect on Social Media: Emotional Responses to Prominent Suicide Deaths on Twitter and Subsequent Increases in Suicide. *Social Science & Medicine* 219: 19–29.
- Figueroa, Rosa L., Qing Zeng-Treitler, Sasikiran Kandula and Long H. Ngo. 2012. Predicting Sample Size Required for Classification Performance. *BMC Medical Informatics and Decision Making* 12(1): 8.
- Goldberg, Yoav. 2016. A Primer on Neural Network Models for Natural Language Processing. *Journal of Artificial Intelligence Research* 57: 345–420.
- Grimmer, Justin and Brandon M. Stewart. 2013. Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis* 21(3): 267–297.
- Haselmayer, Martin and Marcelo Jenny. 2017. Sentiment Analysis of Political Communication: Combining a Dictionary Approach with Crowdcoding. *Quality & Quantity* 51(6): 2623–2646.
- Hopkins, Daniel J. and Gary King. 2010. A Method of Automated Nonparametric Content Analysis for Social Science. *American Journal of Political Science* 54(1): 229–247.
- Hu, Minqing and Bing Liu. 2004. Mining and Summarizing Customer Reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 168–177. ACM.
- Hutto, C. J. and Eric Gilbert. 2014. VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. In *Eighth International Conference on Weblogs and Social Media (ICWSM-14)*. <http://comp.social.gatech.edu/papers/icwsm14.vader.hutto.pdf>, accessed 5 November 2018.
- Laver, Michael and John Garry. 2000. Estimating Policy Positions from Political Texts. *American Journal of Political Science* 44(3): 619–634.
- Loughran, Tim and Bill McDonald. 2011. When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *The Journal of Finance* 66(1): 35–65.
- Maas, Andrew L., Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, Christopher Potts. 2011. Learning Word Vectors for Sentiment Analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 142–150. Portland, Oregon, USA: Association for Computational Linguistics. <http://www.aclweb.org/anthology/P11–1015>.

- Mohammad, Saif M. 2016. Sentiment Analysis: Detecting Valence, Emotions, and Other Affectual States from Text. In Meiselman, Herbert (ed.), *Emotion Measurement*, pp. 201–237. Elsevier.
- Mohammad, Saif M. and Peter D. Turney. 2013. Crowdsourcing a Word-Emotion Association Lexicon. *Computational Intelligence* 29(3): 436–465.
- Nielsen, Finn Årup. 2011. A New ANEW: Evaluation of a Word List for Sentiment Analysis in Microblogs. ArXiv Preprint ArXiv:1103.2903.
- Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, Edouard Duchesnay. 2011. Scikit-Learn: Machine Learning in Python. *Journal of Machine Learning Research* 12: 2825–2830.
- Pennington, Jeffrey, Richard Socher and Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. *In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543.
- Rooduijn, Matthijs and Teun Pauwels. 2011. Measuring Populism: Comparing Two Methods of Content Analysis. *West European Politics* 34(6): 1272–1283.
- Rudkowsky, Elena, Martin Haselmayer, Matthias Wastian, Jenny Marcelo, Stefan Emrich, Michael Sedlmair. 2018. More than Bags of Words: Sentiment Analysis with Word Embeddings. *Communication Methods and Measures* 12(2–3): 140–157.
- Sardinha, Jovan. 2017. An Introduction to Model Ensembling – Weights and Biases. *Medium*. <https://medium.com/weightsandbiases/an-introduction-to-model-ensembling-63effc2ca4b3>, accessed 24 June 2019.
- Silge, Julia and David Robinson. 2017. *Text Mining with R: A Tidy Approach*. O'Reilly Media.
- Smailović, Jasmina, Janez Kranjc, Miha Grčar, Martin Žnidaršič and Igor Mozetič. 2015. Monitoring the Twitter Sentiment during the Bulgarian Elections. *In Data Science and Advanced Analytics (DSAA), 2015*. 36678 2015. IEEE International Conference, pp. 1–10. IEEE.
- Socher, Richard, Alex Perelygin, Jean Y. Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, Christopher Potts. 2013. Recursive Deep Models for Semantic Compositionality over a Sentiment Treebank. *In Proceedings of EMNLP*, pp. 1631–1642.
- Wang, Hao, Dogan Can, Abe Kazemzadeh, François Bar and Shrikanth Narayanan. 2012. A System for Real-Time Twitter Sentiment Analysis of 2012 US Presidential Election Cycle. *In Proceedings of the ACL 2012 System Demonstrations*, pp. 115–120. Association for Computational Linguistics.
- Wiedemann, Gregor. 2018. Proportional Classification Revisited: Automatic Content Analysis of Political Manifestos Using Active Learning. *Social Science Computer Review* 37(2): pp. 135–159.
- Wilson, Theresa, Janyce Wiebe and Paul Hoffmann. 2005. Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. *In Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pp. 347–354. Association for Computational Linguistics.
- Wolpert, David H. 1992. Stacked Generalization. *Neural Networks* 5(2): 241–259.





# Big Relational Data: Network-Analytic Measurement

Ernesto Calvo, Joan C. Timoneda and  
Tiago Ventura

## INTRODUCTION

In less than a decade, we entered an era of massive datasets reporting inter-connected observations. Social media data provides one of the most visible examples, but it represents just the tip of the data iceberg. Wireless phone networks, live transit data, e-cities and e-government data, citation networks, web page hyperlinks, congressional collaboration events, scientific mentorship dyads, are all part of vast, clean and accessible sources of information currently at our disposal. The flood of data is so extensive that empirically minded scholars have become hoarders of unused datasets stored in their computers.

Big relational data (BRD) represents a challenge to scholars trained in social network analysis, a field that developed most of its techniques to describe small and sparsely featured datasets. As is the case for their smaller counterparts, BRD observations are not independent and identically distributed draws from a population. Paraphrasing

Tobler's first law of geography, everything is related to everything else, but connected things are more related to each other than unconnected ones. As we move from small and shallow relational datasets to large and full featured ones, storing and representation, processing time, parallelization, data dependency, model specification, and model estimation need to be considered together. In this chapter, we discuss theoretical, empirical and practical problems that emerge from the study of large networks, where models for data dependency need to overcome significant conceptual and computational constraints.

As we grapple with the technical issues described above, readers also need to consider what type of information is important to retrieve from their large networks. This chapter introduces readers to a variety of strategies to describe and study big relational data (BRD). This includes feature description, influence and propagation, network dimensionality and reduction, as well as a

description of simple alternatives to estimate inferential statistics that address network dependency in large datasets. The chapter prioritizes practical solutions for inference and a theoretically informed discussion of network dependency, in a field that is rapidly evolving and poised to see significant change and growth in the coming years.

The field of computational social science for social networks is undergoing a massive technical and conceptual revolution, from new algorithms and statistical strategies to specialized software and hardware. These new advances are certainly exciting. However, with constant change comes increased academic uncertainty. As concepts, theories, software and techniques undergo rapid change, academic consensus and accepted standards fail to develop.

Such a rate of innovation makes it difficult to present techniques that will survive the test of time. Every month, new versions of R, Python, igraph, SNA and hundreds of related packages are made available to researchers. At a dizzying pace, dozens of new functions and algorithms are proposed and packaged for academic consumption every month. In this chapter, consequently, we provide a conceptual roadmap for the future using today's technologies. While it is inevitable that the current techniques will be superseded by new advances, the conceptual and empirical problems will likely persist.

Each of the sections in this chapter provides solutions to specific problems that often hamper the study of large networks. We begin with practical advice to create objects that make smart use of limited memory and computational capacity. We then present solutions to improve visualization and description of network topologies when researchers work with anything less than a supercomputer, and provide a few suggestions for if and when supercomputers are an available resource.

We provide practical examples using Twitter data, which has become one of the most prevalent sources of ready-to-go

network data for the study of political communication.

The order of presentation is as follows. First, we present readers with some basic principles that will make data collection and processing more efficient. Memory and processing power are important, scarce resources when dealing with large networks. Therefore, we describe some challenges in data processing and variable selection and discuss some strategies to facilitate network reduction and reproducible datasets. Second, we discuss the efficient manipulation of nodes and edges, as well as functions that describe the topology of large networks. As networks are increasing in size, we will provide alternatives to computationally expensive structural terms such as triangles, stars and geodesic forms. Third, we discuss how to model social influence, network activation and propagation using less computationally demanding algorithms. Finally, we discuss simple solutions to model data dependency in large networks.

## STORING AND DESCRIBING LARGE NETWORKS

### *First Rule: Be Efficient!*

When dealing with networks of hundreds of thousands of nodes and millions of edges, inefficiencies scale at a dizzying pace. Consider a moderately sized network of 100,000 nodes. An undirected affiliation matrix for this network will take a huge amount of space,  $100,000 * 100,000 = 1e + 10$  numbers – that is, 10 billion numbers. An affiliation matrix for a network of 120,000 nodes would add another 5 billion to your memory demands. Therefore, every piece of inefficient code will bring your machine to a grinding halt, and even your department's supercomputer will likely exit with an unceremonious 'insufficient memory allocation' message. Therefore, consider simple functions to run your analyses and store in memory only the variables that you will use.

### **Second Rule: Write Simple Functions to Collect Your Variables of Interest**

One of the most common data formats to store complex relational data is JavaScript Object Notation, known by its acronym JSON. The JSON format allows complex hierarchical information to be presented in a string of characters. Every tweet delivered by Twitter's API, for example, needs to include very different types of information, such as the unique id of the tweet, the unique id of the user, nested information about the user (Twitter bio or self-description, color selected for the wall, number of followers, etc.), and nested information about the tweet (time of the tweet, hyperlinks, etc.). While the JSON format is very efficient in storing the relevant information for every tweet, it also duplicates the variable names and stores dozens of variables that we would never use for our analyses. R packages such as *quanteda*, *twitteR* and *streamR* provide off-the-shelf functions that process a subset of important variables that many researchers use. However, this still includes many variables that we may not need, representing millions of observations that occupy valuable memory space. Further, these functions often store many intermediate objects as they process the data, making it likely that the computer system will run into memory allocation problems much sooner. Reading JSON files and parsing the needed variables using your own function will save significant memory space and time.

### **Third Rule: If Possible, Use Optimized Functions in Stable Package Distributions**

Some packages, such as *igraph* (Csárdi and Nepusz, 2006), have extremely efficient functions to deal with massive networks, subject only to RAM and CPU limitations of the system. But even extremely good packages

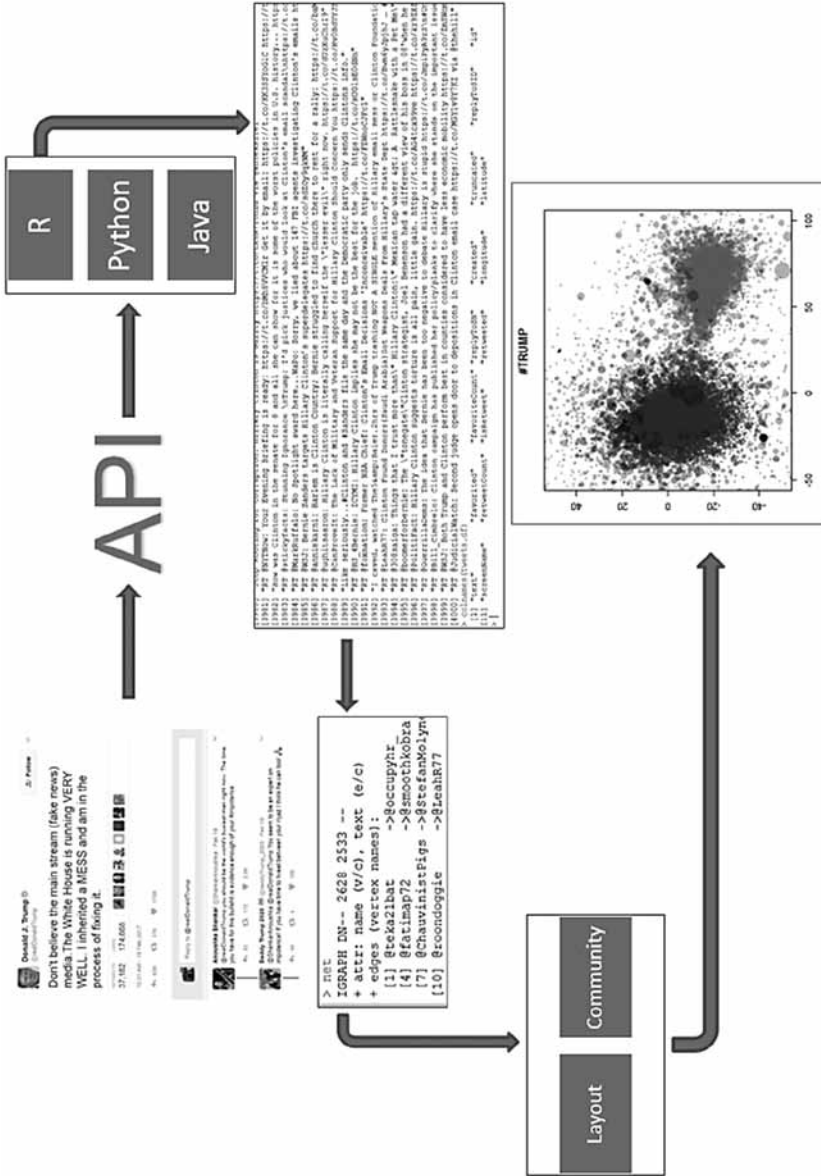
often add wrappers around functions which can increase time and reduce memory. For example, consider `get.adjedgelist()`, one of the most important and useful functions, which retrieves the adjacency matrix for edges in *igraph*. The function that does the heavy lifting is extraordinarily fast and can be easily executed directly using:

```
.Call('R_igraph_get_adjedgelist', net, 1,
      PACKAGE = "igraph")
```

The wrapper in `get.adjedgelist()`, however, adds other functionalities such as the edges' names, with internal calls that slow down the system as networks' size increases. Whenever possible, lose the wrapper and just use the function that executes the required process.

### **Accessing, Selecting and Parsing Data for Big Networks**

We now exemplify a typical data processing exercise using Twitter files. Figure 30.1 describes the usual steps that most researchers follow, reading the JSON file into memory and creating network objects that include a variety of descriptive information, such as the in-degree and out-degree of each node, the nodes' latent locations (layout) and the nodes' membership within subnetworks (community detection). In graph theory, the degree of a node represents the number of edges connected to this node; in the case of directed graphs, where the edges have a direction associated with them, the number of edges coming to the node is called in-degree whereas the number of edges coming out is called out-degree. Twitter networks are obviously directed graphs, since the action of the direction of the connection on Twitter matters. For example, the action user  $v_1$  retweets user  $v_2$  is different from the action user  $v_2$  retweets user  $v_1$ . Nodes with high in-degree are labeled authorities in the network, while nodes with high out-degree are called hubs.



**Figure 30.1 API interface and pre-processing of large Twitter network**

Note: We use one of the dedicated APIs to download JSON files with tweets that include the character 'kavana', on the morning of 9/19/2018. We parse the JSON file to select 25 variables of interest, including the user names ('hub' and 'authority' for each tweet). We estimate a location for each node using the Fruchterman-Reingold layout and identify communities using the walk-trap community detection algorithm in igraph.

Twitter data can be accessed using one of Twitter's dedicated APIs, the forward or 'live' stream and the backward or 'search' stream. The forward API captures tweets in real time on a topic set by the researcher using one or multiple keywords. These tweets are downloaded as they are published by Twitter users. Note that the live stream is throttled once the topic of interest surpasses 1% of total global traffic on Twitter, which can be solved by gaining access to Twitter's Firehose. The search API provides a sample of tweets on a given topic published up to a week before the query. The sample is curated to mix both popular and most recent tweets containing a given keyword. Access to tweets older than one week comes at a cost.<sup>1</sup>

Packages have been developed in languages such as R and Python to access Twitter's APIs. TwittR and streamR in R as well as twarc in Python are prominent examples. All of these solutions store information in a JSON (JavaScript Object Notation) file, which can then be parsed to extract the information required by the researcher.

### ***Parsing JSON Files***

The JSON format organizes information in strings of characters that are easy to parse. They are particularly useful for objects with dynamic and open-ended data structures such as large nested lists and arrays. Consider that one tweet produces over 200 different variables with information about the tweet and the user who wrote it. Functions such as `searchTwitter()`, from the package `TwittR()`, provide only 16 of the most frequently used features in Twitter data, while the function `filterStream()` from the package `streamR()` returns 42 variables. Both programs provide simple functions to transform the downloaded JSON file into a ready to use dataset in R. In Python, the package `twarc` downloads a complete JSON file which can then be processed in R, Java or Python. Downloading and parsing the full JSON file

is always better if a different set of variables is needed for our research or if memory constraints make the use of off-the-shelf functions unfeasible. Code to parse a JSON file is available from the authors' websites.

To improve memory management, it is convenient to create functions that, after uploading JSON files into the system, select a few important variables and collect them into a dataframe. Rather than iterating with a *for* loop, some of the base functions such as `apply`, `lapply`, `sapply`, as well as the map functions from the `purrr` R package will significantly improve computation time as the size of the dataset increases.

### ***Creating a Large Network***

Once the JSON file has been uploaded and variables have been selected into a dataframe, we can create our network using relational aspects of the data such as retweets or replies, which connect user *j* and user *i*. Other variables may also be loaded to the nodes – characteristics intrinsic to each user – and to the edges – characteristics of the relationship between users.

### ***Fourth Rule: When Creating Your Network, Add Your Important Variables at Once!***

As we will show, networks need to be trimmed, pruned, sliced and transformed in multiple ways. You do not want to merge new variables later in the process once a large number of edges and nodes have been selected out and it has become hard or impossible to apply the same selection rules to new variables. In R, packages such as `igraph` and `Network` create efficient objects that can store variables of length (nodes) or length (edges). When a large network object is created, therefore, add all variables of interest to the nodes:

```
V(net)$friends = kavana.df$friends # (igraph)
net %v% friends
  = kavana.df$friends # (network)
```

as well as to the edges:

```
E(net)$text = kavana.df$text #(igraph)
net %e% text = kavana.df$text #(network)
```

A common way to build a network from Twitter data is by analyzing the retweeting behavior of users. In doing so, we act on the assumption that a retweet is a sign of affinity, where users that retweet some users rather than others are closer to the former and more distant from the latter. We may use this information to identify, for instance, more important or influential users, that is, users that are more frequently retweeted (e.g. users with higher in-degree) or users with a larger number of ‘followers’. Later in the chapter, we provide some modeling strategies to understand retweeting behavior of users.

## THINNING LARGE NETWORKS

### *Fifth Rule: The Whole Data is Not Always Better!*

There is little doubt that networks can be visually stunning. Figure 30.3, for example, uses the Fruchterman–Reingold layout and the walk.trap community detection algorithm to describe the relationships between 8,835 retweets about Donald Trump by 7,911 users on the morning of October 8, 2018. The original network (upper plot, Figure 30.3) includes a very significant number of smaller subnetworks in its periphery. As readers may notice, these smaller sets of users are not connected to the primary cluster and, consequently, cannot be placed on the same latent space; nor are they connected to a particular community of the main cluster. Visually, it

will be tempting for readers to perceive these unconnected nodes as if their location in the graph and community membership was meaningful. However, there are no paths that connect nodes in the main cluster to any of those subnetworks.

As networks are loaded, researchers will often have to make decisions regarding the network objects that deserve their attention. This could require working with the primary connected cluster (middle plot), which considers only the largest cluster of connected users (e.g. users with paths to all other users). Oftentimes, filtering for the primary connected cluster will also eliminate isolates that are not truly ‘talking’ about the same events.

Another important question is whether single occurrences by a user (a single connecting edge from one user to one other user in the network) adds relevant visual information. The middle plot in Figure 30.3, which only includes the main connected cluster, still includes a large number of users that participated only once in the network, extending a single link to one other node. Their contribution is already captured by the in-degree of the receiving node, visualized by larger circles. Therefore, it is often a good idea to filter single occurrence nodes and retain in the network object only those users that either have an out-degree larger than 1,  $degree(net_{ij}, mode="out") > 1$ , or an in-degree equal or larger than 1,  $degree(net_{ij}, mode="in") > 1$ . The third and final plot in Figure 30.3 describes a filtered network that eliminates unconnected users as well as users of  $out - degree == 1$ , who would be given an arbitrary location in the network.

## NETWORK TOPOLOGY

Different from spatial models in geography, which locate points in a sphere according to projections that respect the original

```

twList <- function(tw=tw) {
  test <- tryCatch(fromJSON(tw), error=function(e) NULL)
  out <- c(test$id_str, test$text, test$user$screen_name,
  ifelse(is.null(test$user$verified)==TRUE, "", test$user$verified),
  out)}
  library("rjson")
  my.file <- "trump.json"
  my.tweets <- readLines(my.file)
  att <- sapply(my.tweets, function(x) twList(x), simplify=TRUE,
  USE.NAMES = FALSE)
  trump.df <- t(att)

```

**Figure 30.2** Sample code to create a dataframe with selected variables

distances between points, network visualizations are not constrained by a ‘true underlying grid’. Data reduction techniques that locate nodes and edges on the space, *layouts*, are selected by researchers because they summarize and communicate information that is deemed relevant.

### ***Sixth Rule: Know What Information You Wish to Transmit to Your Readers!***

Consider the same dataset used in Figure 30.3, but subject to a variety of algorithms that recover locations for all nodes and edges. Each algorithm represents a compromise between pure data reduction techniques (e.g. multidimensional scaling layout in the left-middle row of Figure 30.4) and pure visual discrimination (e.g. the Random layout in the right-middle row of Figure 30.4).

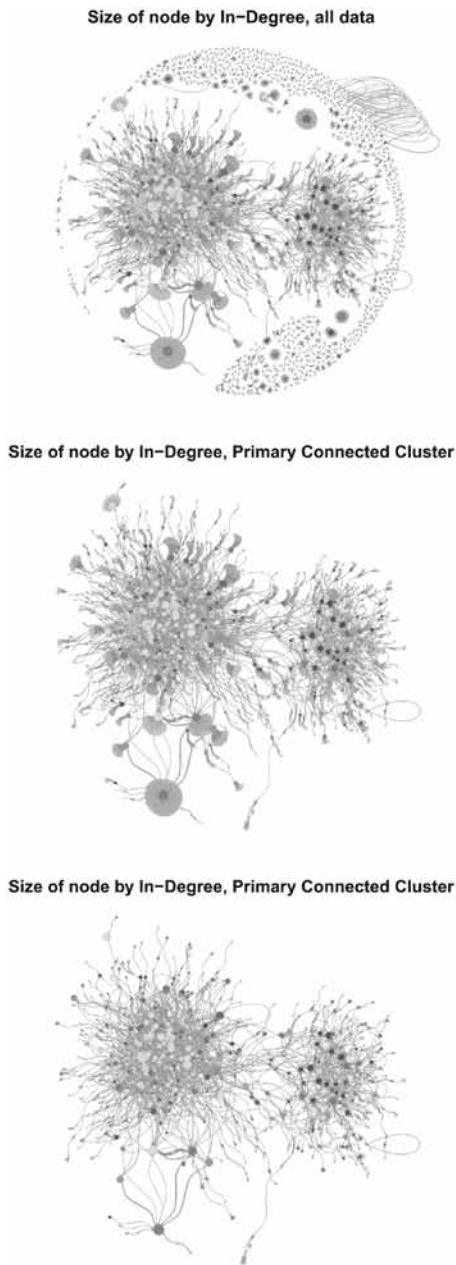
Force-directed algorithms, such as the Fruchterman–Reingold plot in the upper-left corner of Figure 30.4, are a compromise between the pure data reduction retrieved from the adjacency matrix and visual

discrimination among nodes. In the case of the Fruchterman–Reingold plot, discrimination between nodes in close proximity is carried out through *springs* that separate nodes over short distances. This short distance repulsive force is absent in multidimensional scaling. Therefore, better visualization by the Fruchterman–Reingold algorithm will also add information that is not in the adjacency matrix.

Researchers should select algorithms for network description (e.g. both visualization and data reduction) which provide readers with the appropriate information. Researchers need to report both the advantages of the selected layout and the data implications. The selection of a layout, therefore, should make explicit the type of information that researchers wish to communicate to readers. Let us describe some of the most common layouts used for network visualization.

### ***Fruchterman–Reingold***

The Fruchterman–Reingold layout is a force-directed graph drawing algorithm that



**Figure 30.3 Twitter 'Trump' network (1)**

*Note:* The top plot presents all nodes and edges in the data, with node size according to the in-degree of each node. The middle plot retains the primary connected cluster, with unconnected sub-clusters eliminated. The lower plot retains the primary connected cluster and eliminates nodes with out-degree = 1. The information of the eliminated nodes is preserved in the size of the node (e.g. the in-degree in each node)

groups together nodes that share more connections with each other. The length of the edges is minimized, and nodes repulse each other as if electrically charged when they come into contact. When in equilibrium, the process creates a map of the network in which communities with strong ties are easily detected, as they are separated from each other by force. By altering the size of the nodes in accordance with some measure of importance within the subnetwork, we can also broadly assess the hierarchy of the network and the presence of authorities. As a disadvantage, the Fruchterman–Reingold algorithm is particularly slow in networks with more than 1,000 nodes. Precisely because of the force that underpins the graphing technique, nodes with weak connections to other actors in the network will be pushed out toward the periphery (Fruchterman and Reingold, 1991; Kobourov, 2013).

### **Davidson–Harel**

Davidson and Harel's algorithm is based on simulated annealing, a technique used to cool liquid that leaves a crystallized structure by reducing energy slowly. The top-right plot in Figure 30.2 shows precisely this: it's as if the nodes and edges of the network had been slowly set in one non-random but cohesive structure. Notice that communities can be detected by the color of the nodes, and it is therefore not random. Contrast this with the plot immediately below the Davidson–Harel graph, which is indeed random. Both nodes and edges are distributed evenly in the random plot, while nodes and edges are concentrated in communities in Davidson–Harel. It minimizes edge crossings and ensures vertices are not too close to non-adjacent edges. These communities, however, are much less clear to the naked eye than they are, for instance, in the Fruchterman–Reingold layout. As a note of caution, Davidson and Harel themselves admitted that



the layout algorithm underperforms in networks with more than 60 nodes (Gibson et al., 2013).

### ***Multidimensional Scaling***

Multidimensional scaling is a visual representation of networks in which nodes that are more similar to each other are placed closer together. Distances among objects are estimated using a data reduction technique and placed on a map, which can have two or more dimensions. The final orientation of the axes is arbitrary, prioritizing the accurate estimation of relative distance. As opposed to factor analysis, multidimensional scaling is a non-linear dimension reduction technique. As with Fruchterman–Reingold, multidimensional scaling identifies communities well, as it groups together nodes that are similar to each other and separates them from those that are not. But since both procedures use entirely different algorithms to place objects in space, the way they communicate information varies slightly (Gibson et al., 2013; Buja et al., 2008).

### ***Kamada–Kawai***

The Kamada–Kawai model produces a force-directed layout based on a logic of stress minimization. As opposed to Fruchterman–Reingold, it does not ensure that nodes are not too close to each other. Rather, it seeks to reach an equilibrium in which the geometric differences between points closely resemble those in the graph. The Kamada–Kawai algorithm is computationally expensive and may struggle to detect communities as precisely as Fruchterman–Reingold. This is because, in the latter, nodes repel each other such that vertices that are far away from each other in theory are further apart in the graph. Kamada–Kawai, on the other hand, represents a more direct mapping of the

theoretical distance between nodes in a graph layout (Kobourov, 2013).

### ***Large Graph Layout (LGL)***

The large graph layout (LGL) was developed to visualize large networks of hundreds of thousands of nodes and more than a million edges, which is particularly difficult to do with the layouts described above. A force-directed algorithm, the LGL begins by identifying different clusters within the network, that is, it separates the overall network into highly connected clusters. It then lays out each connected set separately in space before bringing the entire network together into the same coordinate system. For smaller networks such as the one used as an example here, the LGL does not appear to make a difference. Indeed, the bottom-right plot in Figure 30.2 provides a less clear visual representation of the two different communities in the network than do the Kamada–Kawai, multidimensional scaling and Fruchterman–Reingold algorithms. However, in a large network, the LGL would provide a more meaningful layout of the network (Adai et al., 2004). Table 30.1 summarizes the different layouts we introduced here.

## **COMMUNITY DETECTION**

Beyond estimating the location of users in a network (layout), we may be interested in identifying groups of nodes that interact more frequently with each other (e.g. regions of the network with nodes that are more densely connected). In Figure 30.4, the  $[x,y]$  coordinates of the nodes (different layouts) summarize information about the latent location of nodes, but such proximity provides little information about the probability that these nodes will be more likely to interact with each other than with the rest of the nodes in the network.

**Table 30.1 Summary of layouts for network visualization**

<i>Layouts</i>	<i>Summary</i>	<i>Keywords</i>
<i>Fruchterman–Reingold</i>	A graph drawing algorithm that separates communities by force. It ensures that nodes are never too close to each other	force-directed, spring forces, minimum separation
<i>Davidson–Harel</i>	Network energy is reduced slowly and minimizes edge crossings. Vertices not too close to non adjacent edges	simulated annealing, slow energy reduction
<i>Multidimensional scaling</i>	Nodes similar to each other are placed closer together. Nonlinear dimension reduction technique. Works with multiple dimensions	multiple dimensions, data reduction
<i>Kamada–Kawai</i>	Force-directed layout based on stress minimization. It represents a more direct mapping of the theoretical distance between nodes in a graph layout	stress minimization, theoretical distance
<i>Large graph layout (LGL)</i>	Used for large networks. It breaks down clusters first, then connects them back into one single coordinate space	large networks, force-directed

### ***Girvan–Newman***

One way to detect communities is by using the Girvan–Newman algorithm. It progressively removes edges from the network starting with those that most likely link different communities. It first calculates the ‘edge betweenness’ of every edge in the network, assessing the number of shortest paths of which it is part. The higher the edge betweenness, the more often a given edge is likely to connect various communities. Clearly separate communities emerge after these edges are removed. However, betweenness is recalculated after removing the edge with the highest score, which makes the procedure intensive and often error prone (Girvan and Newman, 2002).

### ***Fast and Greedy***

The fast and greedy algorithm provides a fast calculation of modularity in larger networks,

using a bottom-up approach. Each vertex is initially independent, and the algorithm progressively merges nodes in such a way that it produces the largest locally optimal increase in modularity in each step. These communities are then aggregated and the hierarchy of the network is revealed from the bottom-up. This is a fast method partly because there are no parameters to estimate, but it requires a certain sample size to work properly (Clauset et al., 2004).

### ***Louvaine***

A similar method to the fast and greedy algorithm is the Louvaine method for community detection, although it is computationally more intensive and slower. As opposed to the Girvan–Newman algorithm, the Louvaine method is a bottom-up approach that first estimates the smallest possible subnetworks by optimizing modularity. After each step,

communities are collapsed into one node, and again the process is repeated. All nodes are allocated in this way until no increases in modularity are possible. Once this process is over, the algorithm builds the network again, with clearly separate communities (Blondel et al., 2008).

### **Random Walk**

The logic of the random walk approach to community detection is that most random walks will stay within the community, as there are few edges that lead to other communities. You can select the number of steps for the random walk and the algorithm will then generate communities using a similar bottom-up approach as in fast and greedy. What random walks lose in speed, they usually gain in the accuracy of the results (Pons and Latapy, 2006; Yang et al., 2016). In Table 30.2, we summarize this discussion about distinct

algorithms for community detections in network analysis.

Once nodes are given a location in the space (layouts) and a group membership (community detection), it is often unnecessary to visualize the edges. This is particularly true once networks grow in size and plotting the edge information would just add visual noise. As networks increase in size, hundreds of thousands of edges will become just a color background, with little information value.

### **Seventh Rule: As Networks Increase in Size, Lose the Edges!**

Of course, we are not suggesting that the information of the edges should be eliminated – far from it. It is just not necessary to plot those edges when the relevant information can be described by adjusting the properties of the nodes. We will exemplify this suggestion in the next section, as we plot

**Table 30.2 Summary of the algorithms for community detection in the network**

<i>Method</i>	<i>Summary</i>	<i>Keywords</i>
<i>Girvan–Newman</i>	A top-down approach that removes those edges that have the highest levels of edge-betweenness first. Computationally more intensive as betweenness is calculated after every step	top-down, edge betweenness
<i>Fast and Greedy</i>	A bottom-up approach to identify communities quickly. Modularity is optimized locally at each step, generating communities. These are then aggregated into a hierarchical network	bottom-up, local optimization
<i>Louvaine</i>	Similar to fast and greedy, slightly slower. Communities are collapsed into nodes after each estimation until no increases in modularity can be achieved	modularity optimization, bottom-up
<i>Random Walk</i>	Most random walks will stay within the community, and fewer will lead to other communities. The algorithm uses this insight to build communities from the bottom up	bottom-up

different hashtags that *activate* content of the Kavanaugh network of tweets in September of 2018.

## INFLUENCE AND PROPAGATION

Large networks differ from small networks not only because there are a very large number of nodes but also because these nodes are arranged in network topologies that are locally distinct. For example, consider the small but complex network in Figure 30.5, which has three different but connected local topologies. There are three edges that connect the different regions, allowing information to flow from one part of the network to the others. Those three edges have high edge betweenness, which bridges the content that circulates between each set of communities.

Influence and propagation, the capacity to elicit responses from nearby nodes and the capacity to activate content in a network

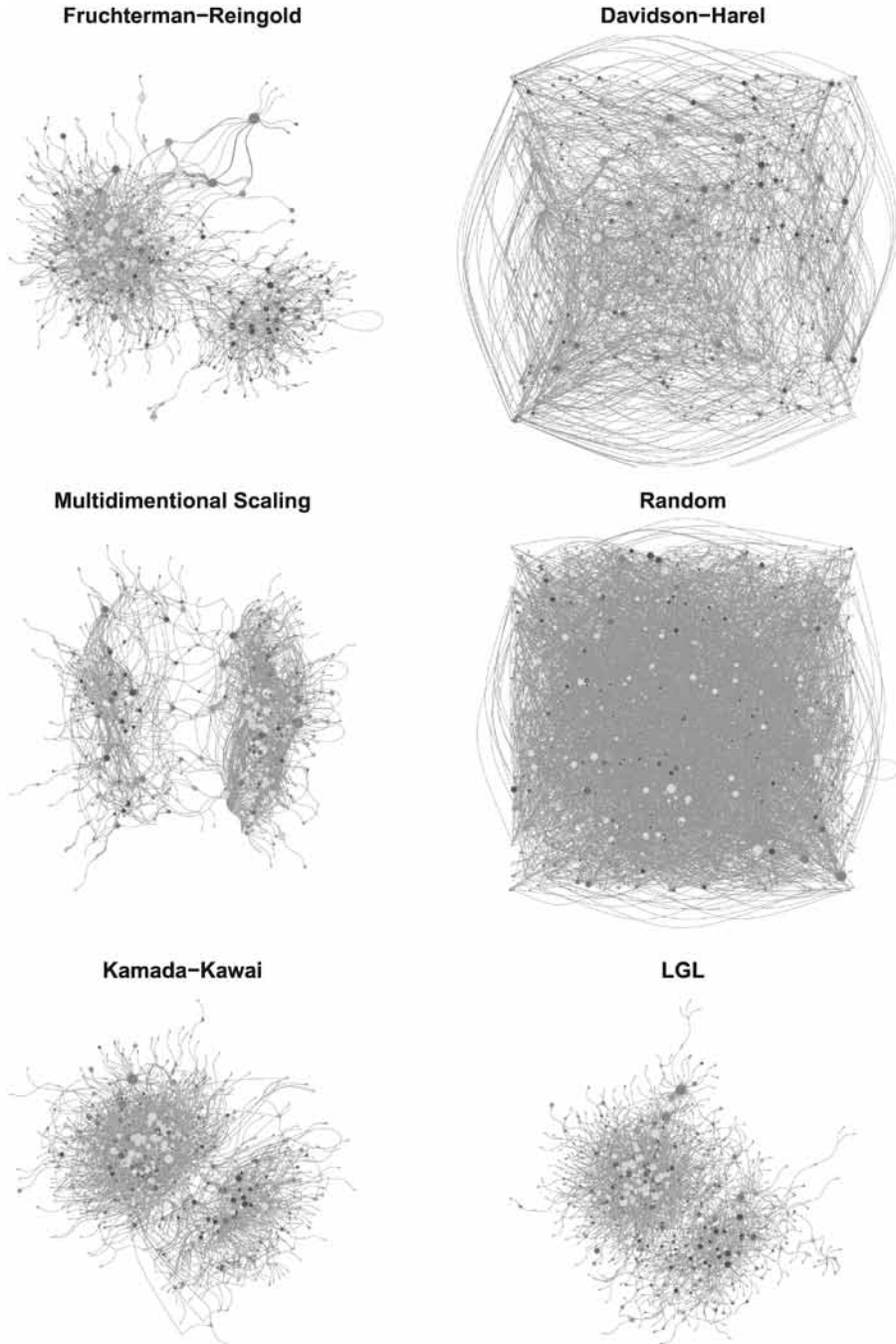
have become an important area of research as we move from small and sparsely featured networks to large and complex ones. Indeed, Table 30.3 describes some of the techniques available to assess the influence of nodes in such social networks.

Figure 30.5 describes the flow of information over multiple rounds in this network, allowing in-degree information to propagate to contiguous nodes until node information is equalized (entropy). In each round, nodes read the average in-degree of the contiguous nodes, allowing for information to converge. After 25 rounds, the in-degree information has equalized and all nodes converge to the average in-degree scores of the network.

However, as shown in Figure 30.5, information propagates rapidly in the fully connected community (top) and slowly in the ring community (lower right). In complex networks, information propagates at different rates in regions of the network, depending on how groups of nodes are connected to each other. Therefore, influence (e.g. the capacity to affect the behavior of contiguous nodes)

**Table 30.3 Influence and propagation**

Source	Definition	Keywords
<i>Rashotte</i>	a phenomenon caused by interaction among individuals that changes their thoughts, sentiments, and behavior	changes in behavior, interaction among actors
<i>Watts and Dodds</i>	influence defined as individual/s who appear in a given percentile $q$ of the influence distribution $p(n)$	top $q\%$ , Poisson distribution
<i>Cha et al.</i>	influence based on three types of actions: following, retweets and dialog	actions
<i>Bakshy et al.</i>	the influence of a given post is calculated first; then a model is fitted that predicts influence incorporating individual attributes and past cascades	reposting, cascading
<i>Tang et al.</i>	the extent to which individuals change their behavior based on their relationships with other people, organizations and society	changes in behavior, perceived relationships



**Figure 30.4 Twitter 'Trump' network (2)**

*Note:* The six different layouts are created applying different algorithms to the network. Different from Euclidean geography in transportation networks, there is no 'right' layout. The different layouts represent compromises between visualization and proximity. Data reduction techniques, such as multidimensional scaling, respect latent distance in the data at the expense of visualization. Force-directed algorithms respect relative proximity and improve visualization by adding short-distance distortion.

and propagation (e.g. the capacity to activate content that reaches contiguous nodes) varies locally in complex networks. As networks increase in size, summary information such as the density, the average degree, the centrality or the overall betweenness scores become less informative.

### ***Eighth Rule: Go Local! Understand Activation Events in Your Network***

Consider the activation of different hashtags in the Kavanaugh network in Figure 30.6. For these plots, we still use the Fruchterman–Reingold algorithm that locates users in the primary connected cluster as well as the random walk algorithm for community detection. Colors for the communities were edited, with blue for those that oppose the nomination of Kavanaugh and red for those that support nomination. Inspection of the largest authorities in each community (e.g. users with the highest in-degree) show that political figures from the Democratic party are prevalent in the opposition community and that President Trump and Republican representatives are prevalent among those that seek confirmation.

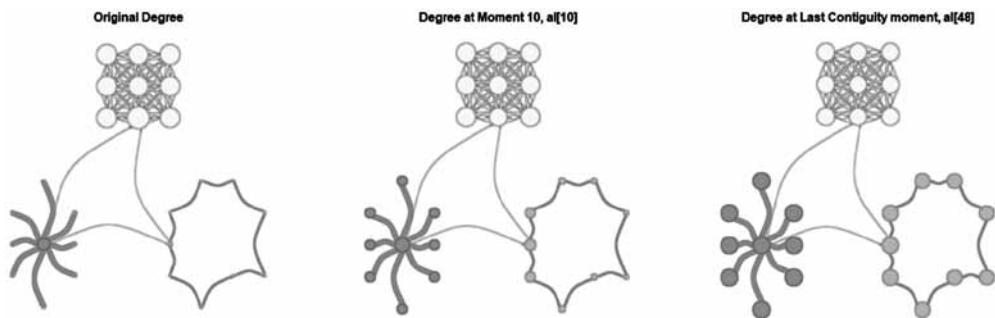
As shown in Figure 30.6, very different hashtags were activated by users in each

community. To explore the activation of different contents in the Kavanaugh network, we collect all the hashtags embedded in tweets of the Kavanaugh network, with tweets searched by the string of characters ‘kavanaugh’ on the morning of September 19, 2018. The upper-left plot describes the full distribution of nodes, with users opposing confirmation in blue and those supporting confirmation in red. Then, we use the edge-contiguity matrix to create a node attribute that counts the number of edges for each node that include our selection of hashtags.

Rather than coloring millions of edges, we modify the size of the nodes to reflect the log-counts of the hashtags, plotting thousands of nodes rather than millions of edges while preserving the information.

The information depicted in the different figures showed that supporters of Kavanaugh’s confirmation were more likely to use #MeToo hashtags than those opposing nomination. Instead of shying away from the issue, supporters of Kavanaugh used hashtags such as #MeTooHucksters to argue that #MeToo accusations were just an attempt to derail confirmation.

As shown in Figure 30.6, modeling how content is activated in different regions of the network is considerably more informative



**Figure 30.5** A complex network with three local topologies: (a) a ring subnetwork, (b) a fully connect subnetwork and (c) a follow the leader subnetwork. Information propagates at different rates within and across networks. In fully connected networks, information moves effortless to all nodes. Meanwhile, information takes several rounds to reach removed nodes in the ring subnetwork

than reporting just the frequency of content on the average node.

## STATISTICAL DEPENDENCY

Activation and propagation are two different, although interrelated, phenomena. In the case of activation, we model how different regions of the network communicate different information. In the case of propagation, however, we hope to understand how nodes influence each other locally – that is, to model the extent to which the probability of observing a local event is affected by contiguous nodes.

When working with big relational data, tools to model network dependency may be computationally untractable (Schmid and Desmarais, 2017). In their recent article, Schmid and Desmarais note that exponential random graph models (ERGMs) may be computationally prohibitive once they reach 1,000 nodes, a modest network size for today's standards. Instead, they argue, for networks of those sizes, maximum pseudolikelihood estimation (MPLE) may be an alternative. However, even MPLE, or simpler structures such as triangles, stars or geodesic forms in an entire network, will be outside MPLE reach when we deal with the smaller 'Trump' network described earlier.

In this final section, we discuss some simple solutions to model data spatial dependency in large networks. We begin with the *Path Weighted Local Regression Model* (hereafter *PWR*). This model extends the *geographically weighted regression (GWR) model* (Fotheringham et al., 2002), allowing researchers to model spatial dependency and heterogeneous effects of big relational networks using a less demanding computational strategy. We then briefly discuss an autoregressive network model that uses the first order contiguous nodes to control for network dependency. Both strategies are computationally cheap, can be easily parallelized

and will see modest increases in computation time as the size of the network grows large.

### *The Path Weighted Local Regression*

PWR is an alternative to model spatial dependency and non-stationary, heterogeneous effects of covariates across large connected networks. Inspired by the geographically weighted regression model of Fotheringham et al. (2002), Lloyd (2010) and Darmofal (2008), the PWR strategy describes the effect of unobserved factors across closely connected nodes.<sup>2</sup> The model takes as input the distance weights matrix of a network and gives more weight to nodes that have shorter paths between them. As a network variation on LOESS (Jacoby, 2000; Keele, 2008), the model is estimated locally, with  $i$  values for each local  $\beta$ , with weighted paths to all nodes in a network.

Consider a linear model on the data of a fully connected network where the dependent (node) variable  $y_i$  is explained by a set of observed covariates  $x_N$  and unobserved parameters  $\beta_N$ :

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_n x_{in} + \varepsilon_i \quad (1)$$

Using ordinary least squares (OLS) we minimize the sum of square residuals and solve for the  $\hat{\beta}$  parameters, so that:

$$MSE(\beta) = \frac{1}{N} \sum_{i=1}^N (y_i - x_i \beta)^2 \quad (2)$$

$$\hat{\beta} = (X'X)^{-1} X'y \quad (3)$$

For any given node, however, we may consider that observations that are more closely connected (lower path distances) will have a larger effect on the dependent variable than observations that are further away. Therefore, for each node in the network we may estimate a separate weighted linear regression model:

$$wMSE(\beta, w_1, \dots, w_N) = \frac{1}{N} \sum_{i=1}^N w_i (y_i - x_i \beta)^2 \quad (4)$$

And solve for the local parameter:

$$\hat{\beta} = (X'WX)^{-1} X'Wy \quad (5)$$

Similar to the standard geographically weighted regression, we are estimating separate models for each node in the network, with observations that are further removed from a node weighing less than those observations that are closer. That is, all PWR estimates are local and the output of the model returns coefficients for each covariate at each node  $i$  in location  $[x, y]$ .

Model results provide a full distribution of  $\hat{\beta}$  parameters for all nodes in the network, which can then be post-processed and visualized. Heterogeneous effects of the independent variables by node location, consequently, allow researchers to understand differences in content propagation at different regions of the network.

In contrast with GWR in spatial models, path distances in networks are shorter, with a relatively small set of discrete values that connect all nodes. Therefore, bandwidths across different networks will have greater similarity compared to bandwidths in GWR. For our implementation of the PWR, we consider the minimum number of paths connecting each pair of nodes, creating a distance weights matrix. In dense areas of a network, therefore, there will be a larger sample of minimum paths connecting all nodes. By contrast, in sparsely connected regions of a network the opposite holds.

More important, shorter paths in densely packed communities will display lower variation across nodes as information will travel fast, while longer distances across communities will result in higher parameter discrimination. That is, more dense networks will reduce local effects, as the distance across all pairs of nodes will be smaller. Meanwhile, local effects will be more distinct in sparse networks, where path distances across nodes are on average larger.

Similar to GWR, building the weights matrix requires that researchers decide the relative contribution of nodes through a decay function. Several alternatives are available, with the Gaussian weighting function being the most common choice in existing literature (Fotheringham et al., 2002; Darmofal, 2015). The Gaussian discount function takes the form:

$$w_{ij} = \exp \left( \frac{-I}{2 \frac{P_{ij}^2}{b}} \right)$$

In the previous equation,  $P_{ij}$  describes the minimum number of paths connecting node  $i$  to node  $j$ , and  $b$  is the bandwidth for the path-decay of the weighting function. A larger bandwidth results in estimates that are more distinctively local. By contrast, a smaller bandwidth results in estimates that are roughly similar across nodes. As in other local polynomial models, there is a trade-off between bias and variance in the choice of the PWR bandwidth. To approximate an optimal bandwidth, one strategy is the leave-one-out cross-validation selection (Fotheringham et al., 2002):

$$CV = \sum_{i=1}^N (y_i - \hat{y}_{i \neq i})^2$$

The cross-validation procedure uses a leave-one-out process in which each local estimation of the observation  $i$ , where the local parameter is centered, receives weight equal to zero. Then, the score takes the square difference of the  $y_i$  and the prediction of the model for the observation where the weight was set to zero. We provide an analytical solution to find the point where the optimal value of the score CV is minimized. Working on big relational data, one can approximate the value by taking a sample of the nodes. The same procedure is suggested in applications of geographical weighted regression (Fotheringham et al., 2002).



### ***A Note on Parallelization***

One of the advantages of the path weighted regression model is that it can be easily parallelized. The technique is running separate regressions for each node, weighing more heavily those vertices that are closer to each other. The model can independently fit as many linear regressions as there are nodes in the network. Since calculations are computed using the same bandwidth on a steady weights matrix, regressions can be run in parallel rather than sequentially, which translates into lower computation times and more efficient use of computing power. The extent of these savings will depend on the number of cores available and the clock speed of each core, virtual or real. In a supercomputer cluster, one can expect much greater gains in efficiency and speed. This is particularly important in large networks, which are computationally demanding.

As important, the weight matrix in a PWR model does not increase exponentially, as the addition of new nodes does not translate into a more than proportional increase in the number of weights. As a result, computation does not become exponentially more difficult when a network size increases. Because more nodes only increase estimation demands linearly, larger networks can be modeled without sacrificing time or computational resources. Thus, combined with parallelization, PWR is a fast and efficient modeling strategy for large networks.

### **APPLICATION: TWITTER NETWORK ANALYSIS OF THE KAVANAUGH CASE**

To provide an example of the PWR, we analyze the Twitter repercussions of Donald Trump's nomination of Brett Kavanaugh to the US Supreme Court. After his nomination to the Senate in September of 2018, allegations of sexual harassment by the nominee were disclosed in the press and in Senate

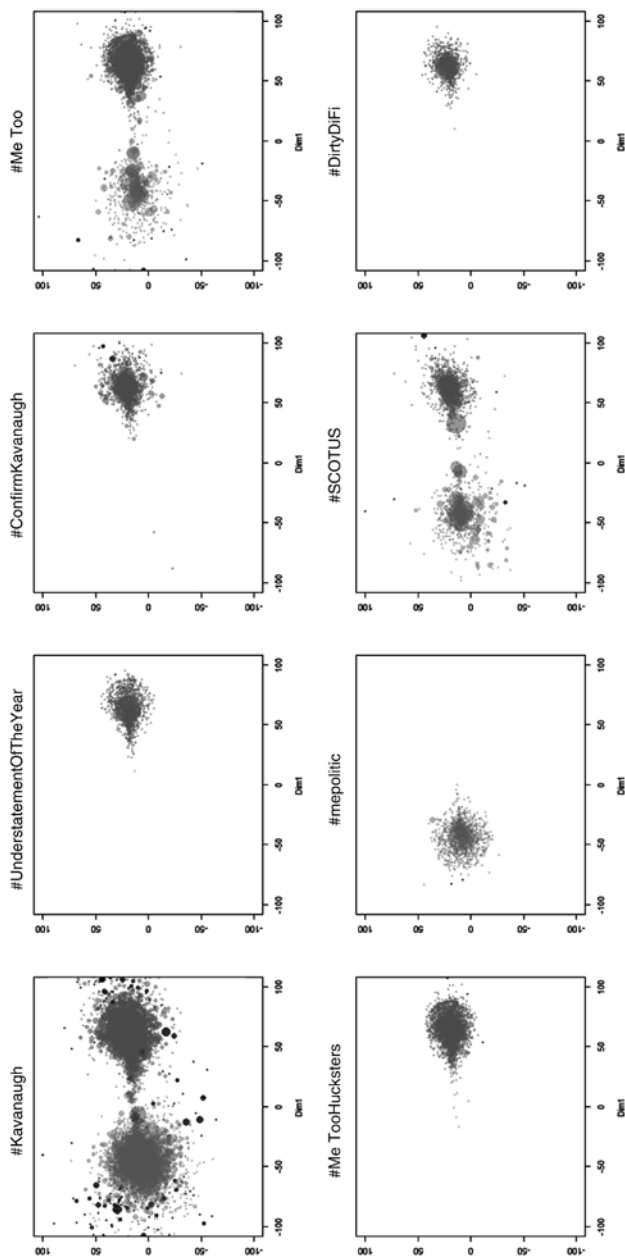
hearings. While Kavanaugh denied all charges, discussions in social media follow deeply polarized fault lines.

Between 8am and 12pm on September 19, 2018, we tapped into Twitter's stream API to collect publications that included the string of characters '*Kavanaugh*'. To provide a visual representation of the data, we estimated a Fruchterman–Reingold layout on the Kavanaugh network, retrieving the [x,y] coordinates for each user described in Figure 30.7. We also ran the walk.trap algorithm in *igraph 1.0* and retrieved an index community value for each node (Csárdi and Nepusz, 2006). The walk.trap algorithm (Pons and Latapy, 2006) sorted nodes in two large communities, which we identified by their highest in-degree users as aligned with opposition to Kavanaugh's nomination (39,631 users) or support of the nomination (54,174 users). The remaining 1,268 accounts were sorted into 532 smaller communities weakly connected to the two larger communities. Figure 30.7 provides a general visualization of the #Kavanaugh network.

We then ran our PWR model on the #Kavanaugh network, regressing each node's *average time to retweet* in seconds on the nodes in-degree. The model seeks to explain the speed of retweets of users that have a larger group of retweets (high in-degree), to model the propagation of information in different regions of a network. The PWR model allows us to explain heterogeneous effects of in-degree on content propagation at the center or periphery of each community, with time to retweet as a valuable measure of content propagation (Aruguete and Calvo, 2018).

### ***Estimation***

We first retrieved an optimal bandwidth for the weight matrix of path distances. To this end, we selected 250 random samples of 100 nodes, using the mean of these distributions to estimate the full cross-validation model. Figure 30.8 plots the sampling distribution of the bandwidth, which returns an optimal



**Figure 30.6** Activation of eight different hashtags during the Kavanaugh Confirmation Hearings, September 19, 2018

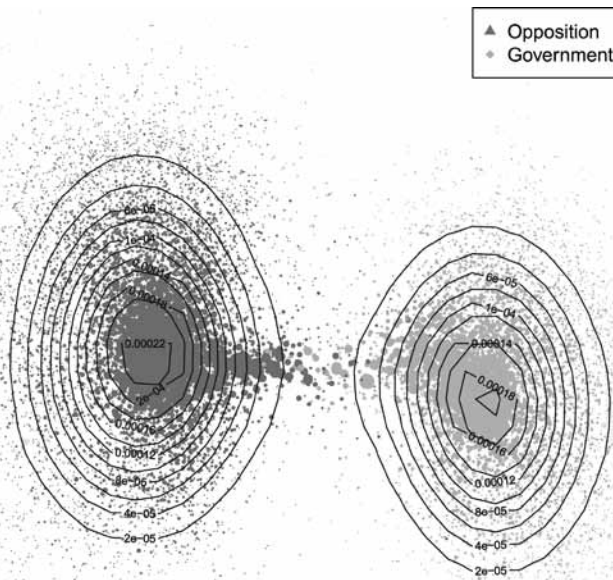
*Note:* The network has 95,073 nodes and 771,665 edges. Therefore, plotting all edges of the Kavanaugh network would be computationally expensive and produce very heavy graphs. However, we can simply adjust the size of the nodes to be proportional to the number of edges that include each hashtag. Therefore, all the relevant information can be presented to readers without plotting hundreds of thousands of edges.

bandwidth of 1.217 to be used on the complete dataset. It is worth noticing that smaller bandwidths will increase local effects while larger bandwidths will provide local estimates that are closer to the overall network mean. Finally, we parallelize estimation as suggested in the previous section.

Figure 30.9 presents the distribution of local  $\beta$  coefficients for the PWR model. We color the nodes by quartile to facilitate visualization. The size of the nodes describes the users' in-degree. The upper plot presents local intercepts while the lower plot presents the slopes of the model for each node. Interpretation of the results is straightforward, with more negative slope coefficients describing slower time to retweet as the in-degree goes up. The fact that authorities (i.e. users with higher in-degree) would

take a longer time to retweet information is expected for two distinct reasons. First, high in-degree authorities will be more risk averse regarding sharing content that may reflect poorly on them. Second, bots, trolls and other managed response systems tend to have a smaller numbers of followers and are set to quickly share content from priority accounts.

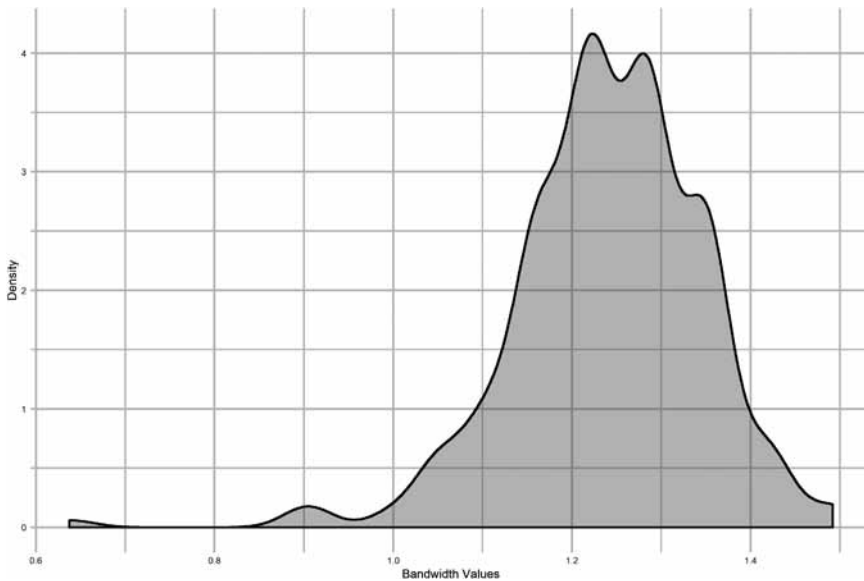
The intercept of the model described in the upper plot of Figure 30.9 shows that the average time to respond by users that are located at the center of each community is slower than the average time of the second periphery. More interesting, the slope estimates displayed in the lower plot show that the effect of in-degree on time to retweet is larger at the center of each community. That is, as in-degree increases, the decline in time to retweet is considerably larger for core nodes



**Figure 30.7 Primary connected network of #Kavanaugh**

*Note:* Blue dots describe users opposing the confirmation of Kavanaugh. Red dots describe users that support confirmation. Layout of users estimated using the Fruchterman-Reingold algorithm in *igraph*.

*Source:* Community detection using walk.trap algorithm in *igraph* (Csárdi and Nepusz, 2006).



**Figure 30.8** Sampling distribution of the bandwidth selection procedure (250 samples with 100 observations each)

in the pro- and anti-Kavanaugh networks than in their peripheries. By contrast, nodes in the periphery of each community are quick to propagate content even when their in-degree increases. The lack of effect of in-degree on content propagation on the second periphery of both communities lends considerable support for the presence of bots and computer managed accounts.

The heterogeneous effects of in-degree on content propagation are large and substantively interesting. As in-degree increases, the model shows, as expected, that nodes will take longer to share messages from other users. Users at the center of each community have a stronger decay function, while, for loosely connected nodes on the periphery, being an authority matters less for quickly propagating news.

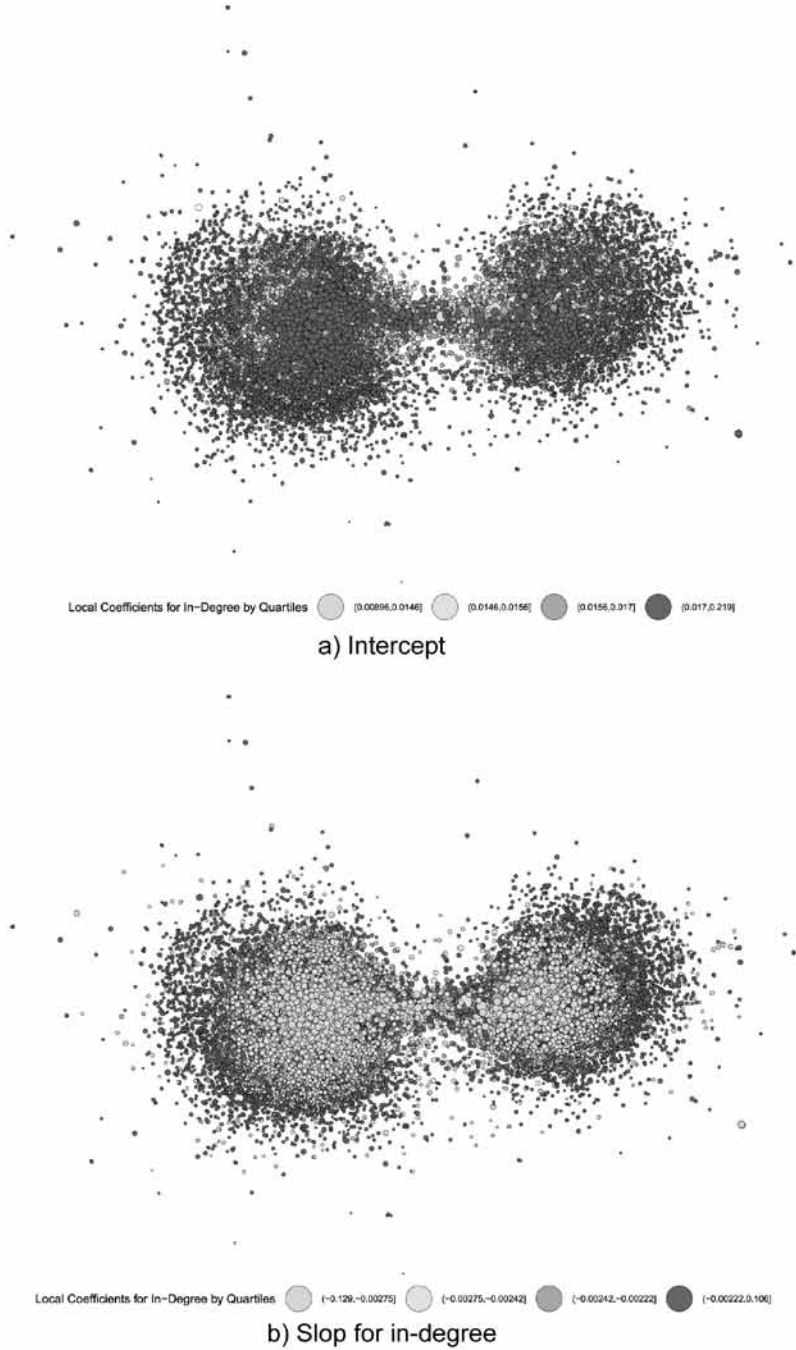
One caveat is important here. As one can see by the size of the nodes in Figure 30.9, authorities tend to be at the center of each community. Therefore, the average in-degree of the nodes is higher in the more populated areas of the network. With that feature in

mind, the PWR model allows us to assume that the effects of popularity exhibit increasing returns on the rate of content activation. One additional retweet of a user on the periphery of the network who rarely receives attention has a negligible effect on one's speed of content activation; the same increase in popularity has more substantial adverse effects in the speed of propagation on authorities located at the center of the network.

## CONCLUSION

As data availability increases, finding novel strategies to model big relational data has become a more pressing issue. In this chapter, we have emphasized the importance of efficient data management skills and the need for statistical models that are computationally efficient.

In the past decade, statistical advances in studying small and relatively homogenous social networks have been remarkable.



**Figure 30.9 PWR results in the network format. The model runs locally the impact of in-degree on time to retweet**

However, powerful new techniques such as exponential random graph models become technically challenging or altogether unfeasible as network size increases. Future methodological contributions to the study of social networks need to provide computationally feasible models for large heterogeneous networks.

The path weighted regression model is but one statistical alternative to extract meaningful information from very large social networks. Adaptation of standard error correction and auto-regressive models that can address network dependency seems to also be a valuable venue for computationally feasible estimation of inferential models for large networks.

## Notes

- 1 Twitter modified the way its APIs operate in August 2018, but the methodology explained here continues to be useful at the time of writing.
- 2 We direct the reader to some applications of GWR in the literature here (Darmofal, 2008; Cho and Gimpel, 2009; Calvo and Escolar, 2003).

## REFERENCES

- Adai, A. T., Date, S. V., Wieland, S. and Marcotte, E. M. (2004). LGL: creating a map of protein function with an algorithm for visualizing very large biological networks. *Journal of Molecular Biology*, 340(1):179–190.
- Aruguete, N. and Calvo, E. (2018). Time to #Protest: selective exposure, cascading activation, and framing in social media. *Journal of Communication*, 68(3):480–502.
- Blondel, V. D., Guillaume, J.-L., Lambiotte, R. and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008.
- Buja, A., Swayne, D. F., Littman, M. L., Dean, N., Hofmann, H. and Chen, L. (2008). Data visualization with multidimensional scaling. *Journal of Computational and Graphical Statistics*, 17(2):444–472.
- Calvo, E. and Escolar, M. (2003). The local voter: a geographically weighted approach to ecological inference. *American Journal of Political Science*, 47(1):189–204.
- Cho, W. K. T. and Gimpel, J. G. (2009). Rough terrain: spatial variation in campaign contributing and volunteerism. *American Journal of Political Science*, 54(1):74–89.
- Clauset, A., Newman, M. E. and Moore, C. (2004). Finding community structure in very large networks. *Physical Review E*, 70(6):066111.
- Csárdi, G. and Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal Complex Systems*, 1695:1695.
- Darmofal, D. (2008). The political geography of the New Deal realignment. *American Politics Research*, 36(6):934–961.
- Darmofal, D. (2015). *Spatial analysis for the social sciences*. Cambridge University Press, New York, NY.
- Fotheringham, A. S., Brundson, C. and Charlton, M. (2002). *Geographically weighted regression: the analysis of spatially varying relationships*. Wiley, Chichester.
- Fruchterman, T. M. and Reingold, E. M. (1991). Graph drawing by force-directed placement. *Software: Practice and Experience*, 21(11):1129–1164.
- Gibson, H., Faith, J. and Vickers, P. (2013). A survey of two-dimensional graph layout techniques for information visualisation. *Information Visualization*, 12(3–4):324–357.
- Girvan, M. and Newman, M. E. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826.
- Jacoby, W. G. (2000). Loess: a nonparametric, graphical tool for depicting relationships between variables. *Electoral Studies*, 19(4):577–613.
- Keele, L. (2008). *Semiparametric regression for the social sciences*. Wiley, Chichester; Hoboken, NJ.
- Kobourov, S. G. (2013). Force-directed drawing algorithms. In R. Tamassia, (ed.). *Handbook of Graph Drawing and Visualization*, Chapter 12, 383–408. CRC Press, Boca Raton, FL.
- Lloyd, C. D. (2010). *Local models for spatial analysis* (2nd ed.). CRC Press, Boca Raton, FL.

- Pons, P. and Latapy, M. (2006). Computing communities in large networks using random walks. *Journal of Graph Algorithms and Applications*, 10(2):191–218.
- Schmid, C. S. and Desmarais, B. A. (2017). Exponential random graph models with big networks: Maximum pseudolikelihood estimation and the parametric bootstrap. In *2017 IEEE International Conference on Big Data*, pp. 116–121. IEEE.
- Yang, Z., Algesheimer, R. and Tessone, C. J. (2016). A comparative analysis of community detection algorithms on artificial networks. *Scientific Reports*, 6:30750.

PART V

# Quantitative-Empirical Methods





*This page intentionally left blank*



# Econometric Modeling: From Measurement, Prediction, and Causal Inference to Causal-Response Estimation

Robert Franzese

One can identify four modes, or purposes, of empirical analyses of positive<sup>1</sup> political science and international relations: *measurement* and description, *testing* (of causal theory or ‘effects’), *prediction* or forecasting, and *estimation* (of causal models and causal responses or effects).<sup>2</sup> These alternative aims or ends one might have in empirical analyses will place emphasis on different methodological challenges and properties over others and so demand different methodological approaches and tools. Econometric modeling<sup>3</sup> is an approach and set of tools that can be useful toward all four ends, but it plays an especially crucial role in the last: causal-response estimation. *Causal responses*, as opposed to *treatment ‘effects’*, refer to how some outcomes of interest (dependent variables) respond to inputs of interest (independent variables or treatments). As such, causal responses are inclusive of the contextual conditioning and effect heterogeneity, of the temporal, spatial, and spatiotemporal dynamics, and of the causal-simultaneity feedback

that *treatment ‘effects’* purposefully exclude (in order to cleanly identify tests for causal-effect *existence*), and these heterogeneities, dynamics, and feedbacks cannot be estimated without modeling of the theoretical and substantive structure. Indeed, the specification and estimation of the empirical model, far from being an unfortunate unavoidable limitation, is, from that perspective and for those aims, the very goal of the exercise. The purpose of the analysis and the aim of the theoretically structured model is for its estimates to provide a ‘useful empirical summary’ of the actual substantive processes under study.

*Measurement* follows (as directly as possible) on *operationalization* – the translation from theoretical concepts,  $X$  and  $Y$ , to observable empirical indicators of those concepts (see Munck et al., Chapter 19, this *Handbook*) – to assign quantitative values gauging the extent or presence of those indicators in some unit of observation. As an end-goal of empirical analysis, measurement is distinct from the other modes in its purely descriptive aims:

scholars might conduct empirical analyses to offer measures of –i.e., to describe empirically– democracy (e.g., Coppedge et al., 2008; Treier and Jackman, 2008), the ideological placement of parties (Poole, 2019), and socioeconomic cleavage-structures (Selway, 2011), to name just three. Measurement is also distinct in being a fundamental prerequisite of any of the other modes of empirical analysis: causal inference, causal estimation, or prediction/forecasting. Econometric modeling is central to some measurement analyses (for examples, see Fariss et al., Chapter 20, this *Handbook*; Leemann and Wasserfallen, Chapter 21, this *Handbook*; and Treier, Chapter 48, this *Handbook*), but the focus of this overview remains the distinction between causal-effect *testing* and causal-effect *estimation* and the essential role of econometric modeling in the latter.

Regarding testing of causal theory and ‘effects’, the aim of the analysis is to evaluate empirically some causal-theoretical claim, i.e., to assess *whether* some posited causal relationship or causal effect *exists* empirically. Because the analyst’s central purpose is to *test* a particular theory, ideally as little as possible from beyond that theory will be brought into the empirical assessment, so as to isolate the ‘empirical existence proof’ of the hypothesized causal effect. An empirical implication is derived from the theoretical argument that some  $X \Rightarrow Y$ , and the empirical analysis aims to evaluate this argument, this causal proposition, this hypothesis; in other words, the researcher wants to verify empirically that  $dx \rightarrow dy$ ,<sup>4</sup> which entails (a) demonstrating that  $dx$  associates with  $dy$  empirically and (b) substantiating that the causal arrow goes from  $X$  to  $Y$  in the expressed direction. Notice that the adverb *empirically* applies only to component (a), empirical association of  $x$  and  $y$ ; it intentionally does not apply to establishing causality because *causality* is a theoretical and not an empirical concept. Thus, the validity of ‘empirical tests of causal “effects”’ rests on the strength of the empirical association and, separately, on the strength of the arguments establishing that

the causal arrow generating that empirical association goes in the theorized direction from  $X$  to  $Y$ . That is why the gold-standard ideal for *causal inference*<sup>5</sup> (see Bowers and Leavitt, Chapter 41, this *Handbook*) is the randomized controlled trial (RCT).

The potential-outcomes framework (POF)<sup>6</sup> proposes as the estimand for causal inference, i.e., for testing the existence of a causal effect of  $X$  on  $Y$ :

$$\text{Causal Effect} = Y_{it}(X = 1) - Y_{it}(X = 0) \quad (1)$$

The *fundamental problem of causal inference* arises immediately: for a single observation on unit  $i$  at time  $t$ , denoted subscript  $it$ , the treatment or causal impetus,  $X$ , either is present ( $X = 1$ ) or is not ( $X = 0$ ). The counterfactual cannot be observed. Empirical designs for causal inference typically then proceed to establish conditions under which the difference in the empirical sample-means of  $y$  under  $x = 1$  vs under  $x = 0$  can be taken as an estimate of (1).<sup>7</sup> In the POF, this involves designing an analysis in which the comparison treatment ( $X = 1$ ) and control ( $X = 0$ ) groups are identical in all ways except treatment status ( $x$  value) and, potentially, the outcome ( $y$  value). The association of  $x$  and  $y$ ,  $dx \rightarrow dy$ , is measured or estimated, and it can be understood as indicating the causal relationship  $dx \Rightarrow dy$  if two alternative causal-relationship possibilities can be ruled out as having instead generated that association: (a) that  $Y \Rightarrow X$  (i.e., *endogeneity*, for instance by *simultaneity* or reverse causality) and (b) that some  $Z \Rightarrow Y$  and  $Z \leftrightarrow X$  (*spuriousness*).<sup>8</sup> Now we can see why the RCT is the gold standard for causal inference. *Experimental control* (of  $dx$ ) assures that movements in  $x$  could not have been caused by  $y$ ; the analysts know  $y$  did not move  $x$  because the analysts themselves moved or manipulated  $x$ .<sup>9</sup> *Experimental randomization* in which unit-times receive  $dx$  rules out spuriousness because if the values of  $dx$  are successfully independently randomized across a very large number of observational units, then  $dx$  will be unassociated

with any alternative causal-factor  $Z$  – observed or unobserved (or even unimagined!) – by definition of *independent*.

The validity of the estimated empirical associations as test statistics for causal inference rests on the strength of the argument that the causal arrow underlying those associations runs as theoretically postulated. In some rare situations,  $Y \Rightarrow X$  may be ruled out *a priori*: e.g., few  $Y$  could logically possibly cause race or gender  $X$ , but experimentation – i.e., *successful* control and randomization (and large samples<sup>10</sup>) – usually offers the strongest possible argument. In these cases, so-called *nonparametric causal inference* – i.e., causal inference that does not rely upon a pre-specified structural model (no model beyond the additive and separable treatment effects inherent in the difference-in-means definition of *causal ‘effect’* in (1)) – may be feasible. Even here, though, the validity of the causal-effect interpretation of any empirical estimate of (1) requires that Stable Unit Treatment Value Assumption (SUTVA) hold. The SUTVA conditions can be understood as the conditions under which the empirical  $dx$  is validly as-if experimental, i.e., controlled and randomized, and they amount practically to the following:

The probability of one unit receiving treatment, the *homogenous* magnitude of the treatment, and the *homogenous* effect of treatment are independent of each other and of any other unit(s) receiving treatment, the sizes of treatments in those others, or effects of treatments in those others.

As one of POF’s founding protagonists suggests, ‘The two most common ways in which SUTVA can be violated [are] when (a) there are versions of each treatment varying in effectiveness or (b) there exists interference between units’ (Rubin, 1990: 282). If SUTVA is violated, for instance by treatment – ‘effect’ heterogeneity or conditionality and/or by spillovers or contagion across units  $i$  and time periods  $t$  – and such heterogeneity and interdependence are ubiquitous in socio-politico-economics – empirical estimates of

(1) by the sample-mean difference of  $y$  given  $x = 0$  or  $x = 1$

$$\widehat{E}(y | x = 1) - \widehat{E}(y | x = 0) \quad (2)$$

will not be valid or, speaking more precisely and generally, will be *inadequate* estimates of the causal effect of  $X$  on  $Y$ , or the expected empirical response of  $Y$ ,  $dy$ , to an exogenous movement in  $X$ , i.e., treatment,  $dx$ . In general, reclaiming valid causal-effect (existence) inference, and *a fortiori* any hope to claim valid causal-effect (response) *estimation*, will require econometric modeling beyond that of estimating (1) by its empirical analog (2).

In sum to this point, for purposes of establishing causality and non-spuriousness in the experimental sample, i.e., for *internal validity*, the RCT is indeed the gold-standard ideal (see Morton and Vásquez-Cortés, Chapter 51, this *Handbook*). Of course, all scholars accept that practical and ethical considerations constrain what can or should be experimentally manipulated in the purview of social science. Even beyond those feasibility limitations, however, except for purely descriptive purposes of determining what was true in the observed experimental unit-times, analysts are more concerned with inferring from that experimental sample to what would be true in new data, outside that observed sample, i.e., with *external validity*.<sup>11</sup> And for that purpose, three significant challenges of *representativeness* – of the experimental sample to the intended population of inference, of the experimental treatment to the causal variable of interest, and of the experimental context to that of the socio-politico-economic outcomes of interest<sup>12</sup> – hinder what can usefully be inferred from these high-internal-validity RCT studies to target populations of interest. As a practical, empirical, applied matter (see also note 11), it seems as though external validity dominates internal validity in mean-squared-error terms, where potentially biased observational studies in proper context generally yield smaller mean-squared errors than do unbiased experimental studies conducted

in *necessarily*<sup>13</sup> incorrect contexts (Pritchett and Sandefur, 2015). Moreover, even an ideal RCT is generally mute on other potential causes and says little about the magnitude of the treatment effects from the study relative to other effects<sup>14</sup> and relative to variation in the outcome variable of interest in the actual (external) context of interest.

Experiment[s] will have nothing whatsoever to say about other causes. What it will do, and do well, is to determine *whether* [...treatment...] had a positive or negative effect, or none at all. (Kellstedt and Whitten, 2009: 70; emphasis added)

Regarding external-validity concerns but staying within this *causal inference* or *testing-for-causal-effects* mode of empirical analysis, there has been much advancement in extra-laboratorial field- and survey-experimental research designs (see Sinclair et al., Chapter 52, this *Handbook*) and observational-study research designs (see Bowers and Leavitt, Chapter 41, this *Handbook*; Nielsen, Chapter 42, this *Handbook*; Keele, Chapter 43, this *Handbook*; Cattaneo et al., Chapter 44, this *Handbook*) to yield pseudo-experimental conditions for this causal 'effect' as defined in the POF (equation (1)). Treatment uptake by subjects is necessarily less strongly controlled in the field than in the lab, therefore, relative to the RCT laboratory experiment, field-experimental studies essentially trade some loss of purity in control and randomization for some enhancement of representativeness, perhaps of all three sorts (see note 12). Survey experiments, somewhat analogously, buy enhanced representativeness of subjects, given a scientific survey-design appropriate to the intended population, at the cost of representativeness of the treatment – mention or emphasis in a survey question, question ordering etc., are generally quite unlike the conceptual cause of interest in the theory – and of the context – answering a survey is usually very unlike the context to which the results are intended to be inferred.

For these reasons, social scientists sometimes must, and often *choose*, to work with

observational data, especially for purposes of inference beyond sample, i.e., of inferring a causal relationship to exist in some target population of interest, and not only that a cause operated in some (specific, observed, and past) experimental sample. As shall be demonstrated below, the move to econometric modeling of observational data becomes especially judicious as the aim of the analysis moves beyond establishing that some causal effect exists (*causal inference*), to estimating causal responses,  $dy/dx$  (*causal estimation*). With these moves beyond internal causal inference to external causal inference and, especially, further beyond to causal estimation, empirical analyses in social-science observational studies confront not one, but at least four, fundamental challenges; namely, in socio-politico-economic reality (Franzese, 2007):

- 1 *Multicausality*: just about everything matters;
- 2 *Causal heterogeneity and context conditionality*: how everything that matters varies – how everything matters depends on just about everything else, i.e., on context;
- 3 *Dynamic causality*: just about everything is temporally, spatially, or spatiotemporally dynamic, not static;
- 4 *Omnicausality*: just about everything causes just about everything else.

Further exacerbating these four challenges is a fifth (or zero-th) challenge, which is that even with the enormous quantities of data now obtainable from internet, social-media, satellite/geospatial, and other big-data sources (Nyhuis, Chapter 22, this *Handbook*; Barberá and Steinert-Threlkeld, Chapter 23, this *Handbook*; Darmofal, Chapter 24, this *Handbook*), observational empirical analysts often find relatively little useful empirical variation with which to surmount these hurdles, even in those oceans of data. In the first instance, this is where econometric modeling becomes essential: given heterogeneity, dynamics, or simultaneity, without some structural model to reduce the parameterization of the problem, the number of quantities

to estimate necessarily grows faster than the number of observations with which to estimate them.

Consider the following representation of empirical relations for one outcome of interest:

$$y_{it} = f_{it}(\mathbf{x}_{js}, \boldsymbol{\beta}_{js}, \boldsymbol{\varepsilon}_{js}); \boldsymbol{\varepsilon} \sim (\mathbf{0}, \boldsymbol{\Sigma}_{it}); \quad (3)$$

$$i, j = 1..N, t, s = 1..T, n = NT$$

In this perhaps most-general possible model,<sup>15</sup> there are  $k$  parameters,  $\boldsymbol{\beta}_{it}$ , linking right-hand-side variables,  $\mathbf{x}_{js}$ ,<sup>16</sup> to the outcome of interest,  $y_{it}$ , plus a mean-zero stochastic component,  $\boldsymbol{\varepsilon}$ , characterized by an  $n \times n$  variance-covariance matrix, itself possessing  $\frac{1}{2}n^2 + \frac{1}{2}n$  (which is greater than  $n$ ) parameters. In total, there are generally  $k + \frac{1}{2}n^2 + \frac{1}{2}n$  parameters to estimate *per function, per observation*. This number of quantities to estimate grows exponentially faster than does the number of quantities observed (a  $k + 1$  vector of  $y, \mathbf{x}_{it}$ ). Thus, without some extremely strong structural-modeling assumptions, there could be no empirical analysis at all. From this perspective, we see that so-called nonparametric causal inference POF approaches applying (2) are actually highly structurally modeled: (a) empirical relations are assumed constant across all observations (within a bin if some heterogeneous effects are allowed) so there is only one function,  $f$ , to estimate; (b) random components are assumed orthogonal and homogenous across observations (or assumed zero with deterministic relationships such that the only randomness enters through the experimental manipulation); and (c) the parameters,  $\boldsymbol{\beta}$ , are also assumed constant across all observations,  $it$  (within a bin).<sup>17</sup> Notice that these are essentially identical to the assumptions of classical regression analysis. Indeed, the typical empirical model in POF-based studies, (2), is in many ways much more restrictive than the typical empirical model in regression-based studies: additive, separable,

homogenous treatment effects (usually of a homogenous treatment:  $X = 1$  or  $X = 0$ ). The assumptions are the same because the logical necessity of some (radical) parameter reduction is the same. Therefore, the *arguments* or claims made about the research design, and not the estimation model, are the basis for POF causal-inference studies' claims to have ruled out spuriousness and simultaneity (again reflecting that causality is theoretical, not empirical).

Much of the econometric modeling deployed in the service of causal-inference studies focuses on ruling out spuriousness, i.e., some  $Z$  related to  $X$  actually causes  $Y$ :  $Z \leftrightarrow X$  and  $Z \Rightarrow Y$ . *Matching*-based inference (see Nielsen, Chapter 42, this *Handbook*), for example, leverages the idea that, if the researcher can observe and measure all relevant  $\mathbf{z}$ , then comparing  $y|x = 1$  to  $y|x = 0$  for 'balanced' groups of data – meaning data for which the empirical *distributions* (sample means, variances, etc.) of all  $\mathbf{z}$  are equal (or statistically indistinguishable) – yields a difference in means between treated and untreated observations that could not possibly be due to those  $\mathbf{z}$ . Note that matching, unlike the RCT, cannot control in this way for unobserved  $\mathbf{z}$ . Notice also that matching control for  $\mathbf{z}$  is exactly like regression control for  $\mathbf{z}$ , except that the former is much more robust. Multiple regression controls effects of  $\mathbf{z}$  that manifest in the manner modeled (e.g., linear effects only in linear regression), whereas matching controls effects of  $\mathbf{z}$  in any manner they may manifest.<sup>18</sup> One might thus say 'Matching control is regression control on steroids'. Finally, also like regression, matching *per se* offers absolutely no address of simultaneity; like regression, claim for the matching-based estimation of (2) to be causal rests entirely on the adequacy of the controls.

Another causal-inference econometric-modeling approach that is focused on eliminating the possibility of spuriousness is the difference-in-difference (DID) design (see Keele, Chapter 43, this *Handbook*), and, relatedly, the difference-in-difference-in-difference

(DID or 3D) design and fixed-effects (FE) designs. The key notion underlying DID econometric modeling is that differencing the data,  $y_{it} - y_{it-1}$ , removes any time-constant factor in  $y$ , including any unobserved (time-constant)  $\mathbf{z}$ . Observed  $\mathbf{z}$ , including time-varying  $\mathbf{z}$ , may be addressed by regression or matching control; scale-variation or other functional-form issues may likewise be addressed by (structural) econometric modeling. Empirical implementation of the DID design is very simple. One needs observations on two groups, both in a pre-treatment period in which the  $X$  of interest has not been applied and a post-treatment period in which  $X$  has been applied in one but not the other group. Regression analysis with an indicator for post-treatment period, an indicator for treatment status ( $x = 1$  or  $x = 0$ ), and the interaction of those two dummies yields a coefficient on the interaction of the (treatment) difference in (time) difference, i.e., of the causal ‘effect’ in (2), under the maintained assumptions. Like matching and regression (after all, DID is commonly implemented by estimating a simple dummy-variable-interaction regression model), DID *per se* offers no address to the possibility of endogeneity. Units may select the treatment because of values of  $y$  or expected values of  $y$ , for instance, invalidating the causal interpretation of the DID estimate.

The (Regression) Discontinuity Design ((R)DD) is also an econometric approach to causal inference (see Cattaneo et al., Chapter 44, this *Handbook*) but one focused on addressing simultaneity bias as well as *unobserved confounds*, a.k.a. omitted-variable bias. The DD capitalizes on situations in which a treatment,  $x = 1$ , is triggered – a discontinuous jump in the probability of treatment suffices – as an observed continuous *index variable*,  $v$ , crosses some threshold value,  $v_c$ . For instance, a candidate wins a plurality-election office when his/her vote share crosses the plurality threshold (e.g., Caughey and Sekhon, 2011) or a party’s probability of entering parliamentary government

jumps discontinuously upward when its seat share crosses the plurality threshold (Hays et al., 2019). Provided (a) there are no systematic differences at the threshold in variables,  $\mathbf{z}$ , other than the treatment variable (which can be evaluated empirically for observed  $\mathbf{z}$ ), and (b) no endogeneity in which observations fall near to either side of the threshold (called *sorting*, in this context), then exactly at, or at least near the threshold, it is completely random, or mostly random, whether the observation receives treatment. The causal ‘effect’ as defined in (1) is thus identified at the threshold as if by RCT: any observed or unobserved  $\mathbf{z}$  should be equal on either side very near the threshold (and this can be verified for observed  $\mathbf{z}$ ), and treatment is ‘applied’ randomly. Of course, like the RCT it aims to mimic, the DD estimate lacks external validity of its estimated treatment effect for conditions unlike those at the threshold (e.g., for not-close elections, which greatly limits the applicability of DD estimates of US Congressional incumbency advantage, e.g.). With some additional assumptions, one can estimate (2) by an RDD regressing  $y$  on a flexible polynomial in index,  $v$ , an indicator for treatment status,  $x = (0,1)$ , and the interaction of the polynomial terms with  $x$ . The coefficient on the  $x$  is then the RDD-identified effect.

The more full-throated econometric-modeling approach to causal inference relies upon instrumental variables (IV) (see Carter and Dunning, Chapter 40, this *Handbook*; see also selection modeling in Böhmelt and Spilker, Chapter 37, this *Handbook*) and, more full-throated still, systems estimation (see, e.g., Jackson, 2008). The causal-identification strategy of instrumentation is well known: given a causal relation,  $y=f(x\beta, \square)$ , about which there may be concerns that  $y \Rightarrow x$  as well, find some  $z$  that (a) covaries with  $x$  but (b) not with  $\square$  – alternatively, with more substantive appeal, this variable  $z$ , called an *instrument*, needs (a) to relate to  $x$  but (b) not to  $y$ , except through that relationship to  $x$  – and then estimate the relationship of  $y$  with  $z$  instead by indirect least-squares (ILS), for example:

$$\begin{aligned}
 & \left. \begin{aligned}
 (i) \quad y_{it} &= \beta x_{it} + \gamma z_{it}^y + \varepsilon_{it}^y \\
 (ii) \quad x_{it} &= \alpha y_{it} + \delta z_{it}^x + \varepsilon_{it}^x
 \end{aligned} \right\} \Rightarrow y_{it} \\
 & = \beta(\alpha y_{it} + \delta z_{it}^x + \varepsilon_{it}^x) + \gamma z_{it}^y + \varepsilon_{it}^y \\
 & \Rightarrow y_{it} = \frac{\beta\delta}{1-\beta\alpha} z_{it}^x + \frac{\gamma}{1-\beta\alpha} z_{it}^y \\
 & \quad + \frac{1}{1-\beta\alpha} \varepsilon_{it}^y + \frac{\beta}{1-\beta\alpha} \varepsilon_{it}^x \quad (4)
 \end{aligned}$$

The ILS coefficient-estimates may be solved to retrieve estimates of  $\beta$  and the other coefficients (once the residual correlation is accounted), but, for solely causal-inference *testing* purposes, the significance of the estimated coefficient on  $z_{it}^x$  may be evaluated directly. Note that instrumentation strategies, even in the causal-inference modality that aims to minimize modeling assumptions, must pre-specify enough about the causal process to at least offer a causal diagram (Pearl, 1995), if not a fully specified system of equations, to establish IV-identification conditions (a) and (b).

Two-stage least-squares (2SLS) is a convenient implementation of IV-estimation. 2SLS estimates equation (i) in (4), for example by regressing  $x$  on (any exogenous variables across the system as given by the model/graph and)  $z$ , or on  $\mathbf{z}$  if more than one instrument is available (stage 1), and then regressing  $y$  on that fitted  $x$  (and any exogenous variables in equation (i)) (stage 2). The 2SLS procedure puts the fitted- $x$  regressor in the same scale as the endogenous  $x$  so the 2SLS estimated coefficient on instrumented- $x$  estimates directly, and if there are multiple instruments, the *regression* of stage 1 is the optimal procedure for projecting multidimensional information  $\mathbf{z}$  to unidimensional  $x$  and  $y$ . Single-equation three-stage least-squares (3SLS), which is asymptotically equivalent to limited-information maximum-likelihood (LIML), gains efficiency relative to 2SLS by accounting the necessary non-sphericity in the stochastic component of the model seen in the compound error term in the last line of (4). For

further efficiency gains, the full system of equations (i) and (ii) can be estimated jointly by multi-equation 3SLS or by (asymptotic equivalent) full-information maximum-likelihood (FIML).<sup>19</sup> Systems approaches also facilitate the incorporation of cross-equation substantive/theoretical knowledge into the estimation, such as. e.g., that some coefficients are equal, proportionate, or oppositely signed across equations.

Strategies for addressing simultaneity and the other challenges for applied empirical social science can perhaps be enumerated from most to least structural (excepting item 0<sup>20</sup>) thusly:

- 0 Time ('the poor man's exogeneity');
- 1 Full-system specification and estimation;
- 2 (Single-equation) instrumentation;
- 3 Matching;
- 4 Difference-in-difference;
- 5 Discontinuity designs;
- 6 Survey and field experimentation;
- 7 Laboratory experimentation.

For *causal-inference* testing purposes, the ordering also lists in generally increasing *credibility*, given their decreasing reliance on information beyond the theory to be tested and the empirical data. Indeed, given the tremendous advantages of controlled randomization against spuriousness and reverse causality, what could possibly argue for econometric modeling, or model-based estimation in general, over the RCT or nonparametric causal-inference strategies in general? At broadest, and in general, the answer is: external validity.

Firstly, prior to any external-validity concerns, note that for description, summarization, and measurement purposes, causality is simply irrelevant. Experimentation would be exceedingly cumbersome, and piles of nonparametric estimates – being necessarily unconnected by any formula – tend to offer only poor summary and poorer understanding.<sup>21</sup> The dominance of model-based approaches of diverse kinds for textual-data analysis, scaling, or classification (see Benoit,



Chapter 26, this *Handbook*; Ergerod and Klemmensen, Chapter 27, this *Handbook*; Bouchat, Chapter 28, this *Handbook*), sentiment or network description (Curini and Fahey, Chapter 29, this *Handbook*; Calvo et al., Chapter 30, this *Handbook*), and latent-concept recovery (Fariss et al., Chapter 20, this *Handbook*; Leemann and Wasserfallen, Chapter 21, this *Handbook*; Treier, Chapter 48, this *Handbook*) is therefore natural and optimal. Rather than the RCT, the gold standard for measurement exercises is *usefulness* in conveying summary description or in subsequent analyses.

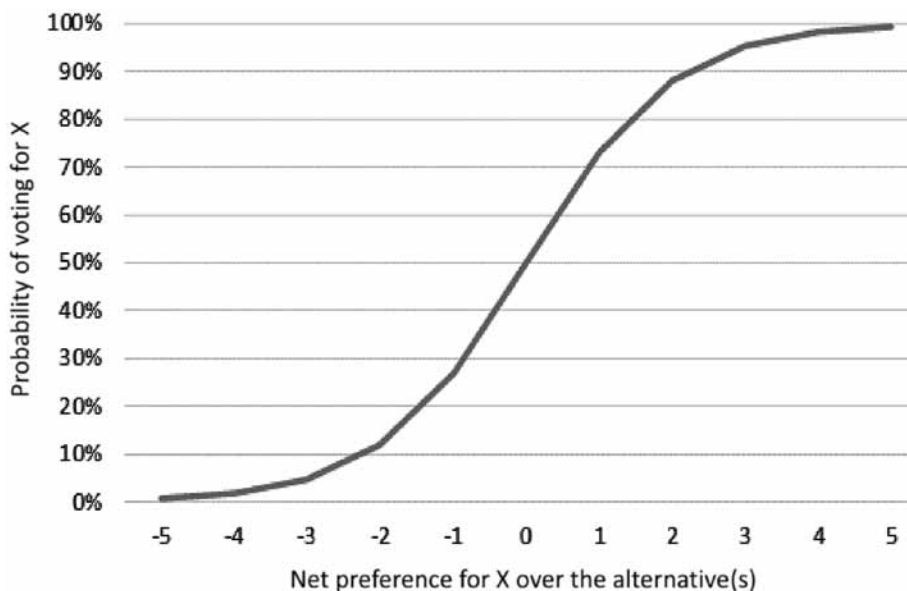
Likewise, for purely predictive and forecasting purposes, the gold standard is not recovery of some ideal-experimental results; it is (obviously) *out-of-sample prediction/forecast error* (see, e.g., Schrodtt and Gerner, 2000; for a similar view but comparing prediction to explanation instead of causal inference, see Ward, 2016). Again, internal validity is irrelevant insofar as the aim is to predict the value of some  $y_{j_s=i_t}$ , full stop; external validity is the only relevant consideration for pure prediction (see note 11). Here, too, econometric model-based strategies dominate, but in this case it is perhaps due more to the inherent limitations of non-parametrics than those of causal inference. Nonparametric estimates, by construction, offer no connection from  $E(y|x = x_0)$  to  $E(y|x = x_1)$ ; consequently, as the possible values that a potentially large number of useful predictors, their interactions, and interdependencies may take grows, the number of nonparametric estimates needed expands at least exponentially (and possibly combinatorially).<sup>22</sup> The forecasting device must somehow dampen this meteoric proliferation of necessary estimands; the preferred methods, having proved most effective, i.e., performing best by the sole relevant criterion, *out-of-sample prediction/forecast error*, include sophisticated econometric modeling with Bayesian methods (see Park and Shin, Chapter 47, this *Handbook*; Gill and Heuberger, Chapter 50, this *Handbook*),

particularly Bayesian model-selection and model-averaging (see Hollenbach and Montgomery, Chapter 49, this *Handbook*), Bayesian structural vector autoregression (Kilian and Lütkepohl, 2017), and machine-learning and artificial-intelligence methods (see Mikhaylov and Chatsiou, Chapter 55, this *Handbook*; Shoub and Olivella, Chapter 56, this *Handbook*).

As analysts' aims move beyond measurement, prediction, and *causal inference* or *testing* for (existence of) causal effects to *causal estimation* or *estimating causal responses*, the limitations of the experimental paradigm in the face of the four fundamental challenges for empirical research in social science become more pronounced. *Multicausality*, that there tends to be many relevant causes of effects, is least problematic, being addressed by control and, in many respects – as just discussed – ideally, by experimental control, at least for causal-inference purposes.<sup>23</sup> The limitations with respect to that first fundamental challenge relate to representativeness and external inference: comparing the treatment in the experimental sample and context to the intended treatments (causes) in their intended population and contexts. Effect heterogeneity raises more serious challenges. The structure of the Neyman–Holland–Rubin (NHR) causal model is of additive, constant, separable effects.<sup>24</sup> Effect heterogeneity or conditionality can, in principle (though see notes 22 and 24), be managed by binning observations with effects that are assumed homogenous within bins. However, to see how limiting this can be given socio-politico-economic reality, consider the simple case of the sigmoidal non-linearity implied in binary-choice and other binary-outcome contexts simply by the nature of probabilities or proportions. The essential substance of the matter dictates that for all binary outcomes, probabilities, or proportions,

$$\Pr(y = 1) \equiv p(y) = f(x, \beta, \varepsilon) \quad (5)$$

$f$ -being sigmoidal with  $0 \leq f \leq 1$



**Figure 31.1** The logically necessarily sigmoidal relation  $p(y) = f(x)$

A model of causal effects on probabilities that does not respect these first principles – the relationship between  $x$  and  $p(y)$  tapers toward its 0 and 1 bounds, (because it surely would not kink at those bounds) and so is steeper in some manner in between such as in Figure 31.1 (the NHR with  $\frac{dp}{dx} = c$  constant is one such non-respecting model) – is unlikely to yield very good *estimates* of those causal effects for external inference, especially for estimates traversing more curved portions of the S-curve and especially not beyond support. If nonlinearities like these are common, or more generally if effects in socio-politico-economic reality are typically heterogeneous and context-conditional as contended here, then the NHR causal model is a poor basis for *causal-effect estimation*, although it may remain a strong model for *causal inference* (see, e.g., Imai and Ratkovic, 2013; Egami and Imai, 2015).

The limitations of nonparametric causal inference in confronting causal heterogeneity and context conditionality are debilitating

for causal-effect estimation, for which purpose econometric modeling is inescapably essential. Far from an unavoidable detraction, however, estimation of an econometric model reflecting the theory and substance of the context is the very goal of the causal-estimation exercise. In this *Handbook*, as examples, see Fukumoto, Chapter 35, for a discussion of appropriate modeling of duration or survival contexts, and see Steenbergen, Chapter 36, for effective empirical-modeling strategies for parameter heterogeneity in multilevel/hierarchical contexts. For a quick illustration of how substantively theoretically specified econometric-model estimation can yield interesting and useful empirical science beyond proofs of causal existence, consider the implications of principle-agent/multi-actor bargaining for policy outcomes (Franzese, 2003, 2010). Equilibrium policy-outcomes in principle-agent and other shared-policy-control situations are some convex combinations of the two (or more) actors' optimal policies, e.g., a linear-weighted average, such as:

$$\begin{aligned}
 y &= \underbrace{c_p(\mathbf{z}) \times f(\mathbf{x}_p)}_{\text{principal control} \times \text{p action}} + \underbrace{[1 - c_p(\mathbf{z})] \times g(\mathbf{x}_a)}_{\text{agent control} \times \text{agent action}} \\
 \Rightarrow \frac{dy}{dx \in \mathbf{x} \equiv (\mathbf{x}_p \cup \mathbf{x}_a)} &= h(c_p(\mathbf{z})) \quad (6) \\
 \text{and } \Rightarrow \frac{dy}{dz \in \mathbf{z}} &= q(\mathbf{x})
 \end{aligned}$$

In words, the effect on the outcome of anything to which principal and agent (the bargainers) would respond differently depends on the degree to which each of the players is in control, which depends in a typical principal-agent model, e.g., on monitoring and enforcement conditions, represented in (6) by  $c_p(\mathbf{z})$ . Conversely, the effect of anything that shapes monitoring and enforcement costs and efficacy,  $z \in \mathbf{z}$ , depends on everything to which the players would respond (differently):  $x \in \mathbf{X}$ . By virtue of the shared influence of the bargainers over the outcome, the effect of any  $x \in \mathbf{X}$  to which they would respond differently depends on all  $\mathbf{z}$  in  $c_p(\mathbf{z})$ , the weight of each actor in determining the outcome, and, *vice versa*,<sup>25</sup> the effect of any  $z \in \mathbf{z}$  that influences the actors' relative control depends on all  $x \in \mathbf{X}$  to which the actors respond differently.

How can empirical researchers effectively estimate complexly context-conditional effects like these? One strategy is to impose the substantively known structure in the empirical model. Franzese (1999) estimates an empirical model like (6) to show how the anti-inflationary effects of central bank independence (CBI) – a situation of shared monetary-policy control, agent central bank and principal government – depend on political-economic conditions that would make governments more inflationary (bigger effect) or less inflationary (smaller effect). The convex-combinatorial form of (6) implies that only one additional parameter needs be estimated to capture all of these theoretically/substantively implied interactions; namely, this parameter is the factor of proportionality by which the central bank independence measure dampens inflation from the government's to the bank's preference as CBI increases. By a further nested pair of weighted

averages, Franzese (2003) extends the central bank/government domestic-actors model of 1999 to the open and institutionalized economy, wherein exchange-rate pegs effectively delegate from these two domestic actors to the peg-currency policy, and infinitesimally small capital-open economies, which effectively constrain domestic policy to the global average. Notice from the estimation model and results in Figure 31.2 that the 'theory-informed' model requires just two more parameter estimates than the linear-additive model, which completely lacks interactions, and 50 fewer than the linear-interactive model requires to generate comparable interactivity. And yet the coefficient estimates on small capital-openness,  $E$ , on single- and multi-currency pegs,  $SP$  and  $MP$ , and on CBI,  $C$ , are easily interpretable as the proportionate constraint each of those measures places on the opposite actors in its convex combination (see model (14) in Figure 31.2). The graphs illustrate two of the (many) rich substantive insights yielded about the context-conditional amplitude of partisan inflation-cycles at the top-right and about the generally declining anti-inflationary bite of CBI since about the 1970s, coinciding with the acceleration of the postwar and current great globalization.<sup>26</sup>

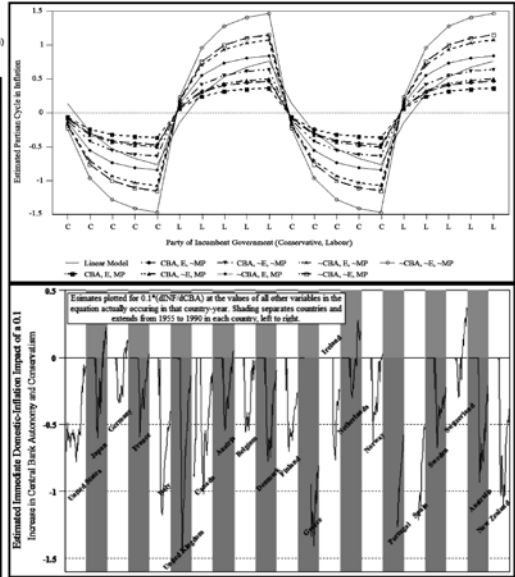
Next, consider how temporal and spatial dynamics highlight the inadequacy of the NHR model to causal-response estimation. Notice that the NHR estimand (1) and its typical empirical estimate (2) yield a scalar estimate, a single number, as the causal 'effect' of  $x$  on  $y$ . In a temporally dynamic context, in contrast, taking the simplest example to illustrate:

$$\begin{aligned}
 (i) \quad y_i &= \rho y_{i-1} + \beta x_i + \varepsilon_i \\
 (ii) \Rightarrow \frac{dy_i}{dx_i} &= \beta \\
 (iii) \Rightarrow \frac{dy_{LR}}{dx_{LR}} &= \underbrace{\beta dx}_{\text{period 0}} + \underbrace{\rho \beta dx}_{\text{period 1}} + \underbrace{\rho^2 \beta dx}_{\text{period 2}} + \underbrace{\rho^3 \beta dx}_{\text{period 3}} + \dots \\
 &= \underbrace{\sum_{s=0}^{\infty} \rho^s \beta dx}_{\text{assuming } |\rho| < 1 \Rightarrow \text{L.R multiplier}} = \frac{1}{1-\rho} \times \beta \times \underbrace{dx}_{\text{perm. shock}} \quad (7)
 \end{aligned}$$

$$E(\pi_t) = \beta_0 + \beta_1 E\pi_t + \beta_2 \pi_{t-1} + (1 - \beta_1) E\left\{ \begin{aligned} &(\beta_{10} GP + \beta_{11} ET + \beta_{12} UP + \beta_{13} BC + \beta_{14} AW + \beta_{15} FS + \beta_{16} TE + \beta_{17} \pi_t) \\ &(-\beta_{18} C) - \beta_{19} C_{t-1} \\ &(-\beta_{20} SP - \beta_{21} MP) + \beta_{22} SP + \beta_{23} MP + \beta_{24} \pi_t \end{aligned} \right\} \quad (14)$$

**Table 1** Alternative models of inflation in 21 OECD democracies, 1957–1990

Independent variable	Linear-interactive model (13)										Theory-informed model (14)
	C = 1, C = 1, C = 1, C = 1, C = 0, C = 0, C = 0, C = 0, C = 0, C = 0		E = 1, E = 1, E = 0, E = 0, E = 1, E = 1, E = 0, E = 0, E = 0, E = 0		P = 1, P = 0, P = 1, P = 0, P = 1, P = 0, P = 1, P = 0, P = 1, P = 0						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	
Intercept	+ .80 (6.1)				+ 5.93 (8.40)					+ .53 (3.0)	
Lagged inflation (%...)	+ .65 (.05)				+ .51 (.06)					+ .55 (.05)	
Twice lagged inflation (%...)	-. 03 (.04)				-. 10 (.04)					-. 12 (.04)	
Government partisanship (GP × X <sub>1</sub> )	-. 14 (.08)								+ 1.82 (.74)	-. 39 (.46)	
Productive year (EY × X <sub>2</sub> )	+ .59 (.30)								+ 4.81 (1.81)	+ 1.81 (1.32)	
Union power (UP × X <sub>3</sub> )	+ 2.19 (.74)								- 11.88 (3.32)	- 3.32 (1.62)	
Coordination of bargaining (BC × X <sub>4</sub> )	- 1.36 (.41)								+ 2.20 (.97)	+ 5.27 (2.35)	
Aggregate wealth (AW × X <sub>5</sub> )	+ .13 (.71)								- 2.24 (1.91)	- 3.43 (.49)	
Financial sector size (FS × X <sub>6</sub> )	-. 15 (.10)								- 1.00 (.71)	+ 4.68 (3.00)	
Trade exposure (TE × X <sub>7</sub> )	-. 04 (.99)								+ 45.70 (103.79)	- 120.2 (4.92)	
Inflation ahead (% × X <sub>8</sub> )	+ .39 (.07)								+ 1.8 (.33)	+ 2.65 (.24)	
Global financial exposure (E)	+ .29 (.73)									+ .44 (.14)	
Single-currency (single) peg (SP)	-. 33 (.49)									+ 1.04 (.05)	
Multi-currency (multi) peg (MP)	-. 37 (.38)									+ .22 (.12)	
Peg on global inflation (%...)										+ .59 (.07)	
Central bank independence (C)	- 1.62 (.68)									+ 1.03 (.13)	
Central bank target (%)										- .59 (.18)	
Obs. (Free)	660 (645)				660 (593)					660 (643)	
R <sup>2</sup> (S.E.R.)	.72 (2.48)				.75 (2.31)					.76 (2.30)	

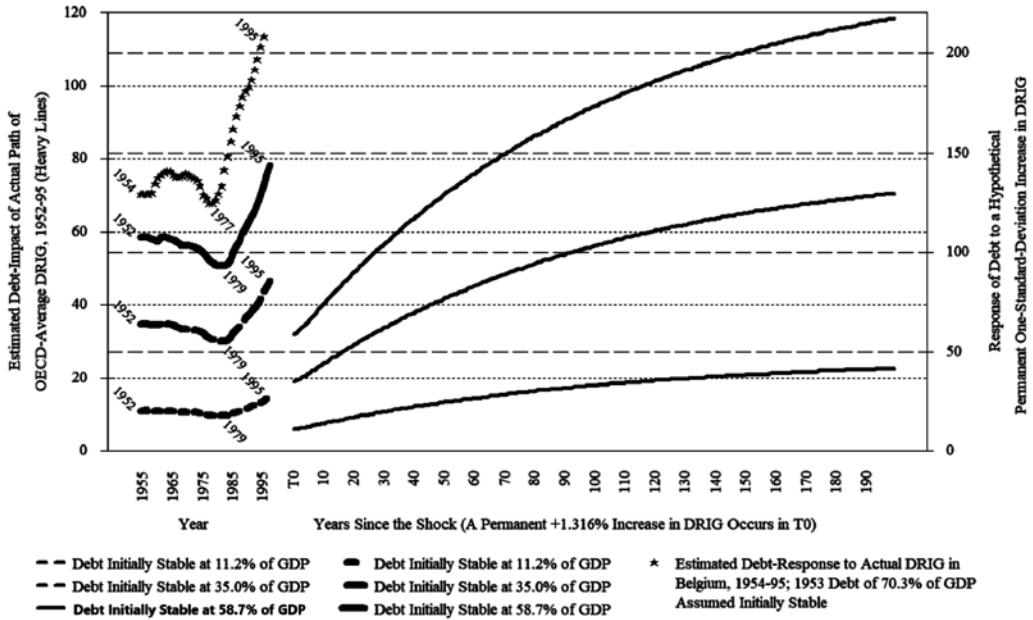


**Figure 31.2** ‘Multiple Hands on the Wheel’ model of complex context-conditionality in monetary policymaking

A causal-inference study designed to test whether  $x$  causes  $y$  will estimate  $\beta$ , which the second line of (7) demonstrates is only the contemporaneous response,  $dy_t$ , to a shock (treatment),  $dx_t$ , in that same period. As the lines (iii) show, if the shock persists (and nothing else occurs), the subsequent period experiences an additional  $\rho\beta$  and the period after that an additional  $\rho^2\beta$ , and so on. If the shock persists infinitely, the long-run steady-state (LRSS) response equals the long-run or temporal steady-state multiplier of  $1/(1-\rho)$  times the initial response,  $\beta$  (for further development and application, see Linn and Webb, Chapter 32, this *Handbook*; for fullest textbook treatment, Hendry, 1995). The single scalar<sup>27</sup> in (2) is obviously an inadequate-answer to the question, ‘what is the effect of  $x$  on  $y$ ?’, in the temporally dynamic context. In fact, the question is underspecified in the dynamic context: ‘the effect of (a movement in)  $x$ , when?, on (movements in)  $y$ , when?’<sup>28</sup>

Temporal dynamics matter greatly for substantive conclusions about causal-effect size.

Consider, e.g., the many well designed causal-inference studies on the effects of voter-registration hurdles, which typically find ‘very small effects’ on turnout (e.g., Hershey, 2009, reviews), but these are impulses,  $\beta$ , not effects,  $dy/dx$ . The considerable evidence that voting is a habit slowly acquired over repeated elections (e.g., Gerber et al., 2003) implies that voter turnout evolves dynamically, as in (7), so the response of voter turnout to registration-easing legislation is not a snapshot-in-time scalar but a vector over time, and with  $\rho$  being large, the long-run cumulative effects,  $\beta/(1-\rho)$ , are many times those previously estimated ‘very small’ causal parameters. Another illuminating example, from Franzese (2002), shows (in Figure 31.3) dynamic estimates from an econometric model of responses of public debt in developed democracies, counterfactually (a) to the actual OECD average real interest-rate (net growth) series 1954 to 1995 and (b) to hypothetical permanent a plus-one standard-deviation shock in real interest-rates proceeding indefinitely into the future, both



**Figure 31.3 Substantive dynamic-effect estimates of real interest-rate net of growth impacts on public debt**

Source: Franzese (2002).

starting from different initial debt-to-GDP ratios. Substantively, these dynamic estimates demonstrate that interest rates, i.e., monetary policy, could have enormous effects on the long-term accumulation of public debt and that, in fact, much of the post-1970s emergence of public-debt crises owed to that stagflationary era’s adverse shocks to growth and unemployment, inducing deficits which were followed by tight monetary-policy that spiked interest rates on those newly accumulating debts.

The inadequacies of the NHR model and estimand are highlighted further and amplified in the time-series cross-section (TSCS) and spatially/spatiotemporally dynamic contexts (see Troeger, Chapter 33, this *Handbook*; Cook et al., Chapter 39, this *Handbook*; relatedly, for dyadic-data and network analyses, see Neumayer and Plümper, Chapter 38, this *Handbook*; Victor and Khwaja, Chapter 45, this *Handbook*; Schoeneman and Desmarais, Chapter 46,

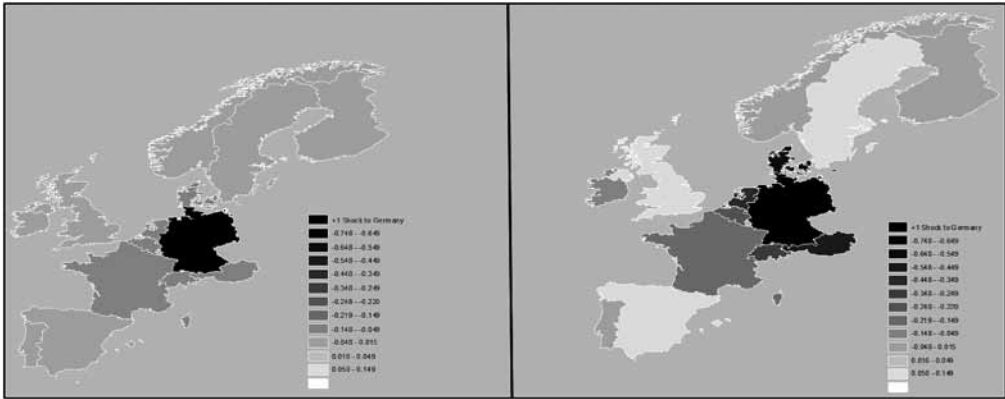
this *Handbook*). A very low-dimensional example, an  $n = 3$ -units cross-section with simultaneous first-order spatial-autoregressive interdependence (i.e., outcome contagion), suffices to demonstrate:

$$\begin{aligned}
 (i) \quad & \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} \\
 = \rho & \begin{bmatrix} 0 & w_{12} & w_{13} \\ w_{21} & 0 & w_{23} \\ w_{31} & w_{32} & 0 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} \\
 + \beta & \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{bmatrix} \\
 (ii) \quad & \mathbf{y} = \rho \mathbf{W} \mathbf{y} + \beta \mathbf{x} + \boldsymbol{\varepsilon} \\
 (iii) \Rightarrow & \mathbf{y} = [\mathbf{I} - \rho \mathbf{W}]^{-1} [\beta \mathbf{x} + \boldsymbol{\varepsilon}] \\
 (iv) \Rightarrow & d\mathbf{y} = [\mathbf{I} - \rho \mathbf{W}]^{-1} \beta \times d\mathbf{x}
 \end{aligned} \tag{8}^{29}$$

Again, the NHR estimand (1) and its empirical estimate (2), which a well designed and conducted experiment would produce, correspond to  $\beta$ , which we can call the ‘pre-dynamic *impulse*’ from  $x$  to  $y$ , and not ‘the effect of  $x$  on  $y$ ’, understood as how  $Y$  responds to movements in  $X$ , which is instead given by line (iv) of (8). Once again, the NHR estimand (1) and estimate (2) are in the wrong dimensionality, and the question is ill posed. Neither treatment nor effect are scalars: both are vectors/matrices and the effect statement is underspecified. In spatial/spatiotemporal contexts, the fully specified statement is ‘the effect of movements in  $\mathbf{x}$ , where (and when)? on movements in  $\mathbf{y}$ , where (and when)?’. In fact, in this case, the *impulse* is not even observable: the response in unit 1 – e.g., to some  $dx$  in, say, unit 1 itself – begins with the *impulse*,  $\beta \times dx$ , but that *instantaneously* induces proportionate movements  $\rho w_{1j}$  in units 2 and 3, which *instantaneously* induces proportionately smaller<sup>30</sup> movements  $(\rho w_{1j})^2$  in units 1 and 3, and so on, reverberating through the units across space analogously to the temporal-dynamics case, but *omnidirectionally* and all *simultaneously*. Thus, the pre-dynamic impulse,  $\beta$ , never manifests observably at all, only its steady-state implications in (iii) and (iv) show empirically. Figure 31.4 maps the estimated responses in a spatiotemporal econometric model from Franzese and Hays (2006), regarding active labor-market (ALM) spending to a hypothetical 1€ (per unemployed worker) spending increase in Germany. The left panel shows the estimated response across all EU countries<sup>31</sup> in the time period contemporaneous to the shock (inclusive of that period’s spatial feedback but exclusive of any time dynamics); the right panel shows the LRSS accumulated response, inclusive of all spatiotemporal feedback. The econometric model uncovered free-riding behavior, i.e., negative spatial interdependence; the characteristic oscillating pattern of negative autoregression is apparent in the right panel.

The case of spatial interdependence also underscores the radically limited scope for nonparametric causal inference, even with regard only to testing, in a socio-politico-economic reality characterized by omnidirectional causality (fundamental challenge number four). Notice from (8) that the effects of  $X$  on  $Y$ , i.e.,  $dy/dx$  in line (iv), impinge in general on all units,  $d\mathbf{y}$ , and vary – the *vector* of effects differs – depending on (a) which units are treated, i.e., the specific allocation of treatments across units,  $d\mathbf{x}$ , and (b) on  $\mathbf{W}$ , the relative connectivity among units, i.e., the specific set of  $\{w_{ij}\}$  connecting units according to which the contagion diffuses. Thus, proceeding nonparametrically, each possible allocation of 1s and 0s across the  $n$  units – there are  $2^n$  such permutations – corresponds to a different treatment; the effect of each such vector of treatments depends further on  $\mathbf{W}$ , which in general has  $n(n-1)$  potentially unique elements (yielding  $2^{n(n-1)}$  possible  $\mathbf{W}$  if connectivity is binary and  $\infty$  if continuous). Thus, there are minimally  $2^n \times 2^{n(n-1)} \gg n$  treatment effects to estimate nonparametrically: obviously impossible without considerable structure (which can be productively imposed in the form of Bayesian hyperpriors in this context, see, e.g., Best et al., 2005). Because at  $\rho = 0$  the allocation of treatments and contents of  $\mathbf{W}$  are irrelevant, a sharp null hypothesis may be formed to test *whether* spatial interdependence is present, but that is the extent of the possible nonparametrically:  $\rho$  and  $d\mathbf{y}/d\mathbf{x}$  are inestimable without considerable structure. Indeed, notice that spatial *association*, i.e., correlation, leaving aside causality entirely, cannot even be measured until the elements of  $\mathbf{W}$  are specified and thereby *proximity* defined.

Simultaneous spatial interdependence is also illustrative as a special case of causal-systems simultaneity, with line (i) of (8) giving a system of equations with three endogenous variables:  $y_1, y_2, y_3$ . Thus the discussion from the spatial-interdependence



**Figure 31.4** Maps depicting the initial (left panel) and LRSS (right panel) spatial ALM-spending responses to +1 shock in Germany

Source: Franzese and Hays (2006).

case applies also, *mutatis mutandis*, to causal systems of simultaneous equations more generally. Socio-politico-economic contexts with cross-unit contagion,  $y_i \leftrightarrow y_j$ , or other simultaneous causality,  $y_i \leftrightarrow x_j$ , imply processes like:

$$\begin{aligned}
 (i) \quad & y = \alpha_0 + \alpha_1 x + \alpha_2 z_y + \varepsilon_y \\
 & x = \beta_0 + \beta_1 y + \beta_2 z_x + \varepsilon_x \\
 \Rightarrow (ii) \quad & y = \alpha_0 + \alpha_1 (\beta_0 + \beta_1 y + \beta_2 z_x + \varepsilon_x) \\
 & \quad + \alpha_2 z_y + \varepsilon_y \\
 y = & \frac{[\alpha_0 + \alpha_1 \beta_0 + \alpha_1 \beta_2 z_x + \alpha_2 z_y + \varepsilon_y + \alpha_1 \varepsilon_x]}{(1 - \alpha_1 \beta_1)} \\
 \Rightarrow (ii) \quad & \frac{dy}{dx} = \frac{\alpha_1}{1 - \alpha_1 \beta_1}, \text{ and not } \frac{dy}{dx} = \alpha_1
 \end{aligned}
 \tag{9}$$

A well designed experiment, or a valid discontinuity or instrumentation design, will identify and estimate  $\alpha_1$ , which is indeed sufficient to test *whether X affects Y*, since

that effect,  $\frac{dy}{dx} = \frac{\alpha_1}{1 - \alpha_1 \beta_1}$ , is zero if  $\alpha_1 = 0$ , but it is clearly insufficient to *estimate* the effect, i.e., the causal response. This is because experiments work to identify the *existence* of causal effects precisely by preventing *estimation* of causal *responses* in the actual simultaneous system of interest. Specifically, causal-inference designs aim to block, *internally*, the feedback from  $y$  to  $x$  that actually occurs *externally* in the inference population. In the contexts of actual interest and intended application, if one ‘moved’  $x$ , this would create *impulse*  $\alpha_1$  to  $y$ , but that in turn would spur  $\beta_1$  further movement in  $x$ , which would move  $y$  some more, which in turn would move  $x$ , and so on.<sup>32</sup> Thus, ironically, experiments and non-parametric causal-inference designs estimate causal *parameters*, like  $\frac{\partial y}{\partial x} = \alpha_1$ , not causal *effects*, like  $\frac{dy}{dx} = \frac{\alpha_1}{1 - \alpha_1 \beta_1}$ .

In other words, notwithstanding that the NHR labels the estimand (1) and its empirical estimate (2) ‘the causal effect’, the aim of studies deploying them is actually *causal inference*, i.e., to establish *whether* a particular causal relation *exists* and not *causal-effect estimation*. The latter is to *estimate how* (not whether) outcomes of interest respond to inputs of interest, i.e., to estimate  $\frac{dy}{dx}$ , where those are expressly total rather than partial derivatives or differences, and the response,  $dy$ , and/or the treatment,  $dx$ , may actually be vectors or matrices of counterfactuals.

As already discussed, causal inference naturally emphasizes internal validity, whereas description and prediction instead stress external validity. Causal-response estimation, for its part, is similar to prediction in that its gold-standard ideal is an out-of-sample performance of the response estimate, emphasizing external validity; but, it is also similar to causal inference in that the external responses it aims to estimate are *causal effects*, not merely to predict  $E(y_{it}|x_{jt})$ , but to predict how  $y_{it}$  would respond (conjunctive tense) causally to hypothetical movements in  $x_{jt}$ : predictive (counterfactual) causal-response estimation. Internal causal validity is also crucial.

Given that the NHR model is inadequate, for causal-response-estimation purposes, to meet the challenge of ubiquitous simultaneity, progress under *omnicausality* (‘just about everything causes just about everything else’) will rely on substantively/theoretically informed econometric modeling, as it did also in fruitfully addressing effect heterogeneity and context conditionality, and spatial, temporal, and spatio-temporal dynamics.<sup>33</sup> To begin, consider the general case of (linear) systems simultaneity, noticing the similarity to the spatial simultaneity in (8) and the bivariate case in (9):

$$\begin{aligned}
 (i) \quad & \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ \vdots \\ y_m \end{bmatrix}'_i \begin{bmatrix} \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1m} \\ \gamma_{21} & \gamma_{22} & \cdots & \gamma_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ \gamma_{m1} & \gamma_{m2} & \cdots & \gamma_{mm} \end{bmatrix} \\
 & + \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_k \end{bmatrix}'_i \begin{bmatrix} \beta_{11} & \beta_{12} & \cdots & \beta_{1m} \\ \beta_{21} & \beta_{22} & \cdots & \beta_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ \beta_{k1} & \beta_{k2} & \cdots & \beta_{km} \end{bmatrix} \\
 & = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \vdots \\ \varepsilon_m \end{bmatrix}'_i \\
 (ii) \quad & \underline{\mathbf{y}}'_i \begin{bmatrix} \Gamma \\ \Gamma \\ \vdots \\ \Gamma \end{bmatrix} + \underline{\mathbf{x}}'_i \begin{bmatrix} \mathbf{B} \\ \mathbf{B} \\ \vdots \\ \mathbf{B} \end{bmatrix} \\
 & = \underline{\boldsymbol{\varepsilon}}'_i \left\{ \begin{array}{l} \text{which normalizing diagonal } \gamma_{ii} \text{ to 1 and} \\ \text{reversing sign of } \gamma_{ij} \text{ can be written for} \\ \text{all } N \text{ obs:} \end{array} \right. \\
 (iii) \quad & \underline{\mathbf{Y}} = \underline{\mathbf{Y}} \begin{bmatrix} \Gamma^* \\ \Gamma^* \\ \vdots \\ \Gamma^* \end{bmatrix} + \underline{\mathbf{X}} \begin{bmatrix} \mathbf{B} \\ \mathbf{B} \\ \vdots \\ \mathbf{B} \end{bmatrix} + \underline{\mathbf{E}} \\
 & \Rightarrow \mathbf{Y} = (\mathbf{X}\mathbf{B} + \mathbf{E})(\mathbf{I} - \Gamma^*)^{-1} \\
 (iv) \quad & \Rightarrow d\mathbf{Y} = (d\mathbf{X})\mathbf{B}(\mathbf{I} - \Gamma^*)^{-1}
 \end{aligned}
 \tag{10}^{34}$$

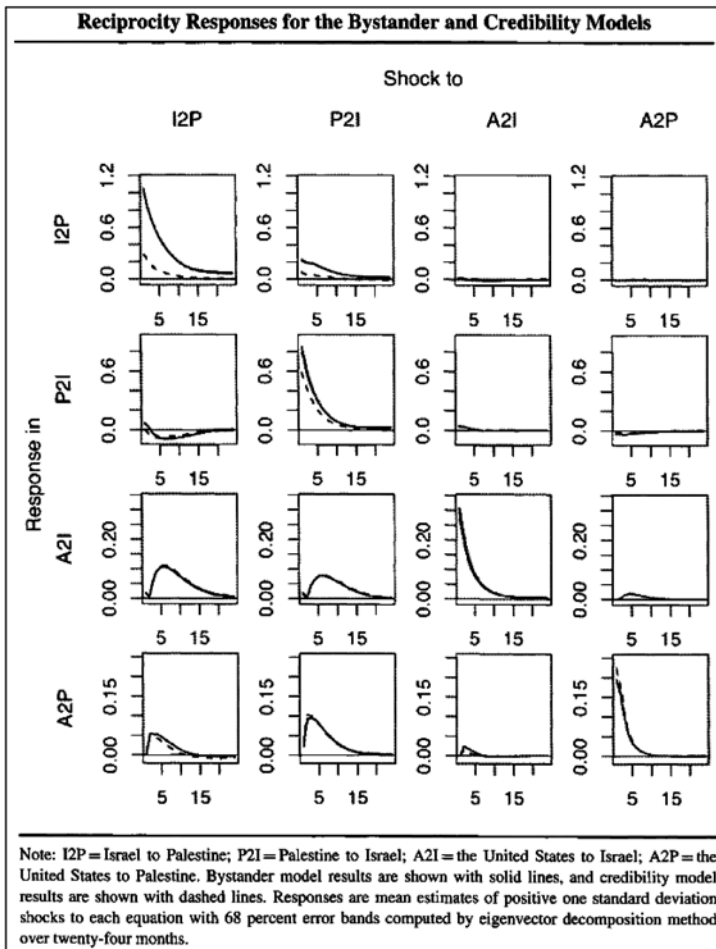
Thus, the translation from the causal-parameter estimates yielded by a well-designed experiment or single-equation causal-inference strategy like discontinuity or instrumental-variable designs to causal-response estimates involves a systems steady-state multiplier,  $(\mathbf{I} - \Gamma^*)^{-1}$ , analogous to the temporal multiplier in (7), spatial multiplier in (8), and bivariate-system multiplier in (9). Thus, causal-effect estimation requires systems estimation (Jackson, 2008, is an excellent exposition<sup>35</sup>), or at least somehow an estimation of all the parameters of the properly modeled system relevant to some desired out-of-sample causal-response estimation. Unfortunately,



political science and international relations rarely focus attention on system estimation, despite *strategic interdependence* being definitionally core to both. The emphasis on theory testing and its consequent idealization of nonparametric causal inference likely bears some blame for this, notwithstanding that systems interdependence greatly complicates even hypothesis testing and makes econometric modeling inescapably essential to causal-effect estimation.<sup>36</sup>

One econometric-modeling approach that does focus squarely on dynamic systems of endogenous equations is the Bayesian

Structural Vector Autoregression (BSVAR) framework. In one illustrative application, Brandt et al. (2008) uncover the reciprocity and other reactions between the Israeli government and military, Palestinian groups, and US official diplomatic and foreign-policy actions (Figure 31.5).<sup>37</sup> The discussions above prove that these rich substantive interrelations could not be estimated in a nonparametric causal-inference approach and that, in fact, even testing for the existence of causal effects related to the alternative reciprocity, accountability, and credibility theories of these actors' strategically interdependent behavior would likely



**Figure 31.5 BSVAR Estimated responses from a system of Israel↔ Palestinian, US→Israel, US→Palestinian actions**

Source: Brandt et al. (2008).

fail using such a framework due to ubiquitous ‘heterogeneity and spillovers’ (see Rubin (1990: 282) again).

## CONCLUSION

Empirical analyses in political science and international relations have at least one of four different goals: description, prediction, causal inference, and causal estimation.<sup>38</sup> The different aims carry with them different sets of weights on desiderata like internal and external validity, robustness, flexibility, efficiency, and richness. For causal-inference theory-testing purposes, the aim is to establish empirically the *existence* of a causal relationship as credibly as possible. For these purposes, the (a) randomized (b) controlled trial, being the strongest-possible guard against spuriousness and reverse causality and with the greatest possible nonparametric purity against ‘model dependence’, represents the gold-standard ideal. However, these causal internal-validity strengths generally come at some costs in terms of representativeness of experimental sample, treatment, or context, thereby limiting the utility of the causal inferences in practical applications, which necessarily require extra-sample inference and therefore external validity, which are not strengths of *laboratory* experimentation. Accordingly, social-science experimentation moves into the field and the survey to enhance representativeness, and from there into econometric-modeling techniques like matching, difference-in-difference, discontinuity, and instrumentation designs in observational data, which tend to maximize representativeness of the sample to the population and context of intended inference. From this perspective, the remaining threats to valid causal inference are effect heterogeneity and spillovers, violations of SUTVA that would bias causal hypothesis tests. In fact, the challenges of multicausality – effect heterogeneity and context-conditionality,

temporal, spatial, spatiotemporal dynamics and interdependence (spillovers), and omnicausality (spillovers) – are ubiquitous, indeed virtually definitionally central, in social-science and in socio-politico-economic reality. From a causal-inference perspective, econometric modeling is aimed to address these challenges to valid designed-based testing of positive social-science theory. Beyond causal inference, however, this chapter has noted that, in the first instance, causality is irrelevant for measurement and prediction empirical purposes. The gold-standard ideal for measurement is neither internal nor external validity, but rather *usefulness*,<sup>39</sup> toward providing summary descriptions and conveying information and understanding thereof. For prediction, the ideal is out-of-sample performance, i.e., prediction or forecast error. Whether the prediction input-output algorithm involves causal relationships or simply associations is irrelevant; external rather than internal validity is crucial. Econometric modeling can play central roles in both prediction and measurement, but its essential and most important role lies in causal-response estimation. As this chapter has demonstrated, valid causal-inference studies can at best provide estimates of causal *parameters*, not causal *effects*. Given the complex causal heterogeneity and context conditionality that characterizes socio-politico-economic reality, in fact, even causal-parameter estimation is impossible without considerable ‘structure’, ideally via theoretically/substantively informed econometric-model specification. The limitations of (so-called) nonparametric causal inference, even solely for testing theories, and the necessity and virtues of econometric modeling become even more pronounced given the temporal, spatial, spatiotemporal, and systems causal interdependence of socio-politico-economics. Careful theoretically/substantively informed specification of econometric models is the essential heart of empirical analysis for purposes of estimating causal effects, and the aim of and gold

standard for empirical modeling is to provide *useful empirical simplifications* of the actual, empirical causal processes of interest.<sup>40</sup> From this perspective, far from an unavoidable detraction, estimation of an econometric **model** reflecting the theory and substance of the context is the very goal of the causal-estimation exercise.

## Notes

- 1 *Positive* here opposes *normative*, positive theory being about how the world works in actuality as opposed to how the world *ought* to work normatively or would work in some fictional *ideal*.
- 2 One can also distinguish two types of empirical questions: factual questions ('what happened or will happen?') and causal questions ('why did or will something happen?'). The former have empirically extant, finite populations and deterministically true answers – 'what percentage of the citizens of certain country approve of the government's performance?' – and the latter have hypothetical populations and uncertainly estimated answers (of theoretical *because*s): 'what characteristics of citizens, governments, and performance affect citizens' approval of governments?'
- 3 To offer a definition, (the purpose and evaluative standard of) an *econometric model*, analogously to a theoretical model in Clarke and Primo (2012), is (to be) a *useful empirical simplification*.
- 4 The variables  $x$  and  $y$  are empirical measures, here assumed to be wholly unproblematic, of the theoretical concepts,  $X$  and  $Y$ ;  $dx \Rightarrow dy$  is the empirical implication derived from the theoretical argument,  $X \Rightarrow Y$ .
- 5 Notice the word used in this testing context is *inference*, and not *estimation*; this is because the central aim is to infer the existence of a causal effect, i.e., to establish that  $dY/dX \neq 0$ , rather than to estimate it. The empirical estimand from a causal-inference design is most usually but not necessarily, a difference in means,  $E(y|x = 1) - E(y|x = 0)$  (see Bowers and Leavitt, Chapter 41, this *Handbook*), which will only in very specific (and likely exceedingly rare in social science) conditions equate to an empirical estimate of the true causal effect of  $x$  on  $y$ , understood as  $dy/dx$ , i.e., how  $y$  responds to a causal impetus from  $dx$ .
- 6 The POF, also called the Neyman–Rubin or Holland–Neyman–Rubin *causal model*, is indeed a model: the causal effects described in (1) are discrete, static, additive, and separable. Indeed, precisely these characteristics of the POF/(H)NR causal model simultaneously make it so powerful for *causal inference* (testing theorized causal-effect existence) and yet so limited for *causal estimation* (estimating empirical causal effects or responses). See also note 17 and the discussion throughout the rest of this chapter.
- 7 Other counterfactually defined estimates have been proposed for causal inference/testing (in political methodology, e.g., see Bowers, 2013), but by far the most common practice is to define the causal quantity of interest as in (2).
- 8 Double-lined arrows indicate causal relationships and single-lined arrows empirical ones, i.e., associations.
- 9 Some scholars go so far as to suggest that if  $x$  cannot be manipulated, *race* for example, then it cannot be causal, but this confuses the empirically implementable with the logically possible. *Causality*, being a theoretical and not empirical concept, involves only the latter; the former is irrelevant (see, e.g., Woodward, 2016 for a fuller discussion).
- 10 If empirical outcomes,  $y$ , are less than perfectly fully determined by the experimentally controlled  $x$ , such that there remains some residual component in  $dy$ , even if orthogonally random but especially if possibly systematically caused or related to alternative causes  $Z$ , then a large sample, successful randomization, and some reliance upon some form of central limit theorem are also essential to proper interpretation of test statistics from a RCT.
- 11 Some scholars contend oppositely, that internal validity has lexical priority over external validity, that internal is more important and without it external has no value. Imbens (2010), e.g., suggests that instances where one could conduct the appropriate experiments and would choose observational data instead are inconceivable. To debate whether internal or external validity is more important or, especially, which is lexically prior is obviously inane: of course, one wants *both*, the aim being to infer (a) validly and (b) from the observed and already known to new contexts. If we must debate priority though, clearly the more defensible position is the reverse: external validity without internal validity (i.e., non-causal empirical associations within sample that obtain also beyond sample) is still useful, e.g., for prediction, whereas an internally validated causal relationship with no external validity has only descriptive value within the already observed and known sample and zero use in any context beyond the study, i.e., for *inference*.

- 12 The representativeness of an experimental sample to the intended population of inference refers to the equivalence of the subjects in the experimental sample and the external units to which the results of the study are to be inferred: college sophomores in a certain class compared to voters in actual democracies, for example. Representativeness of treatment analogously refers to the equivalence of the experimentally manipulated treatment to the concept theoretically understood as causal in the population of inferential interest, e.g., mention of party affiliation in a paragraph that trial subjects are given to read compared to the actual partisanship of actual bill sponsors in actual political contexts. Representativeness of context refers to the equivalence of the situation of the experimental subjects and treatments relative to each other and relative to the relevant socio-politico-economic reality outside the experiment compared with the situations in these regards of the intended inference population: randomized application of a campaign strategy that subjects read about or experience in a media lab in an experiment's contrived campaign compared to campaign *strategies* (by definition) strategically (which means interdependently) chosen by competing parties in actual campaigns where publics are going about their lives, not engaging in a social-science experiment.
- 13 The RCT obtains its strong causal-inference properties precisely by designing an unnatural context: feedback, which exists in nature, is severed by experimental control, and experimentally randomly independently assigned treatments are in nature likely non-randomly and often even strategically assigned. See also note 12.
- 14 Conjoint experiments offer some advances in this specific regard (see, e.g., Hainmueller et al., 2014).
- 15 The  $i, j, s, t$  subscripts are intended to signify that  $y_{it}$  may be a function of  $\mathbf{x}$ ,  $\beta$ ,  $\varepsilon$  in any units  $i$  or  $j$  and periods  $s$  or  $t$ .
- 16 To be as fully general as possible,  $\mathbf{x}_{j_s}$  may include  $y_{j_s}$  and/or temporal and/or spatial lags of  $\mathbf{x}_{i_t}$  as well.
- 17 To elaborate these points more precisely, the POF estimand can be conceived as nonparametric estimate of some *average* treatment 'effect' (ATE), regardless of what functions,  $f$ , may have generated that average difference in means, and this may be adequate for purposes of testing whether this ATE is non-zero (the orthogonality of unobserved random components still seems necessary). However, to interpret this ATE as an *effect*, i.e., as an estimate of how  $y$  would respond to some exogenous  $dx$  *outside of the observed sample*, is to treat it as a model.
- 18 In fact, the similarity of matching and regression control extends further: regression controls  $\mathbf{z}$  to the degree its effects manifest as modeled; matching controls any manifestation of effects of  $\mathbf{z}$  provided, or to the degree, the appropriate form of  $\mathbf{z}$  is included in the matching balancing.
- 19 Jackson (2008) offers a more complete introduction to instrumental-variable and systems estimation.
- 20 *Time* here refers to arguments that 'it happened yesterday, therefore it's exogenous', which is not guaranteed in socio-politico-economic applications, where human foresight can give causal weight to current expectations of futures. Moreover, exclusive reliance on temporal precedence for identification is highly susceptible to specification error.
- 21 *Nonparametric* here references methods that yield large numbers of discrete, unconnected values as distinct from methods explicitly intended to produce (likely graphical) descriptions that are not *a priori* structured but are smoothed descriptions (see Pagan and Ullah, 1999, for far fuller coverage of nonparametric econometrics).
- 22 Again, see Pagan and Ullah (1999) for a much fuller view of nonparametric analyses; here, we intend nonparametric causal inference or causal estimation specifically, which necessarily entail distinct causal 'effect' estimates for each and every context, there being allowed no functional smoothing connections between 'effects' in different conditions. A paradigm labeled *evidence-based medicine*, which carries considerable weight in the biomedical sciences, is illustrative here. The notion is that, if a well and credibly designed RCT yields reliable results that treatments of certain medicines in certain doses to patients with certain conditions, characteristics, and histories produces some estimated net-benefits, then, regardless of whether that RCT-estimated effect has some theoretical explanation, the treatment is to be applied. This is a purely predictive approach but one that attempts to retain the nonparametric foundations of the RCT. As such, the model on which it relies for external validity, i.e., the predictive basis on which to prescribe the treatment, is like that of matching: matching treatments applied to patients with matching conditions will have the same effect. No basis is provided for applying only *similar* treatments to only *similar* patients; that would require more of a model.
- 23 Chapter 2 of the classic text *Statistics* (Freedman et al., 2007 [1978]) extols the two great virtues of experimentation. Even in the examples

- mooted there, though, some doubts of universal unmitigated virtue may be raised. For instance, when double-blind randomization is assumed vindicated because surgeons who know the health of their patients and the nature and severity of their ills yielded significantly beneficial results of an experimental surgery whereas blinded ones insignificantly so, this could, instead of suggesting pernicious bias, suggest effect heterogeneity, about which the non-blinded surgeons in the study know, as would – crucially – surgeons in actual practice. Chapter 3 then warns severely of the dangers of observational studies, lacking those two great experimental virtues. An interesting pattern develops, however: each example observational study's conclusion is overturned later by ...another observational study, plus *arguments* that the latter was better designed...because *causality is ultimately a theoretical, not an empirical, matter*. Finally, the examples have also shifted from primarily clinical-medical in Chapter 2 to primarily epidemiological in Chapter 3, and epidemiology, like '[macro]economics [and most political science and international relations] is not an experimental science' (Sims, 2010).
- 24 In practice, treatments are also nominal,  $x = (0,1)$ . Although claim is often made to straightforward extensions for continuous treatments, in fact the extension is generally complicated and incompatible with nonparametric causal inference, as explained in the surrounding text (see also note 22).
  - 25 All interactions are symmetric in this way: how  $\mathbf{z}$  moderates the effect of  $\mathbf{x}$  on  $y$ ,  $d(dy/d\mathbf{x})/d\mathbf{z}$ , is identical to how  $\mathbf{x}$  moderates the effect of  $\mathbf{z}$  on  $y$ ,  $d(dy/d\mathbf{z})/d\mathbf{x}$ , because interactive effects, i.e., effects on effects, are cross derivatives, and the order of differentiation in a cross derivative is irrelevant.
  - 26 In 'The multiple effects of multiple policymakers'; Franzese (2010) shows how one can leverage the distinct aspects of multiple policymakers – effective (common pool) vs raw numbers (veto actors) of parties, variance (common pool) vs range (veto actors) polarization of parties, and the ideological distribution of parties (bargaining compromise) – along with the different ways these different aspects of multiparty government affect policy outcomes (common pool: proportionate over/under-action; veto actor: adjustment-rate retardation; bargaining-compromise: convex combinations) to separately model, and so to separately identify and estimate, the veto-actor, common-pool, and bargaining-compromise effects of multiple policymakers.
  - 27 So-called 'dynamic' nonparametric-causal-effect estimates are either estimates of  $\beta$  from the static (2) in moving-windows of data, or estimates using static (2) of the period effects in (7) (iii), without the model, i.e., not estimates of the model and its parameters  $\rho$  and  $\beta$  separately, without which they are incapable of generating dynamic response-path or LRSS estimates.
  - 28 Indeed, we could expand here to note that all data, all outcomes of interest, occur in some (space and) time, and so these issues actually arise universally, ubiquitously in all applied empirical analysis, experimental or observational.
  - 29 The elements  $w_{ij}$  of the spatial-weights matrix,  $\mathbf{W}$ , give the relative connectivity from  $j$  to  $i$ , and  $\rho$  the strength of interdependence (contagion) operating in that predetermined pattern.
  - 30 Assuming  $\rho\mathbf{W}$  is the matrix equivalent of 'less than 1', such that  $\|\mathbf{I}-\rho\mathbf{W}\|\neq 0$  so the inverse spatial-multiplier exists. Note that the spatial multiplier derives from an infinite sum of the reverberating spatial feedback analogously to the temporal case (of forward-propagating-only 'feedback'):  $(\mathbf{I}-\rho\mathbf{W})^{-1} = \mathbf{I} + \rho\mathbf{W} + \rho^2\mathbf{W}^2 + \rho^3\mathbf{W}^3 + \dots + \rho^\infty\mathbf{W}^\infty$ .
  - 31 Franzese and Hays (2006) use a modified border-contiguity  $\mathbf{W}$  to define proximity in this application.
  - 32 These feedback reverberations are dampening provided  $\alpha_1\beta_1 < 1$ , so the system is not explosive (see also note 30).
  - 33 Semi- and flexible parametric designs offer a promising way forward for the (likely ubiquitous) combination of causal heterogeneity and causal simultaneity (see, e.g., Marra and Radice, 2011).
  - 34  $\Gamma$  in line (ii) has  $\mathbf{1}$  on its diagonal;  $\Gamma^*$  in line (iii) has  $\mathbf{0}$  on its diagonal and reverses sign of all off-diagonal elements from  $\Gamma$ .
  - 35 Also in this context, see methods specifically designed for complex or high-dimensional systems of endogenous dynamic equations, such as structural vector-autoregression (e.g., Kilian and Lütkepohl, 2017, for textbook exposition, and Pickup, Chapter 34 in this *Handbook*).
  - 36 In this regard, empirical-methodological practices in physics could serve as better exemplar for social science than biomedicine (see note 23). In physics, experimental statistics often yield not only tests of causal theoretical hypotheses but also estimates of the parameters in well-specified theoretical models, and it is the empirical-estimate-calibrated model rather than the experiment's test statistics that are used for causal-response estimates and prediction.
  - 37 The model assumes the United States influences but is not influenced by the other two actors.
  - 38 Furthermore, their empirical questions can be factual – 'who voted for Hitler?' – and so pertain to

defined, finite, and extant *populations* or theoretical – ‘what characteristics of voters and contexts contribute to right-wing populist support?’ – and so have populations of intended inference that are hypothetical and unlimited.

- 39 The analogy to Clarke and Primo’s (2012) declaration of *usefulness* as the aim of theoretical modeling is intentional and perfect.
- 40 Again, the analogy to Clarke and Primo’s (2012) declaration that theoretical models are to be *useful simplifications* is intentional and perfect.

## REFERENCES

- Best, N., Richardson, S., Thomson, A. 2005. ‘A Comparison of Bayesian Spatial Models for Disease Mapping’, *Statistical Methods in Medical Research* 14(1):35–59.
- Bowers, J., Fredrickson, M., Panagopoulos, C. 2013. ‘Reasoning about Interference between Units: A General Framework.’ *Political Analysis* 21(1):97–124.
- Brandt, P.T., Colaresi, M., Freeman, J.R. 2008. ‘The Dynamics of Reciprocity, Accountability, and Credibility’, *Journal of Conflict Resolution* 52(3):343–74.
- Caughey, D., Sekhon, J.S. 2011. ‘Elections and the Regression Discontinuity Design: Lessons from Close US House Races, 1942–2008’, *Political Analysis* 19(4):385–408.
- Clarke, K., Primo, D. 2012. *A Model Discipline: Political Science and the Logic of Representations*. Oxford: Oxford University Press.
- Coppedge, M., Alvarez, A., Maldonado, C. 2008. ‘Two Persistent Dimensions of Democracy: Contestation and Inclusiveness’, *The Journal of Politics* 70(3):632–47.
- Egami, N., Imai, K. 2015. *Causal Interaction in High Dimension*. Unpublished manuscript,, Princeton, NJ: Princeton University Press.
- Franzese, R.J. 1999. ‘Partially Independent Central Banks, Politically Responsive Governments, and Inflation’, *American Journal of Political Science* 43(3):681–706.
- Franzese, R.J. 2002. *Macroeconomic Policies of Developed Democracies*. Cambridge: Cambridge University Press.
- Franzese, R.J. 2003. ‘Multiple Hands on the Wheel: Empirically Modeling Partial Delegation and Shared Policy Control in the Open and Institutionalized Economy’, *Political Analysis* 11(4):445–74.
- Franzese, R.J. 2007. ‘Multicausality, Context-Conditionality, and Endogeneity’, in C. Boix & S.C. Stokes, eds, *The Oxford Handbook of Comparative Politics*, Oxford: Oxford University Press.
- Franzese, R.J. 2010. ‘The Multiple Effects of Multiple Policymakers: Veto Actors Bargaining in Common Pools’, *Rivista Italiana di Scienza Politica* 40(3):341–70.
- Franzese, R.J., Hays, J.C. 2006. ‘Strategic Interaction among EU Governments in Active Labor Market Policy-Making: Subsidiarity & Policy Coordination under the European Employment Strategy’, *European Union Politics* 7(2):167–89.
- Freedman, D., Pisani, R., Purves, R. 2007 [1978]. *Statistics*, 4<sup>th</sup> ed. New York: W.W. Norton.
- Gerber, A.S., Green, D.P., Shachar, R. 2003. ‘Voting May Be Habit-Forming: Evidence from a Randomized Field Experiment’, *American Journal of Political Science* 47(3):540–50.
- Hainmueller, J., Hopkins, D.J., Yamamoto, T. 2014. ‘Causal Inference in Conjoint Analysis: Understanding Multidimensional Choices via Stated-Preference Experiments’, *Political Analysis* 22(1):1–30.
- Hays, J.C., Ornstein, J.T., Franzese, R.J. 2019. *Estimating the Interest-Premium Cost of Left Government by Regression-Discontinuity Analysis of Close Elections*, Unpublished manuscript, St. Louis, MO: Washington University.
- Hendry, D.F. 1995. *Dynamic Econometrics*. Oxford University Press on Demand.
- Hershey, M.R. 2009. ‘What We Know About Voter-ID Laws, Registration, and Turnout’, *PS: Political Science & Politics* 42(1):87–91.
- Imai, K., Ratkovic, M. 2013. ‘Estimating Treatment Effect Heterogeneity in Randomized Program Evaluation’, *The Annals of Applied Statistics* 7(1):443–70.
- Imbens, G.W. 2010. ‘Better LATE Than Nothing’, *Journal of Economic Literature* 48(2):399–423.
- Jackson, J.E. 2008. ‘Endogeneity and Structural Equation Estimation in Political Science’, in *The Oxford Handbook of Political Methodology*, Oxford: Oxford University Press.

- Kellstedt, P.M., Whitten, G.D. 2018. *The Fundamentals of Political Science Research*. Cambridge: Cambridge University Press.
- Kilian, L., Lütkepohl, H. 2017. *Structural Vector Autoregressive Analysis*. Cambridge: Cambridge University Press.
- Marra, G., Radice, R. 2011. 'A Flexible Instrumental Variable Approach', *Statistical Modelling* 11(6):581–603.
- Pagan, A., Ullah, A. 1999. *Nonparametric Econometrics*. Cambridge: Cambridge University Press.
- Pearl, J. 1995. 'Causal Diagrams for Empirical Research', *Biometrika* 82(4):669–88.
- Poole, K. 2019. *Spatial Models of Parliamentary Voting*. Cambridge: Cambridge University Press.
- Pritchett, L., Sandefur, J. 2015. 'Learning from Experiments When Context Matters', *American Economic Review* 105(5): 471–5.
- Schrodt, P.A., Gerner, D.J. 2000. 'Cluster-Based Early Warning Indicators for Political Change in the Contemporary Levant', *American Political Science Review* 94(4):803–17.
- Selway, J. 2011. 'The Measurement of Cross-cutting Cleavages and Other Multidimensional Cleavage Structures', *Political Analysis* 19(1):48–65.
- Sims, C.A. 2010. 'But Economics Is Not an Experimental Science', *Journal of Economic Perspectives* 24(2):59–68.
- Treier, S., Jackman, S. 2008. 'Democracy as a Latent Variable', *American Journal of Political Science* 52(1):201–17.
- Ward, M.D. 2016. 'Can We Predict Politics? Toward What End?', *Journal of Global Security Studies* 1(1):80–91.
- Woodward, J. 2016. 'Causation and Manipulability', in E.N. Zalta, ed., *Stanford Encyclopedia of Philosophy* (Winter). Available at <https://plato.stanford.edu/archives/win2016/entries/causation-mani/> (Accessed on 31 January 2020).

# A Principled Approach to Time Series Analysis

Suzanna Linn and Clayton Webb

Everything is dynamic. Mass attitudes, public policy outcomes, state budgets, and international conflicts operate over, or in, time. The applied time series analyst's goal is to answer questions about dynamic relationships through a process of theory development, model specification, hypothesis testing, and inference. However, political reality is complex, and our theories are rarely rich enough to dictate specific dynamic specifications. Even when theory is rich, model estimation is constrained by the available data.

The challenges posed by dynamic modeling have fostered a number of misguided approaches to time series analysis. One approach treats the dynamic features of the data as a nuisance. Analysts often include lags of the regressand to 'control for' serial correlation or use robust standard errors to purge the dynamic features of the data from the residuals. These approaches treat serial correlation as a disease rather than a symptom of dynamic misspecification.

Another approach posits dynamic specification as a theoretical problem. Proponents of this perspective profess a priori knowledge of the complex details by which mostly continuous data generating processes (DGPs) are sampled and aggregated to produce observed data – but this theoretical 'knowledge' often amounts to little more than folk-logics and pseudo-theories that have become common rationalizations for particular dynamic specifications. Some will include  $y_{t-1}$  because 'investment last year should determine investment this year' or omit  $y_{t-1}$  because 'lagged dependent variables wash out the effects of the other regressors.' Others will include  $x_{t-1}$  'to remove the potential for endogeneity' and will only include one lag because 'one lag is usually enough.' Some of the stories are more elaborate, even referencing the sampling interval and level of temporal aggregation, but these are not theories about the DGP.

These approaches are problematic. A single lag of the regressand suggested by 'theory' may be sufficient with some annual data



but may be insufficient for the same variables sampled at quarterly or monthly intervals. Robust estimation procedures remove serial correlation from the residuals, but estimates will still be biased if the serial correlation was a consequence of dynamic misspecification (King and Roberts, 2015). Analysts may be able to rely on their idiosyncratic approaches in some settings but are likely to face problems if they try to generalize these strategies.

Most political phenomena are dynamic: they change over time. How these dynamics manifest in an empirical model depends on measurement. Time series data are defined by specific sampling windows, sampling intervals, and levels of spatial and temporal aggregation. The empirical model for a political relationship measured at the annual level will have a specific dynamic structure and distinct dynamic features from an empirical model for the same relationship measured at the quarterly or monthly level. These models may change if the data are sampled at different periods of time. This suggests that the following principle should guide time series analysis: dynamic specification is an empirical problem. Theory informs the variables that belong in the empirical model, the data inform the dynamic structure of the empirical model. Our empirical models must be compatible with the features of our data.

Our goal in this chapter is to outline a principled approach for model building and statistical analysis of time series data. We propose a general approach to dynamic specification that incorporates features from several popular schools of econometric practice (Pagan, 1987; Granger, 1990; Spanos, 1990; Hendry, 1995; Campos et al., 2005). The analyst begins with a plausible general model (PGM). The PGM is a tentative empirical specification that encompasses the relationships proposed by the theoretical model. It draws on relevant information from prior theory and evidence, our knowledge of the data and its measurement, and our research goals. The analyst conducts diagnostic tests to ensure the PGM is dynamically complete.

If the PGM has good statistical properties, the analyst iteratively tests downward to produce a parsimonious representation of the relationship that is also dynamically complete. The selected empirical model is then estimated, and the results are used for a variety of goals including hypothesis testing, description, policy analysis, and forecasting. While this process will not resolve all the challenges inherent to dynamic empirical modeling, it will produce empirical models capable of generating reliable inferences.

The chapter is organized as follows: we begin by outlining our approach to dynamic specification; we explain how the analyst specifies the PGM, tests the statistical adequacy of the PGM, and simplifies the model; we briefly discuss the different strategies analysts can use to estimate dynamic models before highlighting the diverse set of tools that are available for interpretation and inference for time series analysis. Throughout, we illustrate our approach with reference to general questions about presidential approval. We close by discussing additional challenges analysts are likely to confront when building dynamic models and offering concluding comments that fit time series analysis into the broader social-scientific enterprise.

## **A PRINCIPLED APPROACH TO DYNAMIC SPECIFICATION**

Our empirical models must be compatible with the features of our data. If data are sampled at regular intervals over time, the empirical model must be built to capture the dynamic features of the data. With this in mind, time series analysis proceeds in four stages. The analyst (1) develops a PGM, (2) uses diagnostic tests to verify the PGM is dynamically complete, (3) simplifies the PGM to arrive at a parsimonious model that makes the most efficient use of the information in the data, and (4) draws inferences from the model. If analysts follow this

procedure, they can be confident in their inferences and conclusions. In this section we discuss the first three steps of this process.

### ***How to Specify a General Model***

The PGM is an empirical model that is *plausible* because the primary feature of the model is that it is reasonable given the information available to the analyst and *general* because it is not restricted in any way that is not explicitly determined by theory. The analyst has two types of information to use in the specification process. There is information the analyst brings with them: their goals, their theoretical arguments – as well as complementary and competing theoretical arguments – and previous findings. There is also information in the data: the details about the measurement and the observed features of the data. Some varieties of information are more useful for some decisions than others and there is no direct mapping from information to the PGM. Ultimately, the specification of the PGM will result from an interplay of theory and data (Pagan, 1987).

Analysts must make a variety of choices when identifying the PGM. They must: (1) select the variables to include in the model; (2) choose whether to use a single equation or multivariate model; and (3) identify the model. Identification of the model involves the selection of a functional form, a parameterization, and a general lag specification that encompasses all plausible dynamic relationships that exist in the data.

Among the most influential pieces of information available to analysts is their motivation for the analysis. Generally, our goals lie in describing linkages and testing hypotheses derived from theory, but we may also wish to forecast behavior. Our specific interest can influence whether the PGM is a single or multiple equation model, as well as the variables in the model.

Theory determines the variables that enter the PGM. By variables we mean constructs in the theory that must be measured and included in the statistical model – concepts like trust, optimism, or presidential approval. Analysts should include all the variables necessary to test their theories and account for competing theories and previous findings. If we are to believe that one theory explains the finding of another, both theories must be represented in the empirical model.

Data determine the structure of the empirical relationships in the PGM. Three features of measurement are particularly important: the sampling interval of the data, the levels of spatial and temporal aggregation, and the temporal window of the series. Highly aggregated data mask causal relationships and higher frequency data tend to exhibit more complex dynamic properties (Freeman, 1989). One example is seasonality. Patterns of time dependence often vary predictably over the calendar year.<sup>1</sup> Monthly data may exhibit this type of dependence every twelfth period. Consider a monthly time series of automobile accidents, accidents tend to increase in the winter months when roads are icy. The PGM should be sufficiently general to capture plausible patterns given the periodicity of the data. A PGM for monthly data should have at least 12 lags. A PGM for quarterly data should have at least four lags. Analysts should consider multivariate PGMs in cases where the data are sampled or aggregated at levels that could induce simultaneity. The interaction between the sampling window and the sampling interval is particularly important. The series sampled over 100 weeks will contain different information than the same variable sampled over 100 years.

In addition to information about measurement, analysts can use empirical information about the individual time series to guide specification of the PGM. Time series plots, unit root tests, tests for structural breaks, and period-to-period correlations can highlight dominant features in the data and identify dynamic patterns that need to be accounted

for in the PGM. If the data are stationary – that is, the data have stable long run means and variances – we suggest that the analyst choose a general lag structure for the PGM that, at least, encompasses the highest order of autocorrelation observed in the PACF for the dependent variable and that it be sufficiently large to capture potential seasonal patterns in the variable.<sup>2</sup> If the data contain unit roots – that is, the means or variances of the data are unstable – the analyst should test for cointegration and, in the absence of cointegration, may need to difference the data.

We use two examples to illustrate how available information can inform specification of the PGM. First, consider an analyst interested in the relationship between economic performance and presidential approval. Drawing on previous theory and evidence, our analyst expects that both subjective evaluations and objective economic outcomes shape evaluations of the president and that controls must be included to account for dramatic events, presidential administrations, and the like (MacKuen et al., 1992; Erikson et al., 2002; De Boef and Kellstedt, 2004). The analyst decides to use monthly inflation ( $I$ ), unemployment ( $U$ ), and the Michigan Index of Consumer Sentiment ( $CS$ ) to model presidential approval ( $App$ ). The PGM might be given by the following:

$$App_t = \alpha_0 + \sum_{i=0}^p \alpha_i App_{t-i} + \sum_{i=0}^q \beta_i^{CS} CS_{t-12} + \sum_{i=0}^q \beta_i^U U_{t-12} + \sum_{i=0}^q \beta_i^I I_{t-12} + BX_t + \varepsilon_t$$

Where  $X$  is a series of intervention variables included to capture differences between presidencies and salient events and  $B$  is a vector of parameters that describe those effects. The PGM includes 12 lags to capture any periodicity or seasonality that may exist in the data.

A second analyst might argue that the weak exogeneity assumption of the first

analyst is overly restrictive because evaluations of the president influence economic perceptions. In this scenario, the analyst might begin with a monthly vector autoregression (VAR) PGM. Given the large number of parameters that must be estimated in a VAR, the analyst adopts a single measure of economic performance ( $E$ ): the Conference Board's Lagging Economic Indicator Index. This VAR PGM could be written as:

$$\begin{aligned} App_t &= \alpha_0 + \sum_{i=0}^p \alpha_i App_{t-i} + \sum_{i=0}^q \beta_i^{CS} CS_{t-12} \\ &\quad + \sum_{i=0}^q \beta_i^E E_{t-12} + BX_t + \varepsilon_{1t} \\ CS_t &= \gamma_0 + \sum_{i=0}^p \gamma_i CS_{t-i} + \sum_{i=0}^q \theta_i^A App_{t-12} \\ &\quad + \sum_{i=0}^q \theta_i^E E_{t-12} + \Theta X_t + \varepsilon_{2t} \\ E_t &= \psi_0 + \sum_{i=0}^p \psi_i E_{t-i} + \sum_{i=0}^q \phi_i^A App_{t-12} \\ &\quad + \sum_{i=0}^q \phi_i^{CS} CS_{t-12} + \Phi X_t + \varepsilon_{3t} \end{aligned}$$

where the different parameters, variables, and errors are given for each equation or more compactly as:

$$Y_t = \Gamma_0 + \Gamma Z_t + \varepsilon_t$$

where  $Y_t$  is a matrix containing each of the endogenous variables at  $t$ ,  $\Gamma_0$  is a vector of intercepts,  $Z_t$  is a matrix containing all the other lags and contemporaneous values of the variables,  $\Gamma$  is a matrix of parameters that describe or constrain these effects, and  $\varepsilon_t$  contains the random errors from each equation.<sup>3</sup>

Once an analyst has specified a PGM, they must conduct an initial round of tests to determine whether the model is statistically adequate given the data. An adequate PGM is a model that is dynamically complete, it includes all the information necessary to

accurately characterize the relationships of interest. A dynamically complete PGM will have white noise residuals. There are several tests for this null hypothesis (e.g., Ljung-Box and Breusch-Godfrey). Analysts should also consider tests for parameter constancy and heteroskedastic errors; we will discuss these issues in more detail below.

Diagnostic tests (not reported) suggest that the single equation PGM and the VAR PGM are sufficiently general. While we could stop here and use the PGMs for estimation and inference, the general models are not well suited for inference because they are overparameterized. In the next section we explain the rationale for, and the process of, simplifying the PGMs to arrive at specifications that are dynamically complete and parsimonious.

### ***How to Simplify the Models***

An acceptable PGM incorporates a sufficient number of variables, including their lags, to encapsulate every plausible dynamic specification. There may, however, be more parsimonious forms of the model that are also dynamically complete. With other things equal, analysts should prefer parsimonious models. Multicollinearity in most PGMs will undermine the efficiency of estimates and compromise hypothesis tests.

How does the analyst select a simplified model? The analyst moves from a PGM to a parsimonious model by iteratively testing both the validity of successive model restrictions and the dynamic completeness of the restricted model. If the restrictions are reasonable and the resulting model is dynamically complete, the model is acceptable. If either is not true, the restricted model should be discarded. The iterative testing procedure allows the analyst to identify a set of dynamically complete models. Measures of model fit can be used to select the model for analysis and interpretation from this restricted set.

There are a number of ways the analyst can simplify the PGM. The most obvious tools are *t*-tests and *F*-tests. Some might criticize this practice as data mining. This criticism is misguided. Data mining refers to an unstructured search of a data set. A general-to-specific simplification search is structured by theory and data through the specification of the PGM (Granger, 1990: 10). The simplification search is not ‘p-hacking’ because the tests of interest in each step are not the *t* or *F* statistics that will eventually be used to test theoretically relevant hypotheses. Instead, the analyst is interested in the results of the diagnostic checks for dynamic completeness conducted after each restriction.

Where should the analyst begin? There are a number of potential starting points. No strategy is best, but some are particularly bad. In lieu of a specific set of instructions, we offer a set of admonitions and suggestions: (1) do not test too many restrictions, iteratively or jointly, on highly collinear regressors; (2) do not remove all lags of any variable from the model: ‘A good model should not only explain the data, but it should also explain both the successes and failures of rival models’ (Gilbert, 1990: 289); (3) do not be overly reliant on measures of model fit, fit statistics tell us nothing about the statistical adequacy of the model; (4) consider multiple starting points;<sup>4</sup> and (5) allow variables to have unique lag structures, including different onset effects.

We demonstrate the general-to-specific simplification search in the context of the PGMs outlined in the last section. Table 32.1 presents a subset of models estimated based on simplifications of Equation 1. We present the *p*-values for the Breusch-Godfrey (BG) serial correlation tests. We include results for 12 and 24 lags for each model. We present three measures of model fit: Akaike’s information criterion (AIC), the Bayesian information criterion (BIC), and  $\bar{R}^2$ .

Our strategy for paring down the single equation PGM began with block *F*-tests. We began with 12 lags of each variable and restricted

**Table 32.1 Model simplification in the single equation model of approval**

	<i>ADL</i>	<i>PA</i>	<i>DS</i>	<i>FDL</i>	<i>Mixed 1</i>	<i>Mixed 2</i>	<i>ECM</i>
Intercept	-2.74 (3.63)	-4.29 (3.40)	1.00 (3.43)	-17.68* (8.70)	-2.80 (3.61)	-3.40 (2.27)	-3.40 (2.27)
Approval <sub>t-1</sub>	0.86** (0.02)	0.86** (0.02)	0.87** (0.02)		0.86** (0.02)	0.86** (0.02)	-0.14** (0.02)
ICS <sub>t</sub>	0.16 (0.04)**	0.10** (0.26)		0.37** (0.10)	0.16** (0.04)	0.16** (0.04)	
ICS <sub>t-1</sub>	-0.08+ (0.04)		0.05+ (0.03)	0.27* (0.10)	-0.08+ (0.04)	-0.07+ (0.04)	0.09** (0.02)
Unemployment <sub>t</sub>	0.84 (1.04)	0.04 (0.17)		10.50** (2.46)	0.81 (1.03)		
Unemployment <sub>t-1</sub>	-0.88 (1.06)		-0.17 (0.18)	-9.00** (2.38)	-0.85 (1.00)		
Inflation <sub>t</sub>	0.29 (0.42)	0.24* (0.12)		0.54 (1.00)	0.22+ (0.12)	0.23+ (0.12)	0.23+ (0.12)
Inflation <sub>t-1</sub>	-0.08 (0.41)		0.13 (0.12)	-0.22 (0.98)			
ΔICS <sub>t</sub>							0.16** (0.04)
ΔUnemployment <sub>t</sub>						0.87 (0.99)	0.87 (0.99)
ICS LRM	0.60** (0.19)	0.68** (0.17)	0.39** (0.18)	0.64** (0.06)	0.61** (0.18)	0.63** (0.13)	0.63** (0.13)
Unemployment LRM	-0.29 (1.31)	0.31 (1.22)	-1.26 (1.37)	1.49 (1.31)			
Inflation LRM	1.53+ (0.89)	1.69+ (0.88)	0.98 (0.92)	0.32 (0.30)	1.55+ (0.89)	1.59+ (0.86)	1.59+ (0.86)
$\bar{R}^2$	0.9124	0.9122	0.9102	0.4934	0.9126	0.9128	0.3137
AIC	2544.72	2543.03	2553.99	3388.11	2542.75	2540.80	2540.80
BIC	2644.94	2630.72	2641.68	3484.15	2638.80	2632.67	2632.67
Breusch-Godfrey(24)	0.144	0.173	0.172	0.000	0.145	0.147	0.147
Breusch-Godfrey(12)	0.789	0.836	0.879	0.000	0.791	0.783	0.783

Note: \*\*  $p < 0.01$ , \*  $p < 0.05$ , +  $p < 0.10$ .

the model to include 11 lags of each variable, 10 lags, 9 lags, and so on; instead of, removing one set of lags at a time. If a restriction imposed no loss of information, we tested for serial correlation. The restriction was accepted as long as the model remained dynamically complete. In the final iteration of this initial testing process, the block  $F$ -test rejected the null associated with the static specification in favor of the ADL(1,1) model. This is the first

model presented in Table 32.1. The BG tests suggest this specification is dynamically complete. From here, we could have taken many paths, choosing to either test restrictions on this simplified model or testing alternative sets of restrictions on the PGM. We chose the former strategy. From the ADL(1,1), the partial adjustment (PA) model restricts all the lags of all the regressors to zero, the dead start (DS) model restricts the contemporaneous values

of all the regressors to zero, and the finite distributed lag (FDL) includes the contemporaneous and lagged values of the regressors but omits the lag of approval. These models are presented in columns 3, 4, and 5. The PA and DS model are dynamically complete but the FDL model is not.

We also considered more nuanced lag specifications. The ADL, PA, DS, and FDL models impose common dynamic restrictions on the regressors. The Mixed 1 and Mixed 2 models allow the restrictions to vary among the independent variables. Comparing the ADL and PA models, the coefficient on inflation is significant when lagged inflation is omitted from the model. Mixed 1 removed lagged inflation. The coefficients for contemporaneous and lagged unemployment have similar magnitudes but opposite signs in the ADL. This is consistent with changes in unemployment included in Mixed 2. The results of the iterative simplification process would be the same if one began with an ECM parameterization rather than an ADL. This is reflected in the similarity of the results from the Mixed 2 model and the ECM representation of that model presented in column 8.

To select a final model, we balance the tradeoffs of fit (achieved in the PGM) and

parsimony. The  $\bar{R}^2$  and information criteria capture this trade-off by reporting penalized fit statistics. The penalty is based on the number of regressors in the model. In our example, the AIC and  $\bar{R}^2$  select the Mixed 2 specification in column 7 (or equivalent ECM in column 8). The BIC selects the PA model in column 2.<sup>5</sup> With the exception of the FDL, any of the models presented in Table 32.2 could be used for hypothesis testing.

The VAR is simplified using similar logic but the simplification strategies we entertained are more straightforward in the context of VAR.<sup>6</sup> We successively tested restrictions on the order of the lag length in the VAR using likelihood ratio tests. In each step, we also confirmed that the system was stable by estimating the largest root in the system and dynamically complete by testing for serial correlation in the residuals.

We selected the VAR with four lags. The likelihood ratio test for the VARs including 11 and 10 lags suggest a (nearly) significant loss of information. All other successive restrictions, up to four lags, resulted in no other significant loss of information. At the same time, the Portmanteau (Ljung-Box) tests (12 lags) for each VAR (p) provide no evidence of serial correlation in the residuals

**Table 32.2 Model simplification in the VAR of presidential approval, economy, and consumer sentiment<sup>a</sup>**

<i>Model lags</i>	<i>Portmanteau (12 lags, p-value)</i>	<i>Stable root</i>	<i>Likelihood ratio test (p-value)</i>	<i>AIC</i>	<i>BIC</i>
12	0.167	Yes	0.998	5120.48	5727.94
11	0.371	Yes	0.633	5118.22*	5689.57
10	0.419	Yes	0.052	5122.62	5657.81
9	0.268	Yes	0.498	5131.15	5630.13
8	0.375	Yes	0.875	5129.63	5592.37
7	0.464	Yes	0.150	5128.58	5555.03
6	0.407	Yes	0.247	5136.01	5526.13
5	0.326	Yes	0.371	5141.73	5495.48
4	0.346	Yes	0.013	5144.54	5461.87
3	0.192	Yes	0.420	5167.22	5448.10
2	0.205	Yes	0.000	5171.80	5416.18
1	0.016	Yes	0.007	5198.01	5405.85*

until only one lag was included in the system. The roots of the system were less than one in all cases. Notice that the information criteria select vastly different models and that the BIC selects a model that is not dynamically complete. This latter problem illustrates the danger of making specification choices based on fit alone.

Having settled on simplified versions of our exemplar PGMs, we are nearly ready to turn to the problems of inference and interpretation. Before doing so, we briefly discuss issues in estimation.

## APPROACHES TO ESTIMATION

The ‘dynamic’ in dynamic model refers to the specification of the relationships among the variables over time, it does not suggest a particular estimation strategy. The three primary methods of estimation are ordinary least squares (OLS), maximum likelihood estimation (MLE), and Bayesian estimation methods. Each approach has unique benefits and limitations. Which method is most appropriate often depends on the preferences and goals of the analyst.

Dynamic linear models are often estimated using OLS. The primary benefit of OLS is simplicity. A large number of linear models can be estimated using OLS in a short period of time. Advances in computing have narrowed the gap between least squares and the other methods, but these efficiency gains can be non trivial for large or complex models. Optimization and sampling algorithms can be more time consuming and alternative estimation procedures may require additional diagnostic tests.

The primary limitations of OLS are the strong assumptions OLS makes about the error process. OLS residuals must be white noise. Bayesian and likelihood-based models make it easier to accommodate more complex error structures like the autoregressive, integrated, moving average (ARIMA)

models popularized by Box and Jenkins (1976). Bayesian and likelihood-based methods are also necessary to estimate models of the conditional variance of the series – a process we will discuss in more detail below.

There are a handful of additional advantages offered by Bayesian estimation. Bayesian priors are simultaneously a cost and benefit of Bayesian dynamic modeling. On the one hand, Bayesian priors offer analysts a means of incorporating previous knowledge into the modeling process. Limiting the parameter space using prior information may make estimation more efficient in some circumstances and may facilitate analyses that would otherwise be intractable. For example, Brandt and Sandler (2012) develop a Bayesian Poisson VAR model for multivariate count models and Brandt et al. (2008) use Bayesian Structural VAR models to compare different recursive structures suggested by competing theories about public opinion and international conflict. On the other hand, the iterative procedure we outline in the last section often requires a large number of models to be estimated. Specifying the priors for all the parameters in the PGM can be tedious and time consuming.

Most models can be estimated using OLS, MLE, or Bayesian estimation methods. Unless the analyst has a specific rationale for choosing one method over another, it is usually best to choose the method that is most convenient and best suits the analyst’s needs. As long as the analyst is able to recover estimates from the parameters, how the analyst arrived at the estimates is only of secondary importance. In the next section, we describe how these estimates can be used for inference and interpretation.

## INFERENCE AND INTERPRETATION

Static and dynamic regression models can be estimated using the same procedures, but there are important differences in

interpretation. The dynamic modeler has a more robust arsenal of interpretational tools that can be used to understand dynamic relationships. In this section we discuss these tools and highlight them in the context of our examples.

The most dramatic change in interpretation as one moves from a static model to a dynamic model pertains to the interpretation of the regression coefficients. A regression coefficient in a static OLS regression gives the average change in the regressand associated with a one unit increase in the regressor. In a dynamic model, the coefficient, or impact multiplier, is interpreted as the instantaneous change in the regressand. If there are  $q$  lags of a regressor, the effect of the shock is distributed over  $q$  periods of time. If a lag of the dependent variable is included in the model, the effects of shocks propagate over time. The full effect of a variable in a dynamic regression is represented by the long run multiplier (LRM). The LRM is equal to the sum of the impact multipliers divided by one minus the sum of the coefficients on the lagged values of the dependent variable

$$\sum_{i=0}^q \hat{\beta}_q / \left( 1 - \sum_{i=1}^p \hat{\alpha}_p \right).$$

De Boef and Keele (2008) describe how the formula can be modified to accommodate parameterizations from the GECM.

The impulse response functions (IRFs) and cumulative impulse response functions (CIRFs) can be used to illustrate how shocks to the regressors propagate into the regressand over time. The first value of the IRF is the impact multiplier. It reflects the instantaneous change in the outcome associated with a shock. The next  $t+1, t+2, \dots, t+p$  steps of the IRF depict the rate at which the shock dissipates from the series over time. The effect of a shock dissipates at a rate that is inversely proportional to  $\sum_{i=1}^p \hat{\beta}_p$ . The CIRF for a model shows how the effect of the shock accumulates over time. The first value in a CIRF is also the impact multiplier but the next  $t+1,$

$t+2, \dots, t+p$  steps of the CIRF show how the effect increases from this initial level to the final value given by the LRM. One depicts these patterns by plotting the values over time.

The median and mean lag lengths are useful quantities. The median lag length is the first lag at which at least half of the shock has dissipated from the series, or at which half of the adjustment back toward the long-run equilibrium series has occurred (De Boef and Keele, 2008). The mean lag length is the amount of time it takes for the series to adjust back to its equilibrium, or the amount of time it takes for a shock to play out. The mean and median lag lengths are calculated by dividing the IRFs for a variable by the LRM for that variable and summing those standardized IRFs until one reaches 50% of the LRM (median lag) and 99% of the LRM (mean lag).

We illustrate the utility of these interpretational tools in the context of the presidential approval models presented in Table 32.1. Two points bear repeating. First, one should only interpret results from models that are dynamically complete. One should not interpret the results from the FDL model in Table 32.1. Second, one can use fit criteria to arbitrate among dynamically complete models. Both the PA and Mixed 2 specifications can be used for interpretation and inference. These specifications produce similar inferences because the models contain the same information. A notable exception is the DS model. The fit for the DS model is substantially worse than the fit for the PA and Mixed 2 models. This reflects that there is some amount of information loss with the DS model.

The results presented in Table 32.1 speak to the question raised by the first analyst. Consumer sentiment has a positive and statistically significant effect on presidential approval. The impact multipliers for consumer sentiment sum to 0.9 ( $0.9 = 0.16 - 0.7$ ). The LRM for consumer sentiment is 0.63. These quantities reflect the average response in approval to a one-unit shock in sentiment. While this is illuminating, the



effect of a one-unit shock may not be substantively interesting to our analyst. The analyst could calculate the expected effect of a one standard deviation change in consumer sentiment. A standard deviation change in consumer sentiment is 13 points. In the long run, a one standard deviation change in consumer sentiment is expected to increase presidential approval by eight points. The analyst can learn even more about the dynamic features of the relationship by looking at the IRF and CIRF.

Figure 32.1 shows the IRF and CIRF for a one-unit change in consumer sentiment from the Mixed 2 model in Table 32.1. The left panel of Figure 32.1 shows the IRF. The right panel of Figure 32.1 shows the CIRF. The first bar in the IRF plot corresponds to the coefficient on  $ICS_t$  from the mixed model. The 0 denotes that the positive 0.16 shift in approval occurs in the period where the shock occurred. There is a dramatic drop in the next period because the coefficient for  $ICS_{t-1}$  is negative. After these initial periods, the effects of the change in consumer sentiment and lagged consumer sentiment continue to propagate into approval over the next 24 periods. The cumulative effect

is depicted in the right panel. The first bar in the CIRF also corresponds to the coefficient on  $ICS_t$  but the value of the CIRF in the next period is greater than the value in the first period. The second value in the CIRF is the accumulated effect over the first two periods. The rate at which these effects accumulate is the same as the rate at which the shocks dissipate in the IRF. We can summarize the information contained in the impulse response functions using the median and mean lag lengths.

The IRF and CIRF are based on the Mixed 2 regression from Table 32.1:

$$App_t = \alpha_0 + \alpha_1 App_{t-1} + \beta_0^{CS} CS_t + \beta_1^{CS} CS_{t-1} + I_t + \Delta U_t + \epsilon_t$$

The responses are based on a one unit increase in consumer sentiment. The responses are plotted for the first 30 periods following the change.

The median lag for consumer sentiment is 4 and the mean lag for consumer sentiment is 25. The short median lag and long mean lag suggest that approval error corrects quickly after a shock to consumer sentiment, but that it takes a long time for the full effect of

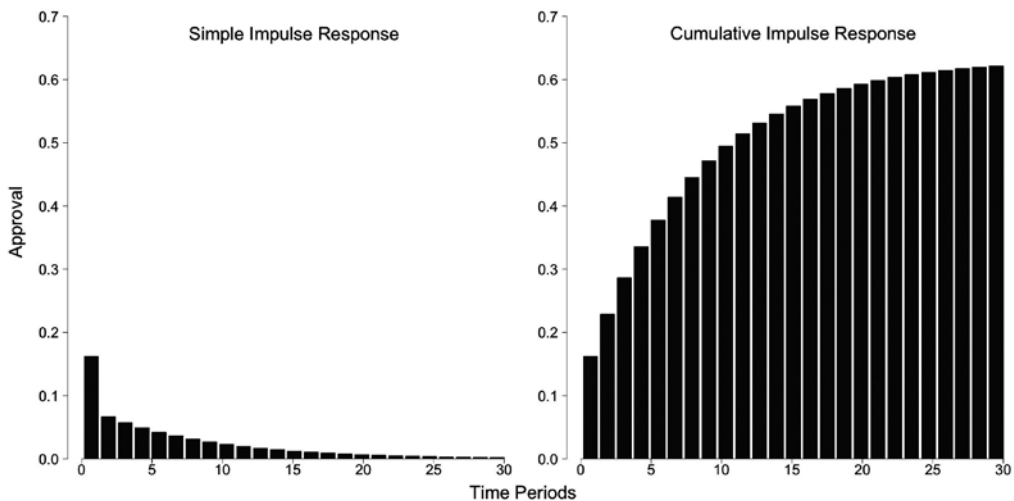


Figure 32.1 IRF and CIRF for the effect of consumer sentiment on presidential approval

that shock to be realized. People's economic perceptions change their perceptions of the president quickly, but these attitudes are sticky. These features of the dynamic relationship are born out in the plots presented in Figure 32.1.

The interpretation of multivariate time series models relies on many of the same tools but there are notable differences. The individual regression coefficients and the associated  $t$ -statistics are not useful for inference in most circumstances because multicollinearity among the variables and their lags inflate the standard errors for the individual coefficients. This collinearity does not affect the performance of  $F$ -tests. Block  $F$ -tests on the groups of lags associated with different variables in the component regressions of a VAR can be used for inference. If a block of lags of one variable enters significantly into the component equation for another, that first variable is said to 'Granger cause' the other variable (Enders, 2015: 305).<sup>7</sup> These Granger causality tests are the most basic inferential tool available in VAR analysis. In the four lag VAR including economic performance, consumer sentiment, and presidential approval; consumer sentiment ( $p = 0.035$ ) and approval ( $p = 0.008$ ) Granger cause economic performance but none of the other relationships in the system reach conventional levels of significance. While Granger causality tests are a useful tool, these tests are limited because they do not reflect the direction and strength of the relationships among the variables.

The primary tools for the interpretation and inference of VAR models are vector moving average (VMA) impulse response functions and forecast error variance decompositions (FEVD). Collectively, these tools are referred to as 'innovation accounting' (Box-Steffensmeier et al., 2014: 113). A VAR process can be rewritten as a VMA process where the variables in the system are expressed in terms of the current and past values of the errors. This representation allows one to trace the time path of shocks

to the system using impulse response functions similar to the IRFs calculated for single equation models. Confidence intervals can be calculated for the IRFs in a number of ways (Brandt and Freeman, 2006). A response is interpreted as statistically significant as long as the interval does not include zero.

Figure 32.2 presents the VMA IRFs for the four lag VAR model of lagging economic performance (Economy), consumer sentiment (ICS), and presidential approval (Approval). The IRFs depict the responses of the system to one standard error of regression shocks in the endogenous variables. The column headings indicate the variables being shocked. The row headings indicate the response variables. The diagonal plots demonstrate the rates of decay in each variable following the shock to itself. The solid black lines are the impulse responses and the dotted black lines are the 90% confidence intervals for the shocks.

A set of restrictions must be placed on the contemporaneous relationships among the variables to interpret the results. The IRFs are sensitive to these restrictions. The default set of restrictions used in most software packages is the Choleski decomposition that produces the triangular pattern observed in Figure 32.2. The instantaneous responses in the functions plotted below the diagonal are restricted to zero. These restrictions only affect the initial values of the IRFs. There may be theoretical reasons to prefer one set of restrictions over another. Brandt and Freeman (2009) describe strategies analysts can use to test competing restrictions using Bayesian Structural VAR models.

The results in Figure 32.2 offer a comprehensive depiction of the relationships among the variables. The plots in the second and third rows of the array depict the relationships between presidential approval and consumer sentiment. The shock to consumer sentiment produces a positive and statistically significant response in presidential approval (row 3, column 2). Like the results from the single equation model, the effect is moderate but persistent. A shock to presidential approval

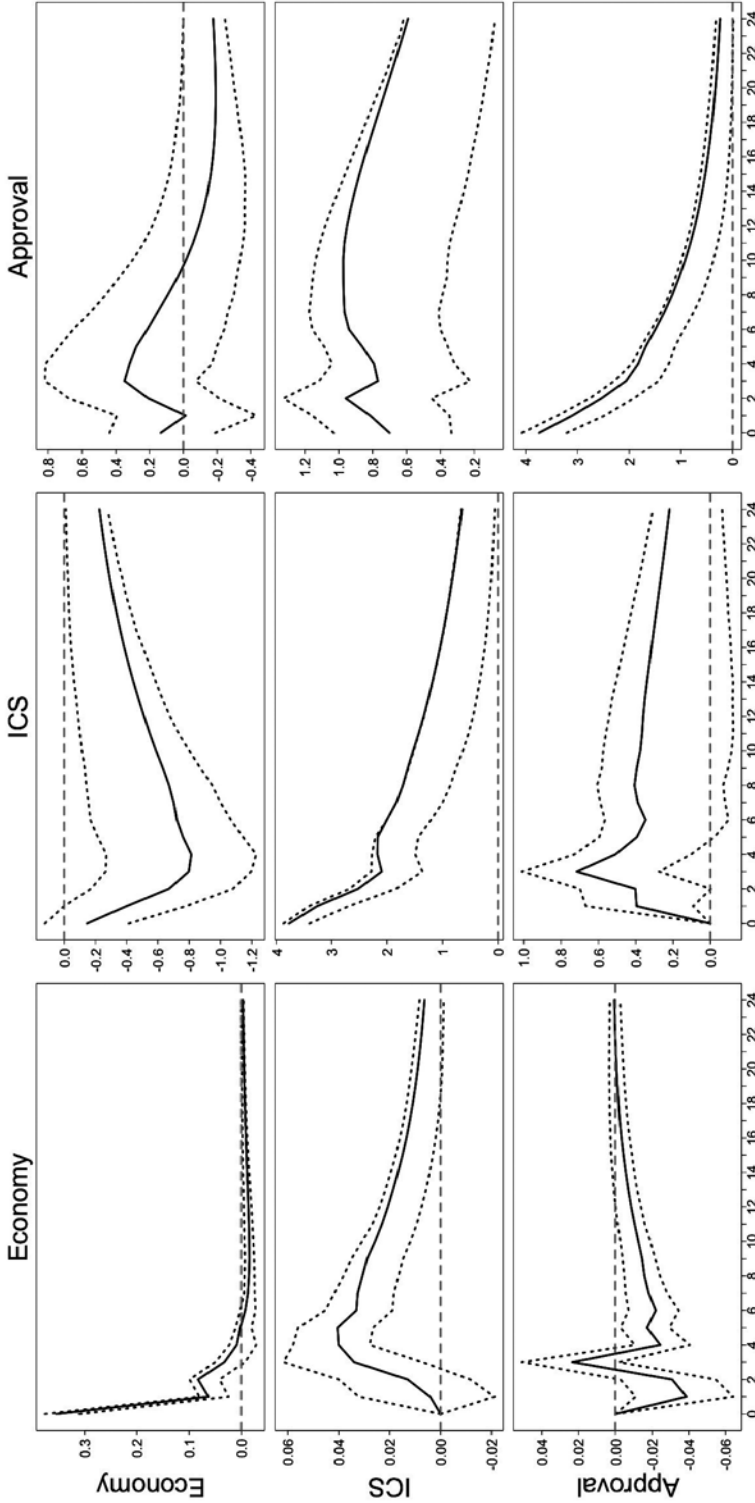


Figure 32.2 The IRFs for three variable VAR model of economic performance, consumer sentiment and, presidential approval

also produces a positive and statistically significant response in consumer sentiment (row 2, column 3). This suggests that public perceptions of the president influence public perceptions of the economy. This effect is more persistent, the response is still reliably different from zero two years after the initial shock.

The array of IRFs allows one to identify indirect relationships among the variables. Consistent with expectations, improved economic performance increases consumer sentiment (row 2, column 1). The shock in economic performance produces an erratic response in approval (row 3, column 1), but the economic performance has a positive, indirect effect on approval through consumer sentiment.

FEVDs offer another tool analysts can use to understand the relationships among the variables in the system. The forecast error variance can be calculated for the  $t$ -step ahead forecasts from a VAR model and the forecast error variance can be calculated as a function of these estimates. The variance of these forecasts can then be categorized in terms of the proportion of the variation caused by each variable. Even at long time-horizons, most variables explain a majority of their 'own' forecast error variance, but the proportions of the variance explained by other variables offer analysts clues about the importance of specific relationships. In the case of the lagging indicators series, presidential approval and consumer sentiment jointly explain less than 10% of the variation in economic performance after 12 periods and only 11% after 24 periods.<sup>8</sup> This is consistent with the small and insignificant responses in the IRFs. There does appear to be an important relationship between consumer sentiment and approval. Consumer sentiment explains more than 25% of the variation in presidential approval after 24 periods, compared with the 1% explained by lagging indicators. Combined with IRFs presented in Figure 32.2, this evidence suggests that consumer sentiment is more important

for understanding changes in presidential approval than changes in objective economic indicators.

Both our analysts have a lot of information they can use to assess their expectations. The dynamic models each analyst estimated offer a richer view of the relationships between approval, the economy, and consumer sentiment than can be reflected with a regression table. Which interpretational tools are most useful depends on the analysts' goals. While the analyses presented in this section seem straightforward, there are a number of things that could complicate these analyses that should be considered. In the next section we discuss some of these considerations.

## ADDITIONAL CONSIDERATIONS

The model specification strategy we outline in this chapter provides a principled approach to the challenges of dynamic specification. The analyst chooses a PGM that encompasses the relationships of interest, tests that the PGM is dynamically complete, and simplifies the PGM to identify an estimable model. This procedure only works if the analyst is able to identify a dynamically complete model. In some cases, a PGM will be inadequate because the general lag structure chosen by the analyst is insufficient but, in many cases, the PGM will fall short because the analyst has not accurately modeled all the features of the relationship of interest. In this section, we briefly describe some of those features and describe how these features of the data can be addressed.

Dynamic modeling can be complicated by nonstationary time series. A stationary series is characterized by a predictable long-run equilibrium. Shocks to a stationary series can move the series away from its equilibrium, but these shocks tend to dissipate and the series returns to its long-run mean. The mean and variance of a stationary series are constant over time. In contrast, the mean and/or

variance of nonstationary series change over time. This can occur because a series is trending or because random shocks to the series do not dissipate like shocks to stationary series. Instead, the shocks integrate into the series causing the series to wander unpredictably. These 'random walk' series are said to be integrated of order  $d = 1$ , or  $I(1)$ , where  $d$  is the number of times the series must be differenced to produce a stationary series. Random walk series do not exhibit mean-reverting behavior. This makes these series difficult to model.

A dynamic model of a nonstationary series can only be dynamically complete if the model is balanced. That is, the regressors must account for the dominant features of the regressand. If the regressand is an  $I(1)$  variable, one or more of the regressors must also be  $I(1)$  variables for the equation to exhibit balance. If such a relationship exists, the variables are said to be cointegrated. Cointegrating variables exhibit equilibrating behavior with respect to one another. Balance is a necessary but not sufficient condition for dynamic completeness. If one regresses one  $I(1)$  variable on another and the residuals from that regression are not white noise, the model is not dynamically complete. The standard errors from such regressions will be incorrect and the limiting distributions for the test statistics will be nonstandard. Variables from these kinds of regressions will often appear to be related when they are not. This is the classic spurious regression problem. Standard practices dictate that analysts apply a series of pre-tests to determine whether the series in a model are  $I(1)$ . There are also tests for cointegration, but these tests rely on the assumption that the series are  $I(1)$  (Philips, 2018). The unit root and stationarity tests used to classify series as  $I(1)$  or  $I(0)$  processes play a critical role in applied time series analysis because classification dictates decisions about model choice and influences interpretation and inference.

This is a problem because unit root and stationarity tests are notoriously unreliable

(Campbell and Perron, 1991; Banerjee et al., 1993; Perron and Ng, 1996; Juhl and Xiao, 2003). The tests have low power in optimal circumstances. The performance of these tests depend upon the analyst accurately identifying the features of the data and modifying the test to accommodate these features. Results, also often hinge on the analysts's chosen level of statistical significance. Different tests often produce competing results. The uncertainty associated with these pre-testing procedures are not reflected in final analyses. This uncertainty plagues standard approaches to applied time series analysis but a number of methods have been developed that can accommodate uncertainty about the univariate properties of the data. Webb et al. (2019; 2020) develop a hypothesis testing procedure that uses critical value bounds for the LRM  $t$ -statistic to accommodate this uncertainty. Brandt and Freeman (2006, 2009) also discuss how uncertainty about univariate dynamics can be incorporated into structural VAR models using Bayesian priors.

Dynamic modeling is also complicated by the presence of structural breaks. A structural break is an abrupt change in the mean of a series. A break may reflect a fundamental change in the underlying parameters of the DGP or a simple change in the mean caused by a shock or intervention. For example, the number of international skyjackings changed after metal detectors were introduced in airports. Structural breaks represent another dominant feature of a variable that must be accounted for in the model. Structural breaks will generate non random patterns in model residuals and can frustrate the already sensitive unit root tests just mentioned. Ignoring a structural break is tantamount to ignoring a critical omitted variable and omitting the structural break creates omitted variable bias problems. Analysts can test for structural breaks and can include variables in models to partial out the effects of these changes on the series or test hypotheses about the breaks.

Dynamic modeling can also be complicated by changes in the variance of a series over time. The models described in this chapter, and many of the diagnostic tests used in the prescribed modeling procedure, rely on the assumption that errors are white noise. White noise residuals are not only random with respect to one another, they also have constant variance. Many time series are characterized by autoregressive conditional heteroskedasticity (ARCH). The variance of the series changes as a function of time. Autoregressive conditional heteroskedasticity affects standard error estimates and can influence the results of diagnostic tests. More importantly, a model cannot be said to be dynamically complete unless it captures these features of the data as well. Engle (1982) developed a test, and model, for ARCH residuals. This model has been extended by Bollerslev (1986) and others to accommodate different dynamic patterns in the conditional variance.

Finally, single equation models rely on the assumption that the regressors are weakly exogenous to the regressand. This assumption is difficult to verify. There are no direct tests of the assumption but analysts can use a number of tools that shed light on the veracity of the assumption. If one variable Granger causes another, the latter variable cannot be exogenous to the former. This will also be reflected in IRFs and FEVDs. Violations of parameter non constancy in single equation regressions may also suggest violations of weak exogeneity but may indicate other kinds of omitted variables.

Unit roots, structural breaks, ARCH errors, and endogeneity create challenges for dynamic modeling because these phenomena constitute important information that must be accounted for in a dynamically complete model. Analysts should strive to identify the important features of their data and incorporate this information into the model selection process. This is true of all empirical models, not just time series analysis.

## CONCLUSION

Quantitative political scientists endeavor to make general statements about the relationships that hold in the complex political world. We are armed with abstract theory and left to fend with imperfect data. This is true of virtually all domains of quantitative analysis. But time series analysis is unique in two aspects. First, there is typically no theory-guided map for dynamic specification. That is, theory is silent as to whether economic performance affects evaluations of the president immediately, the next month, or two months into the future. Second, temporal aggregation and other features of time series measurement complicate identification of the appropriate lag structure. The same construct measured over different time periods will exhibit different dynamic structures. For example, if economic performance influences approval in the following month, the effects will change when the data are aggregated to quarters. We need to draw not only on theory but also on information in the data in order to arrive at acceptable empirical models of the relationships we care about.

Traditionally, analysts have buried their heads in the sand, maintaining the illusion that theory dictates an empirical specification that can be validated (or not) with the available data, no matter how it is aggregated, if we just 'fix' the statistical problems that manifest with dynamic misspecification. But ignoring the realities of the data and eschewing the evidence of dynamic misspecification means such models cannot tell us about the dynamic relationships in the data. Perhaps more problematically, they give rise to unexpected incongruencies between theory and data that limit scientific progress.

The principled approach we outline in this chapter provides analysts a strategy for dynamic specification that incorporates information from theory and data. The end product is a statistically adequate model that can be used to produce inferences about theorized

relationships that tell us something about the more complex political world. The approach does not guarantee the model is useful. The final model will only be as good as the information we bring to the modeling process. If we find our inferences implausible, we may lack relevant theoretical information because we simply do not understand the world well enough, our measures may poorly tap the theoretical constructs of interest, the level of aggregation in the data may obscure relationships, or the time period we have chosen may be too short or too unusual. In these cases, we need to gather more information. This search for additional information to incorporate in our models of the world is the hard work of a progressive scientific program. Dynamic specification should reflect this basic principle: use all the information available in theory and data.

## Notes

- 1 Regular patterns of autocorrelation need not be tied to the calendar year. For example, one might observe weekly patterns in daily financial time series data.
- 2 Various unit root and stationarity tests that have been proposed to classify series as stationary or non stationary. For a more in-depth discussion of unit root tests and the limitations of unit root tests, see Choi (2015).
- 3 For a more complete treatment on Vector Autoregression, see Brandt and Williams (2007).
- 4 Some software packages offer automated procedures that select models using information criteria. We caution against these strategies because these processes only consider the statistical information in the models and can produce specifications that are not dynamically complete.
- 5 Note that the AIC and BIC presented for the ECM cannot be compared to those for the other models as the dependent variable is different.
- 6 If one chose to estimate a system of seemingly unrelated regressions with varying lag structures, more options would exist.
- 7 A Granger-causal relationship does not imply a causal relationship in the classic sense, only a specific type of statistical relationship.
- 8 For the sake of brevity, we do not present the FEVDs.

## REFERENCES

- Banerjee, A., Dolado, J. J., Galbraith, J. W., and Hendry, D. (1993). *Co-integration, error correction, and the econometric analysis of non-stationary data*. Oxford: Oxford University Press.
- Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of econometrics*, 31(3), 307–327.
- Box, G. E., and Jenkins, G. M. (1976). *Time series analysis: forecasting and control*, 2nd edn. San Francisco, CA: Holden-Day.
- Box-Steffensmeier, J. M., Freeman, J. R., Hitt, M. P., and Pevehouse, J. C. (2014). *Time series analysis for the social sciences*. New York, NY: Cambridge University Press.
- Brandt, P. T., Colaresi, M., and Freeman, J. R. (2008). The dynamics of reciprocity, accountability, and credibility. *Journal of Conflict Resolution*, 52(3), 343–374.
- Brandt, P. T., and Freeman, J. R. (2006). Advances in Bayesian time series modeling and the study of politics: theory testing, forecasting, and policy analysis. *Political Analysis*, 14(1), 1–36.
- Brandt, P. T., and Freeman, J. R. (2009). Modeling macro-political dynamics. *Political Analysis*, 17(2), 113–142.
- Brandt, P. T., and Sandler, T. (2012). A Bayesian Poisson vector autoregression model. *Political Analysis*, 20(3), 292–315.
- Brandt, P. T., and Williams, J. T. (2007). *Multiple time series models*. London: Sage Publications.
- Campbell, J. Y., and Perron, P. (1991). Pitfalls and opportunities: what macroeconomists should know about unit roots. *NBER Macroeconomics Annual*, 6, 141–201.
- Campos, J., Ericsson, N. R., and Hendry, D. F. (2005). *General-to-specific modeling: an overview and selected bibliography*. FRB International Finance Discussion Paper No. 838. Available at SSRN: <https://ssrn.com/abstract=791684> or <http://dx.doi.org/10.2139/ssrn.791684>
- Choi, I. (2015). *Almost all about unit roots: foundations, development, and applications*. New York, NY: Cambridge University Press.
- De Boef, S., and Keele, L. (2008). Taking time seriously. *American Journal of Political Science*, 52(1), 184–200.

- De Boef, S., and Kellstedt, P. M. (2004). The political (and economic) origins of consumer confidence. *American Journal of Political Science*, 48(4), 633–649.
- Enders, W. (2015). *Applied Econometric Time Series*, 4th edn. Hoboken, NJ: Wiley.
- Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica: Journal of the Econometric Society*, 50(4), 987–1007.
- Erikson, R. S., MacKuen, M. B., and Stimson, J. A. (2002). *The macro polity*. New York, NY: Cambridge University Press.
- Freeman, J. R. (1989). Systematic sampling, temporal aggregation, and the study of political relationships. *Political Analysis*, 1, 61–98.
- Gilbert, C. L. (1990). Professor Hendry's econometric methodology. In C.W.J. Granger (Ed.), *Modelling economic series: readings in econometric methodology*. (pp 279–303). New York, NY: Oxford University Press.
- Granger, C. W. J. (Ed.). (1990). *Modelling economic series: readings in econometric methodology*. New York, NY: Oxford University Press.
- Hendry, David F. (1995). *Dynamic econometrics*. New York, NY: Oxford University Press.
- Juhl, T., and Xiao, Z. (2003). Power functions and envelopes for unit root tests. *Econometric Theory*, 19(2), 240–253.
- King, G., and Roberts, M. E. (2015). How robust standard errors expose methodological problems they do not fix, and what to do about it. *Political Analysis*, 23(2), 159–179.
- MacKuen, M. B., Erikson, R. S., and Stimson, J. A. (1992). Peasants or bankers? The American electorate and the US economy. *American Political Science Review*, 86(3), 597–611.
- Pagan, A. (1987). Three econometric methodologies: a critical appraisal. 1. *Journal of Economic Surveys*, 1(1–2), 3–23.
- Perron, P., and Ng, S. (1996). Useful modifications to some unit root tests with dependent errors and their local asymptotic properties. *The Review of Economic Studies*, 63(3), 435–463.
- Philips, A. Q. (2018). Have your cake and eat it too? Cointegration and dynamic inference from autoregressive distributed lag models. *American Journal of Political Science*, 62(1), 230–244.
- Spanos, A. (1990). Towards a unifying methodological framework for econometric modelling. In C.W.J. Granger (Ed.), *Modelling economic series: readings in econometric methodology*. (pp 335–364). New York, NY: Oxford University Press.
- Webb, C., Linn, S., and Lebo, M. (2019). A bounds approach to inference using the long run multiplier. *Political Analysis*, 27(3), 281–301.
- Webb, C., Linn, S., and Lebo, M. (2020). Beyond the unit root question: uncertainty and inference. *American Journal of Political Science*. DOI: <https://doi.org/10.1111/ajps.12506>





# Time-Series-Cross-Section Analysis

Vera Troeger

## INTRODUCTION

The use of pooled time series cross section (PTSCS) data has become ubiquitous in observational analyses across the social sciences. Even the identification revolution that swept through empirical social science in the last two decades finds some merit in using data that combines observations across units and over time, because it allows identification through differences-in-differences approaches exploiting within unit variation. In *Mostly Harmless Econometrics* (Angrist and Pischke, 2009) – which has obtained cult-like status in applied empirical economics – the authors recommend the use of diff-in-diff or unit fixed effects approaches to identification, requiring pooled data, if an experiment is infeasible, a meaningful discontinuity, or exogenous variation in the form of an instrument cannot be found.

Pooled data analysis has become the standard for analyzing observational data in quantitative political analysis. This is particularly true in sub-disciplines like International

Relations, Comparative Politics, and Comparative Political Economy, but it has even extended to fields that use micro data, such as Political Behavior or American Politics. Because more survey data over time is available, these fields are using more PTSCS data. Panel data pool cross-sectional information (number of units  $N$ ) with information over time (number of time points  $T$ ), for example, data on individuals or firms at different points in time and information on countries and regions over time etc. Thus, panel data consist of repeated observations on a number of units. We can distinguish between cross-sectional dominant data (Cross-Section Time-Series (CSTS)), time-series dominant data (Time-Series Cross-Section (TSCS)), or pooled data with a fixed number of units and time-points. The data structure has implications for the model choice, since asymptotic properties of estimators for pooled data are either derived for  $N \rightarrow \infty$  or  $T \rightarrow \infty$ . In addition, violations of full ideal conditions and specification issues have more or less severe

effects for bias and efficiency, depending on whether the number of units exceeds the number of observations over time, or vice versa. Below, we discuss the strengths and weaknesses of this method and various ways to cope with some of the inherent problems.

Some have argued that TSCS and CSTS data consist of observations at different points in time for fixed units of theoretical interest, such as countries or dyads. In contrast, in panel data, the units, mostly individuals in surveys, are of no specific interest and are randomly sampled from an underlying population with all inferences dedicated to uncovering the relationships in the population. Textbooks and articles, however, use these terms quite loosely. This entry will follow this trend and discuss general estimation procedures and specification issues with respect to different kinds of data pooling cross-sectional and time series information.

For each specification issue, this entry briefly discusses the solutions presented in commonly used textbooks like Wooldridge (2010) and then turns to more recent discussions in the political methodology literature.

## ADVANTAGES AND DISADVANTAGES OF PTSCS DATA ANALYSIS

Panel data pool observations for units ( $i$ ) and time periods ( $t$ ). The typical data generating process can be characterized as:

$$y_{it} = \alpha + \sum_{k=1}^K \beta_k x_{kit} + \epsilon_{it}, \quad i = 1, \dots, N; \quad t = 1, \dots, T \quad (1)$$

with  $k$  independent variables  $x$  which have observations for  $N$  units ( $i$ ) and  $T$  periods ( $t$ ). The dependent variable  $y$  is continuous (though in principle it can be a limited dependent variable, which requires non-linear estimation procedures) and also observed for  $i$  and  $t$ .  $\epsilon_{it}$  describes the error term for observations  $i$ ,  $t$  and we can assume a  $NT \times NT$  Variance-Covariance Matrix  $\Omega$  of the error term with the typical element  $E(\epsilon_{it}, \epsilon_{js})$ . In case all

Gauss–Markov assumptions are met (the error term is iid) this model can be straightforwardly estimated by OLS. Since PTSCS data combine time-series and cross-section information, this is rarely the case. However, the analysis of PTSCS data offers significant advantages over the analysis of pure time series or pure cross-sectional data. First, using pooled data increases the number of observations and therefore the degrees of freedom which means that more complex arguments can be tested by employing more complex estimation procedures. More importantly, most theories in the social sciences generate predictions over space and time, and it seems imperative therefore to test these hypotheses by using data providing repeated information for theoretically interesting units. PTSCS data analysis can be used to model dynamics, which is impossible when examining pure cross-sections and may lead to spurious regression results. Finally, it is possible to control for unit heterogeneity when analyzing pooled data, beyond the inclusion of additional right-hand side (RHS) variables. Accordingly, we use pooled data to eliminate some kinds of omitted variable bias, make the best of the available information, test theories that predict changes, and test theories that predict parameter heterogeneity.

The most obvious disadvantage of panel data analysis is that an econometrically sound model specification is typically hard to find, since the data structure combines all of the problems of cross-sectional and time-series data, but these problems typically arise simultaneously. Specification problems in pooled data analysis can be summarized as follows:

- 1 The residuals are typically serially correlated and not independent of each other
- 2 The residuals have different variances for different units (panel heteroskedasticity)
- 3 The residuals of different units are contemporaneously correlated
- 4 The residuals of unit  $i$  co-varies with residuals of unit  $j$  for different points in time
- 5 The expected mean of the error term deviates from zero for different units

While each single violation of the underlying model assumptions is often straightforwardly accounted for by existing econometric measures, combinations of problems might not be solved simultaneously in a satisfying manner. Econometric solutions are often incompatible with theories. Sometimes it is hard to find models that are at the very same time econometrically sound (unbiased, efficient) and provide an appropriate test of the theory. When weighing advantages and disadvantages of pooled data analysis, the positive aspects certainly prevail – especially because the analysis of pooled data allows testing complex arguments over space and time, which is characteristic for the social sciences. From this perspective, the steep increase in the popularity of panel data analysis does not seem surprising. However, the big challenge using pooled data is how to deal with simultaneously occurring violations of the underlying assumptions.

First, I will discuss common violations of underlying assumptions and solutions for each misspecification separately and will later turn to attempts to account for simultaneously occurring specification problems.

### **HETEROSKEDASTICITY AND CONTEMPORANEOUS ERROR CORRELATION IN PTSCS DATA**

Heteroskedasticity in pooled data presents a more complex problem than in pure cross-sections since (a) the error term can have unit-specific variances (panel heteroskedasticity), (b) the error term can be contemporaneously correlated, that is, the error term of unit  $i$  is correlated to that of unit  $j$  in the same year, and (c) the error term of one unit  $i$  can be correlated with the error term of unit  $j$  at different points in time. In addition, the error term can have time dependent error variances (autoregressive conditional heteroskedasticity). Panel heteroskedasticity mainly occurs if the model specification fits different units with a

different degree of accuracy. Correlations of the errors across units are determined by unobserved features of one unit that are linked to another unit. Both features violate Gauss–Markov assumptions: while they leave the simple estimators consistent, such estimators are now inefficient and standard errors may be incorrect. More importantly, both heteroskedasticity and error correlation often signal omitted variables bias, since in both cases something that should have been included into the structural part of the equation was left out.

This problem can be solved in a substantive way by identifying the causes of the omitted variable bias and including these variables into the right-hand side of the models. Often, this approach is not feasible because the sources for heteroscedasticity are not known or excluded factors cannot be measured. In this case, several econometric solutions have been proposed. Parks (1967) and Kmenta (1986) were the first to propose a Feasible Generalized Least Squares (FGLS) estimation which is characterized by a  $NT \times NT$  block diagonal matrix with an  $N \times N$  matrix  $\Sigma$  that contains contemporaneous covariances along the block diagonal. Parks (1967) and Kmenta (1986) also suggest an  $\Omega$  matrix with panel specific AR1 error structure and contemporaneously correlated errors, but, in principle, FGLS can handle all different correlation structures. Because the true structure of  $\Sigma$  and  $\Omega$  is unknown, this procedure requires estimating a very large number of parameters in order to obtain the error covariances, which in turn leads to very inefficient, and therefore unreliable, results. Beck and Katz (1995) show that the Parks' method highly underestimates standard errors and therefore induces overconfidence in estimation results. As a result, this estimation procedure has fallen into disuse in recent work using pooled data.

Beck and Katz (1995) suggest a different way of dealing with panel heteroskedasticity. They argue that coefficient estimates of OLS are consistent but inefficient in pooled data

and that the degree of inefficiency depends on the data and the exact error process. They suggest using OLS and correcting the estimated standard errors by taking the specific panel structure of the data into account:

$$Var[\beta] = (X'X)^{-1} X' \Omega X (X'X)^{-1} \tag{2}$$

with

$$\Omega = (E'E / T) \otimes I_t \tag{3}$$

This method is dubbed panel corrected standard errors. Other violations of Gauss–Markov assumptions, such as serial correlation of the error term, have to be treated beforehand. Since this approach only manipulates the standard errors of an OLS model, the coefficients are biased whenever OLS is biased.

### DYNAMICS IN POOLED DATA ANALYSIS

As pooled data combine information across units and over time, another problem arises if dynamics are present and the error term is serially correlated. The error term in  $t$  is dependent on the error term in  $t-1$ :

$$\epsilon_{it} = \rho_i \epsilon_{it-1} + \xi_{it} \tag{4}$$

From a formal econometric point of view, violating the independence assumption only influences the efficiency of the estimation. Yet, since the residual of a regression model picks up the influences of those variables that have not been included, persistence in excluded variables is the most frequent cause of serial correlation. Several remedies for serial correlation are available, all of which have different consequences for the model specification and interpretation of the estimation results.

A substantive solution to the problem of serial correlation is the inclusion of a lagged dependent variable  $y_{it}$  (LDV) to the right-hand side of the regression equation:

$$y_{it} = \beta_0 y_{it-1} + \beta_k x_{kit} + u_i + \epsilon_{it} \tag{5}$$

In many cases this is enough to eliminate serially correlated error terms. However, there are also many perils to adding an LDV to the list of regressors. One of the main problems arises because the inclusion of an LDV makes it very hard to interpret the effects of the substantial RHS variables directly and correctly, since the conditional effect of  $x$  on  $y$  is dynamic and aggregated over all periods. It can be described by the following polynomial:

$$y(x)_{t_1 \rightarrow t_p} = \beta_1 x_{it} + \sum_{p=1}^{t_p} (\beta_0^{t-p} \beta_1 x_{it}) \tag{6}$$

the long-term effect of  $x_k$  reduces to:

$$\frac{\widehat{\beta}_k}{(1 - \beta_0)} \tag{7}$$

Unfortunately, the standard errors of the function in Equations 6 and 7 cannot be easily calculated. Since including an LDV resembles a shortened distributed lag model, we implicitly assume that all variables exert an equally strong one period lagged impact on the dependent variable. Therefore, finding a non-significant coefficient of a theoretically interesting explanatory variable in an LDV model does not necessarily mean that this variable has no effect. It only tells us that this variable does not affect the dependent variable contemporaneously, but it might still have a lagged effect. From this, it follows that the coefficient of the LDV estimates at best the average dynamic effect of all substantive RHS variables, rather than the actual dynamic effect of each explanatory variable. When including an LDV into the specification, it is very important to calculate or simulate the short- and long-term effects of all RHS variables to correctly interpret effects, their size, and significance. Recent work by Williams and Whitten (2011) shows how to simulate long-term dynamics for autoregressive series in pooled data.

Another problem occurs when combining an LDV with the estimation of unit-specific effects by a fixed effects specification or a

least squares dummy variable model (see the next section for a more detailed description). This leads to biased estimates since the LDV co-varies with the time-invariant part of the error term. This problem is called Nickell-bias (Nickell, 1981). The Nickell-bias is sizable in panel data with very short time periods (Pickup, 2018) but becomes negligible as  $T$  increases. The best known suggestions tackling the problem of Nickell-bias are the instrumental variable approach by Anderson–Hsiao (Anderson and Hsiao, 1981) (AH), the differenced GMM (Generalized Methods of Moments) model by Arellano–Bond (Arellano and Bond, 1991) (AB), and the Kiviet (1995) correction, which proposes a corrected within estimator that subtracts a consistent estimate of the bias from the original fixed effects estimator. The first two approaches solve the bias problem by taking the first difference of both sides of the regression equation and instrumentation of the LDV with higher order lags of the LDV. Therefore, AH is only using the two periods lagged LDV as an instrument while AB can make use of all possible lags of the LDV and all exogenous variables in the model. Both approaches generate asymptotically consistent estimation results, but AB produces more efficient estimates due to the exploitation of all moment conditions. In finite samples, however, both estimators are problematic with regard to efficiency, as recent Monte Carlo experiments examining the finite sample properties reveal (Pickup, 2018). Higher lags of the LDV provide good instruments only in the case where  $y$  is highly persistent over time. Unfortunately, in such a case, the probability that the instruments also co-vary with the error term remains high. From this perspective, both estimators cannot solve the problem of Nickell-bias if  $y$  is highly persistent or solve the problem very inefficiently, in case of low persistence. More recent research by Pickup (2018) shows evidence that transformed-likelihood estimators – for example, the orthogonal reparameterization (OPM) fixed-effects approach proposed by

Lancaster (2000) and the quasi-maximum likelihood (QML) fixed-effects approach by Hsiao et al. (2002) – outperform the so called dynamic panel models discussed above by far, especially for very short  $T$ .

A Prais–Winsten (PW) transformation of the model offers another solution to serial correlation. The advantage of the Prais–Winsten approach lies in the transformation of both the left and the right-hand side of the equation, which allows a direct interpretation of the regression coefficients. Prais–Winsten is estimated by GLS and is derived from the AR(1) model for the error term. First a standard linear regression is estimated:

$$y_{it} = x_{it}\beta + \epsilon_{it} \quad (8)$$

An estimate of the correlation in the residuals is then obtained by the following auxiliary regression:

$$\epsilon_{it} = \rho\epsilon_{it-1} + \xi_{it} \quad (9)$$

A Cochrane–Orcutt transformation is applied for observations  $t = 2, \dots, n$

$$y_{it} - \rho y_{it-1} = \beta(x_{it} - \rho x_{it-1}) + \zeta_{it} \quad (10)$$

And the transformation for  $t=1$  is as follows:

$$\sqrt{1 - \rho^2} y_{i1} = \beta(\sqrt{1 - \rho^2} x_{i1}) + \sqrt{1 - \rho^2} \zeta_{i1} \quad (11)$$

Equation 11 shows that another advantage of the Prais–Winsten transformation is the preservation of the first period. The differences between a PW and an LDV model might be substantial depending on  $\rho$  and the serial correlation in both  $y$  and  $x$ .

Dynamics in pooled data estimation presents one of the major challenges, because the sources of the dynamics often remain unknown to the researcher. Therefore, specifying dynamics in panel data is difficult. A researcher may believe that the data generating process (DGP) for their data contains a lag of the dependent variable, lags of the independent variables, and/or serial correlation in the errors. If researchers could observe

the DGP, they could select the corresponding model and, in doing so, account for the dynamics. Unfortunately, they cannot and if residuals show clear evidence of serial correlation this could indicate missing dynamics in either the mean or the error equation (or both). There is no straightforward statistical test that can determine which is one it is. This leaves the researcher uncertain how to proceed. Some scholars suggest always adding dynamics to the mean equation (Wilkins, 2018) or at least ensuring the data model includes many lags of the dependent and independent variables as there are in the data generating process (Wooldridge, 2010: 194). However, the consequences of doing so when the DGP dynamics are in the error equation remain largely unknown. Pickup and Troeger (2019) explore theoretically and empirically whether the source of the dynamic process makes a difference for bias and efficiency of the estimation. They also show that the consequences for bias can be severe when dynamics are included in the mean equation when they should be in the error equation and vice versa.

Whether the dynamics in a particular set of panel data need to be accounted for in the error term or in the mean equation can be unclear. There are four common DGPs that might create uncertainty for the researcher as to whether they require a static model with a control for serial correlated errors or a dynamic model: (a) slow dissipation of effects due to covariates included or not included in the estimation equation; (b) lags of the covariates that are included in the estimation; (c) errors in the DGP that are serially correlated, often due to measurement artifacts; and (d) autoregressive covariates omitted from the estimation.

The first two types of DGPs are dynamic in the mean equation. The first is dynamic in the dependent variable and the second is dynamic in the independent variable. The third and fourth types are static in the mean equation. The third is dynamic in the error equation. The fourth is static in the error equation but if the autoregressive covariate is excluded from the data model, the DGP for the residuals

will be dynamic. Given that the DGP may be dynamic in the mean equation or in the error equation and given that the researcher does not typically observe the DGP, it is important to consider the consequences of: (a) using a model that is dynamic in the mean equation; or (b) static in the mean equation with a control for dynamics in the error equation: first, if a dynamic model is used (correctly or incorrectly), it must be dynamically complete. That is an autoregressive distributive lag (ADL(p,q)) model with p lags of the dependent and q lags of the independent variables needs to be estimated to avoid bias:

$$y_{i,t} = \alpha_1 y_{i,t-1} + \dots + \alpha_p y_{i,t-p} + \beta_1 x_{i,t} + \beta_2 x_{i,t-1} + \dots + \beta_q x_{i,t-q} + u_i + \epsilon_{i,t} \quad (12)$$

Second, if the dynamic model is incorrectly used, it will either not lead to incorrect inference (in the case of AR errors and assuming dynamic completeness) or it will result in MA serially correlated errors. In the latter case, the estimation will likely generate some degree of bias. The magnitude of this bias will depend on  $T$  and the degree of serial correlation. Clearly, it is of utmost importance to specify the dynamic processes as closely to the true DGP as possible. Otherwise, biased estimates will occur. As a consequence, testing for serial correlation and detecting misspecification in dynamic panel models is very important.

However, in reality, applied researchers often perceive serially correlated errors as noise rather than information (DeBoef and Keele, 2008). Yet, serially correlated errors clearly indicate a potentially severe model misspecification, which can result from various sources and occur either in the mean- or error-equation. Perhaps most obviously, serially correlated errors are caused by: incompletely or incorrectly modeled persistency in the dependent variable, time-varying omitted variables or changes in the effect strengths of time-invariant variables, or misspecified lagged effects of explanatory variables. Conditionality makes modeling dynamics

more complicated (Franzese, 2003a, 2003b; Franzese and Kam, 2009). Few empirical analyses model all potential conditioning factors of the variables of interest. If, however, treatment effects are conditioned by unobserved time-varying factors then treatment effects vary over time, and the strength of these effects also changes over time as unmodeled conditioning factors change. Finally, serially correlated errors may result from misspecifications that, at first sight, have little to do with dynamics – for example, from spatial dependence. Yet, spatial effects are certainly misunderstood if they are perceived as time-invariant: ignoring spatial dependence causes errors to be serially correlated (Franzese and Hays, 2007).

In an ideal world, these model misspecifications would be avoided: dynamics should be directly modeled to obtain unbiased estimates. This proves difficult in reality. Since dynamic misspecifications are manifold and complex, econometric tests for dynamics, at best, reveal serially correlated errors, but they are usually unable to identify the underlying causes of autocorrelation. Often, these tests are also weak and do not reveal the true dynamic structure of the DGP, which may lead to overfitting of the data (Keele et al., 2016). Thus, empirical researchers often try to simplify their empirical model and to treat problems such as serially correlated errors with straightforward econometric textbook solutions, such as lagged dependent variables, period dummies, and simple homogeneous lag structures.

Plümper and Troeger (2019) show that econometric textbook solutions are not correct per se because they are usually not modeling the true dynamic process in the underlying DGP. In addition, if dynamics occur in combination with other misspecifications, for example, unit heterogeneity, treating one problem can render the effects (bias) of another violation worse.

One strategy that may reduce the size of the problem is to use less constrained econometric solutions. Distributed lag models, models with a unit-specific lagged dependent

variable, panel co-integration models, models with heterogeneous lag structure (Plümper et al., 2005), more attention to periodization (Franzese, 2003a), and better specified spatial models (Franzese and Hays, 2007) may all reduce the size of the problem. However, as the number of possible dynamic specifications increases, a higher order problem of model selection arises: since all of these different models generate different estimates and often demand different inferences, the next question is how empirical researchers select their preferred model. To eliminate, or at least reduce, the arbitrariness of model selection, DeBoef and Keele (2008: 187) suggest a testing down approach, starting with a full autoregressive distributive lag model and stepwise removing parameters according to predetermined criteria, often the significance of parameters. This procedure will result in a dynamic specification that maximizes the variance absorbed by the minimum number of parameters. As with all testing down approaches, this approach suffers from the arbitrariness in the choice of a starting model because we do not have an infinite number of degrees of freedom. Pickup (2018) similarly suggests a general-to-specific approach to modeling dynamics especially for panel data with small  $T$  to find a plausible dynamic specification before dealing with other misspecifications, such as unit heterogeneity.

The issue of specifying dynamic processes in pooled data becomes even more complicated if this problem is coupled with other potential misspecifications such as unit heterogeneity. The Nickell-bias discussed here is the least of the problems that occurs when addressing both issues separately, as is discussed below.

## HETEROGENEITY

The identification revolution has not failed to impact the analysis of PCSTS data. Since Angrist and Pischke (2009) published *Mostly*

*Harmless Econometrics*, the so-called Fixed Effects Model has become the workhorse specification when using pooled data of any kind. Clearly, one of the advantages of analyzing pooled data is the possibility of controlling for heterogeneity across units. When examining cross-sectional data, it is impossible to tell whether the estimated effects are contingent on unobserved effects that are specific to each unit, and therefore biased. PTSCS data analysis rests on the assumption that units are similar enough to be pooled together. If that is not the case, we can still find appropriate specifications that allow accounting for differences across units, which might influence the estimation results. Textbooks usually discuss this problem under the header unit-heterogeneity and offer remedies such as fixed effects or random effects models. However, these models only deal with time invariant unit-specific effects: units can also be heterogeneous with respect to slope parameters, dynamics or lag structures. The following sections discuss different versions of unit heterogeneity, approaches to dealing with them, and their advantages and disadvantages.

### Unit Heterogeneity

When units have specific characteristics which cannot be measured and are time invariant, they offer different initial conditions which might bias the estimated coefficients. For example, geography is often considered time invariant – a country or region can be landlocked or on the European continent, cities have a certain distance to the next port, etc. Other examples are inheritance, being a former colony, the sex of an individual, or her genetic pool. These are inherited specific to this unit and do not change over time, especially if these time invariant unit-specific effects are correlated with any of the RHS variables, for example, if the gender of a person determines specific behavior such as party identification or voting, coefficient estimates are distorted by

omitted variable bias. If that is the case and we do not control for unit-specific effects, the Gauss–Markov assumption of  $x$  being deterministic is violated:

$$y_{it} = \sum_{k=1}^K \beta_k x_{kit} + \sum_{m=1}^M \gamma_m z_{mi} + u_i + \epsilon_{it} \quad (13)$$

where  $u_i$  denotes the unit-specific effects and  $z$  other explanatory variables that are time invariant but can be measured and are of theoretical interest. If  $u_i$  is excluded from the estimation, it becomes part of the overall error term, and will make the model less efficient in the case that it does not co-vary with any of the  $x$  or  $z$  but induces bias if  $u_i$  co-varies with any of the regressors. Econometrically, we can solve for correlated unit-specific effects by including a dummy variable for each unit into the right-hand side of the model which generates unit-specific intercepts. This estimation procedure is called a Least Squares Dummy Variable (LSDV) model.

$$y_{it} = \sum_{k=1}^K \beta_k x_{kit} + \gamma_{n-1} D_i + \epsilon_{it} \quad (14)$$

The unit-specific dummy variables ( $D_i$ ) are multi-collinear to any time invariant variable  $z$ , the coefficients for  $z$  are therefore not identified. We also can employ fixed effects (FE) specification which is econometrically equivalent to a LSDV model. The fixed effects model first de-means all variables in the model by subtracting the unit-specific mean and then estimates the transformed equation by OLS.

$$y_{it} - \bar{y}_i = \sum_{k=1}^K \beta_k (x_{kit} - \bar{x}_{ki}) + \epsilon_{it} - \bar{\epsilon}_i + u_i - \bar{u}_i \quad (15)$$

$$\equiv \dot{y}_i = \sum_{k=1}^K \beta_k \dot{x}_{kit} + \dot{\epsilon}_i \quad (16)$$

with

$$\bar{y}_i = \frac{1}{T} \sum_{t=1}^T y_{it}, \quad \bar{x}_i = \frac{1}{T} \sum_{t=1}^T x_{it}, \quad \bar{\epsilon}_i = \frac{1}{T} \sum_{t=1}^T \epsilon_{it}$$

The fixed effects transformation eliminates the unit-specific effects, but also time invariant variables that might be of theoretical



interest. FE can become highly inefficient because it only uses the within information of all variables. Yet, not controlling for unit-specific effects leads to biased estimates if unit effects exist and are correlated with any of the regressors.

If unit-specific effects do exist but do not co-vary with any of the RHS variables, not controlling for unit effects does not bias the estimates but instead increases the sampling variation of the OLS estimator, and therefore generates less efficient estimates. A straightforward remedy is a random effects (RE) specification which treats the  $u_i$  as a random unit-specific part of the error term. The random effects model only quasi-demeans the variables: rather than removing the time average from the explanatory and dependent variables at each  $t$ , RE removes a fraction of the time average. The RE estimator generates more efficient results than the FE estimator but the RE model produces biased estimates if RHS variables co-vary with the unobserved unit-specific effects. RE resembles a feasible GLS estimator where the  $\Omega$  matrix (VC matrix of the error term) has a specific RE structure which only depends on two parameters:  $\sigma_u^2$  and  $\sigma_e^2$ . RE and FE estimates tend to grow similar if  $T$  gets large or the variance of the estimated unit effects increases, as compared to the error variance.

Since the RE estimator is more efficient than FE if the unit effects are uncorrelated with the regressors, it is useful to determine which of the two specifications should be used. Textbooks typically suggest employing the Hausman test. The Hausman test (Hausman, 1978) is based on the following logic: since the RE estimator is biased if unit-specific effects are correlated, differences between FE and RE estimates are interpreted as evidence against the random effect's assumption of zero covariance between  $x$  and  $u_i$ . Econometricians attest that the Hausman test has good asymptotic properties. Nevertheless, in finite samples, the test results are influenced by the trade-off between bias and efficiency. The Hausman

test is only powerful in the limit: since FE is consistent, the difference of RE and FE estimates can only be caused by biased RE estimates. In finite samples, however, the differences can result from two sources: biased RE estimates and unreliable FE point estimates due to inefficient estimation of variables with low within variation. The Hausman test actually mirrors this trade-off since it divides the difference between RE and FE estimates by the difference in the asymptotic variances of the RE and FE estimates. From this, it follows that the test results are especially unreliable if the estimation equation contains regressors which are both correlated with the unit-specific effects and are rarely changing over time. Recent research (Pickup and Troeger, 2019; Plümper and Troeger, 2019) shows that the Hausman test is highly unreliable, especially when the estimation also suffers from dynamic misspecifications or uses econometric fixes to model dynamics. The Hausman test is generally biased towards a fixed effects specification. Pickup and Troeger (2019) also show that the Mundlak formulation of the Hausman test is always preferable (especially when the estimation is dynamically more complete) because it does not rely on estimating the differences in variance between the fixed effects and random effects estimates.

The estimation of time-invariant or nearly time-invariant variables is highly problematic in a FE specification. It is easy to see that including completely time-invariant variables would be a problem, but it is less obvious that estimating rarely changing variables would be problematic, because FE specifications generate an estimate. However, this estimate might be very inefficient since FE eliminates all cross-sectional variation and only the variance over time is used to compute the coefficient. If this within unit variation is very small, the sampling variation of FE estimates increases dramatically, which leads not only to large standard errors but also to very unreliable point estimates. In empirical analyses across the social sciences, we are often

interested in the effect of variables that only change once in a while, for example, the level of democracy in a country, electoral rules, central bank independence, marital status, family income.

In the case of time invariant variables, applied researchers often resort to a simple pooled OLS or a RE model which permits the estimation of coefficients for time-invariant variables. These estimates are biased if the unit-specific effects co-vary with the regressors. Hausman and Taylor (1981), as well as Amemiya and MaCurdy (1986), propose estimators that use the uncorrelated RHS variables as instruments for the correlated regressors. The models are based on a correlated random effects model (Mundlak, 1978) and use instrumental variables for the endogenous RHS variables. The underlying assumption is that only some of the time-varying ( $x_{it}$ ) and time-invariant ( $z_i$ ) variables are correlated with the unit-specific effects  $u_i$ . The uncorrelated  $x_{it}$  and  $z_i$  therefore can be used as instruments for the correlated RHS variables. The within transformed  $x_{it}$  serve as instruments for the correlated  $x_{it}$  (these are estimated by FE) and the unit means of the uncorrelated  $x_{it}$  ( $\bar{x}_i$ ) as well as the uncorrelated  $z_i$  serve as instruments for the correlated  $z_i$ . However, if the instruments are poor, Hausman–Taylor produces highly inefficient parameter estimates.

Of course, a fixed effects estimator generates clean estimates for the within effect. However, theories often do not tell us whether we should observe an effect between or within units and whether these effects should be the same. Given that the fixed effects estimator generates highly unreliable estimates if RHS variables are slow moving, and specification tests are highly unreliable, more recently there has been a lot of focus on the conditions under which one should use a fixed or random effects formulation. Plümper and Troeger (2007, 2011) show that it depends on the ratio of within to between variation of the RHS variables whether a fixed or random effects formulation produces better (that

is, estimates with lower Root Mean Squared Error) results. Clark and Linzer (2015) focus on the relationship between the number of units and within unit time points. Bell and Jones (2015), in comparison, demonstrate that it is important for the choice of estimator whether a panel is balanced or not.

### ***Parameter Heterogeneity***

We observe parameter heterogeneity if the coefficient of an explanatory variable differs significantly across units or over time. If parameters change across time or units, we are likely to deal with unobserved, and therefore excluded, interaction effects or we have assumed the wrong functional form of the statistical relationship. If the source of parameter heterogeneity is known or our theoretical model even predicts differences in parameters across units or time periods, we can straightforwardly specify the correct model by including interaction terms between time periods or groups of units and the specific right-hand side variables.

In case the source of parameter heterogeneity is unknown, seemingly unrelated regressions (SUR) or random coefficients models (RCM) offer an econometric solution to the problem. SUR models estimate a single regression for every unit but exploit the panel structure of the data by assuming a joint error process for all units. This increases efficiency of estimation by borrowing strength. SUR models only generate acceptable parameter estimates for long time-series – that is, when  $T$  largely exceeds  $N$ . SUR models employ a GLS type estimator for the VC matrix, which weights the standard errors by the cross-section specific mean squared errors. The random coefficients estimator (Beck and Katz, 2007) provides a compromise between estimating the fully pooled model and a fully unpooled estimate (separate OLS for each unit). Pooled OLS depends on the stark assumption of unit homogeneity, whereas separate OLS estimation for each unit produces inefficient results. The RCM borrows strength

by shrinking each of the individual unit OLS estimates back to the overall (pooled) estimate. It is, therefore, also a good test for poolability of the data. The RCM generalizes the RE estimator from the intercept to all parameters of interest:

$$y_{it} = \alpha + \beta_{ki} \sum_{k=1}^K x_{kit} + \delta_m \sum_{m=1}^M w_{mit} + \epsilon_{it} \quad (17)$$

$$\beta_i \sim N(\beta, \sigma_\beta^2)$$

The RCM can be made more useful by allowing the  $\beta_i$  to be functions of other unit-specific variables.

### ***Heterogeneity of Dynamics and Lag Structures***

In pooled data, not only coefficient estimates but also dynamic effects can vary across units. In addition, different RHS variables might exert a differently lagged impact on the dependent variable and the lag length can differ across units. Different dynamics can be straightforwardly incorporated into RCM or SUR models by including an LDV with unit-specific coefficients. Prais–Winsten specifications also allow for unit-specific autoregressive processes in the error term.

Since statistical tests for heterogeneous dynamics or lag structures are not readily available, specific dynamics should be defined on theoretical grounds. Unit-specific dynamics then can be more directly modeled by interacting the relevant regressors or the LDV with dummies for specific groups of units. Since different explanatory variables can have differently lagged effects across units as well, it is not plausible to just vary the estimates of the LDV, because the marginal effect of an RHS variable at time  $t > 1$  partially depends on the estimate for the LDV.

In summary, different kinds of unit heterogeneity do not prevent pooling of information over time and across units. Theoretically, unit heterogeneity leads to interesting research questions which can be empirically analyzed

with the appropriate model specification. The possibility of controlling for unit heterogeneity renders pooled data analysis more attractive than pure cross-section or time series analysis.

### ***Dynamics and Unit Heterogeneity***

Since PCSTS data, by definition, pool information over different dimensions, in this case space and time, the main challenge when using pooled data is dealing with simultaneously occurring violations of the underlying assumptions. Textbooks usually discuss the underlying assumptions of an estimator and, thus, solutions to violations of individual assumptions, separately. However, it is highly unlikely that only one assumption is violated – for example, in pooled data, we usually observe dynamics and unit heterogeneity at the same time. The main problem results from the expectation that solving one specification issue will reduce the overall bias. Unfortunately, this is not the case, and treating one problem might make another problem worse – that is, increase the overall bias.

Plümper and Troeger (2019) demonstrate this theoretically and with Monte Carlo experiments for the simultaneous occurrence of correlated unit-specific effects and dynamic misspecifications. They look at three common dynamic DGPs that all lead to serially correlated errors: omitted time-varying variables, omitted trends, and misspecified lag structures of the RHS variables. They then estimate six different dynamic specifications for each DGP (including an LDV, Prais–Winsten, Period fixed effects, ADL) and include unit fixed effects. They show that if the econometric solution does not match the DGP, a fixed effects approach increases the bias substantially, especially when compared to not including fixed effects. This exercise shows that misspecification issues need to be addressed simultaneously and focusing on one issue (unit heterogeneity) but ignoring another (dynamics) can increase the overall bias.

Testing for dynamics turns out to be especially challenging in pooled data because individual effects, autoregressive errors, and moving average errors can all present as serial correlation, and the sequence of testing for different violations of Gauss–Markov severely affects the conclusions we draw. For example, as shown by Plümpner and Troeger (2019), the Hausman test (1978) for correlated unit-specific effects performs very poorly under different dynamic DGPs and specifications.

As we know, a model with individual effects has composite errors that are serially correlated by definition. The presence of the time-invariant error component gives rise to serial correlation that does not die out over time. As a consequence, standard tests applied to pooled data always reject the null of spherical residuals. As discussed above, dynamic processes in the mean equation that are not specifically modeled, and serial correlation in the error term, can both result in biased estimates. Therefore, researchers need to test for these misspecifications. However, given the likelihood of individual effects in pooled data (whether correlated or not with RHS variables) testing for serial correlation in concurrence with individual effects is indispensable.

Pickup and Troeger (2019) analyze the performance of the most relevant specification tests for pooled data to determine under what conditions researchers can identify whether dynamic misspecifications and unit-specific effects (correlated or not) occur simultaneously. Since unit-specific effects generate serially correlated errors, the sequence of testing for fixed effects and serial correlation and/or the performance of joint tests becomes very important. They look at a large set of dynamic and static DGPs and estimation specifications for pooled data with (correlated) unit-specific effects and test the performance of the Wooldridge (2010) test for serial correlation, the Born–Breitung test (Born and Breitung, 2016) for serial correlation under FE, the Inoue–Solon (2006) test for serial correlation under FE, the Baltagi–Li test (Baltagi and Li, 1995) for serial correlation

under random effects and vice versa, and the Hausman (1978) and Mundlak (1978) tests for correlated unit-specific effects. None of these tests are designed to uncover the source of potential dynamics. The authors show that while most of the tests underperform for a large set of DGPs and estimation specifications, some guidance for applied researchers can be drawn from this exercise:

- 1 The Inoue–Solon test for serial correlation is extremely biased towards rejecting the Null of no serial correlation.
- 2 When the estimation includes a lagged dependent variable, most tests for serial correlation in pooled data are oversized.
- 3 Tests for serial correlation in pooled data perform best when lags of the RHS variables but not the LDV are included.
- 4 The Mundlak test always outperforms the Hausman test.
- 5 Both Hausman and Mundlak tests are biased towards rejecting the Null of uncorrelated unit effects when an LDV is included.
- 6 The Mundlak test performs relatively well when the estimation includes lags of the RHS variables and is more dynamically complete.
- 7 The Baltagi–Li test for random effects works best with more dynamically complete specifications.

In general, as has been discussed previously, a careful, dynamically complete specification is of utmost importance before dealing with other issues such as unit heterogeneity.

## PCSTS DATA WITH LIMITED DEPENDENT VARIABLES

In principle, all typical limited dependent variable models (Binary choice, Count, Tobit, etc.) are also applicable to PCST Data. As for linear models, the pooled data structure adds problems for specification that are sometimes hard to solve. I will briefly discuss specification issues in binary choice models with pooled data. The implications for other limited dependent variable models are similar.

In binary choice models, the presence of unit-specific effects complicates matters significantly, whether they are correlated with explanatory variables or not. As in the linear case, it is possible to add  $N - 1$  unit dummies to the RHS of the estimation equation to control for correlated unit-specific effects, but only when estimating a logistic or logit model. This is called the unconditional fixed effects logit model. Since the normal CDF used in the probit binary choice model has no closed form representation, adding  $N - 1$  dummies will not allow for the model to be identified. The unconditional FE logit estimator is generally inconsistent, but as Katz (2001) demonstrates, the bias is negligible for  $T \geq 20$ .

Another alternative is the conditional logit fixed effects model (Chamberlain, 1980). This estimator uses a conditional likelihood where only units that switched from zero to one or vice versa are used for estimation. This eliminates the unit-specific effect  $u_i$ . Chamberlain derives this procedure for  $T = 2$ . There has been some discussion (Greene, 2004) whether this model can easily be applied to cases where  $T > 2$ . Chamberlain (1980) proposes a solution in the form of maximizing a conditional version of the likelihood function. The intuition is that the  $u_i$  disappear from the likelihood if the likelihood of a given period of peace (i.e., of a given country) is calculated conditioning on the total number of periods (years) of peace for that country. The unit fixed effects are eliminated from the conditional logit likelihood via a transformation that is analogous to first differencing in linear pooled data models. The units for which the outcome is always 0 or always 1 do not contribute to the likelihood. In other words, the information that they provide is not used to estimate  $\beta$ . These units are unaffected by the explanatory factors. However, if, for example, 99% of the sample is in this situation, we may still estimate a significant  $\beta$  using the 1% of the sample which changed outcome during the observation period. No weight would be given to the fact that, for the vast majority

of the sample, the explanatory factors do not affect the outcome. This is similar to the linear case with slowly changing explanatory factors. Finally, this conditional likelihood approach cannot be adopted in the presence of lagged dependent variables.

Conditional fixed effects probit estimation is not feasible because the conditional likelihood approach does not yield computational simplifications for the FE probit. The  $u_i$  cannot be swept away, and maximizing the likelihood over all the parameters including the fixed effects will, in general, lead to inconsistent estimates for large  $N$  and fixed  $T$ .

In case the unit-specific effects remain uncorrelated with the RHS variables, a Random Effects Probit model can be estimated. MLE, in this case, yields a consistent and efficient estimator of  $\beta$ , but MLE is computationally costly since one has to compute the joint probabilities of a T-variate normal distribution, which involves T-dimensional integrals. This, however, becomes infeasible if  $T$  grows very large. However, it is possible to reduce the computation to a single integral.

There has been a lot of recent research focusing on fixed effects in binary choice and rare events data. Most methodologists dealing with pooled binary TSCS data agree that unobserved unit-specific effects generate a problem for estimation, but there is no clear consensus on how to deal with this problem. As discussed above, it is commonly believed that one of the major problems with rare events data is the fact that estimating a conditional fixed effects model generates inefficient estimates. Cook et al. (2020) revisit this issue and demonstrate that the main concern with fixed effects models of rare events data is not inefficiency, but rather biased estimation of marginal effects. They argue that only evaluating switching units generates a biased estimate of the baseline risk and thus incorrect estimates of the effects. The authors propose a penalized maximum likelihood fixed effects (PML-FE) estimator, which retains the complete sample by providing finite estimates of the fixed effects for each unit.

As with linear models, serially correlated errors violate the independence assumption of most MLE, Logit, and Probit models. Therefore, serially correlated errors encounter similar problems as in the linear case. Chamberlain (1993) shows that Binary Choice Models using a Lagged Dependent Variable are not identified. Beck et al. (1998) argue that Binary Choice PCSTS data are grouped duration data and suggest adding a series of period dummies to account for time dependency. These dummies account for the time elapsed since the last failure (1):  $k_t - t_0$ . This is equal to assuming duration dependence in a hazard model and measuring the length of non-eventful binary spells. Thus corrected, the logit resembles a BTSCS event history model. One can use a Likelihood Ratio test to check whether all included period dummies are jointly zero. This means that if  $T$  grows large, there will be many dummy variables, which could pose a degrees of freedom problem. In this case, period dummies can be transformed into cubic splines instead. The researcher has to specify a number of knots that define what segments of the time variable will have a cubic polynomial fit to it. These splines can be interpreted as hazard rates: the estimated coefficients measure the effect of the calculated base vector on the probability of and outcome, for example, war. Carter and Signorino (2010) advocate the use of the exponential of the time variable instead ( $t, t^2, t^3$ ). They argue that important elements of the splines are ad hoc and have no theoretical justification, which could lead to bias. Exponentials of  $t$  are more readily interpretable.

## CONCLUSION

This entry gives a short overview of basic estimation procedures and specification issues in PTSCS data. Due to space constraints, other important topics such as spatial effects in pooled data, non-stationarity, and the usefulness of error correction models, as well as

a detailed discussion of estimation procedures and specification issues in limited dependent variable models for pooled data, have to remain untouched. While estimators and statistical tests are discussed in most econometric textbooks, a thorough discussion of specification problems in pooled and panel data remains important. This is also true for the question of how asymptotic properties of estimators translate to finite sample analysis. There have been many recent exciting developments in PTSCS data analysis in the social sciences. For example, much has been done to properly model spatial effects and spatio-temporal effects both in linear (Franzese and Hays, 2007; Franzese et al., 2010) and limited dependent variable models (Franzese et al., 2016) for pooled data. In addition, the identification revolution in empirical social sciences has influenced the analysis of pooled data. Blackwell and Glynn (2018), for example, use potential outcomes to define causal quantities of interest in dynamic PTSCS models, and clarify how standard models like the ADL can generate biased estimates of these quantities because of post-treatment conditioning. They propose inverse probability weighting and structural nested mean models to deal with these post-treatment biases. The major challenge to the analysis of data that combines different dimensions – for example, pooling cross-sectional and time-series information – remains the treatment of simultaneously occurring misspecifications, because existing tests cannot discriminate effectively between different sources of misspecifications.

## REFERENCES

- Amemiya, Takeshi and Thomas E. MaCurdy 1986. Instrumental-Variable Estimation of an Error-Components Model. *Econometrica* 54: 869–881.
- Anderson, T.W. and Cheng Hsiao 1981. Estimation of Dynamic Models with Error Components. *Journal of the American Statistical Association* 76: 598–606.

- Angrist, Joshua D. and Joern-Steffen Pischke 2009. *Mostly Harmless Econometrics*. Princeton, NJ: Princeton University Press.
- Arellano, Manuel and Stephen Bond 1991. Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations. *Review of Economic Studies* 58: 277–297.
- Baltagi B.H., Li Q. 1995. Testing AR(1) Against MA(1) Disturbances in an Error Component Model. *Journal of Econometrics* 68: 133–151.
- Beck, Nathaniel 2001. Time-Series-Cross-Section Data: What Have We Learned in the Past Few Years? *Annual Review of Political Science* 4: 271–293.
- Beck, Nathaniel and Jonathan Katz 1995. What to do (and not to do) with Time-Series Cross-Section Data. *American Political Science Review* 89: 634–647.
- Beck, Nathaniel and Jonathan N. Katz 2007. Random Coefficient Models for Time-Series-Cross-Section Data: Monte Carlo Experiments. *Political Analysis* 15: 182–195.
- Beck, Nathaniel, Jonathan N. Katz and Richard Tucker 1998. Taking Time Seriously: Time-Series-Cross-Section Analysis with a Binary Dependent Variable. *American Journal of Political Science* 42(4): 1260–1288.
- Bell, Andrew and Kelvyn Jones 2015. Explaining Fixed Effects: Random Effects Modeling of Time-Series Cross-Sectional and Panel Data. *Political Science Research and Methods* 3(1): 133–153.
- Born, Benjamin and Joerg Breitung 2016. Testing for Serial Correlation in Fixed-Effects Panel Data Models. *Econometric Reviews* 35(7): 1290–1316.
- Carter, David B. and Curtis S. Signorino 2010. Back to the Future: Modeling Time Dependence in Binary Data. *Political Analysis* 18 (3): 271–292.
- Chamberlain, G. 1980. Analysis of Covariance with Qualitative Data. *The Review of Economic Studies* 47: 225–238.
- Chamberlain, G. 1993. Feedback in Panel Data Models. Harvard Institute of Economic Research Working Papers 1656, Harvard – Institute of Economic Research.
- Clark Tom S. and Drew A. Linzer 2015. Should I Use Fixed or Random Effects? *Political Science Research and Methods* 3(2): 399–408.
- Cook, Scott J., Jude C. Hays and Robert J. Franzese 2020. Fixed effects in rare events data: a penalized maximum likelihood solution. *Political Science Research and Methods* 8(1): 92–105.
- DeBoef, Suzanna and Luke J. Keele 2008. Taking Time Seriously: Dynamic Regression. *American Journal of Political Science* 52(1): 184–200.
- Franzese, Robert J. 2003a. Multiple Hands on the Wheel: Empirically Modeling Partial Delegation and Shared Policy Control in the Open and Institutionalized Economy. *Political Analysis*: 445–474.
- Franzese, Robert J. 2003b. Quantitative Empirical Methods and the Context-Conditionality of Classic and Modern Comparative Politics. *CP: Newsletter of the Comparative Politics Organized Section of the American Political Science Association* 14(1): 20–24.
- Franzese, Robert J. and Cindy Kam 2009. *Modeling and Interpreting Interactive Hypotheses in Regression Analysis*. Ann Arbor, MI: University of Michigan Press.
- Franzese, Robert J. and Jude C. Hays 2007. Spatial-Econometric Models of Cross-Sectional Interdependence in Political-Science Panel and Time-Series-Cross-Section Data. *Political Analysis* 15(2): 140–164.
- Franzese, Robert J., Jude C. Hays and Aya Kachi 2010. A Spatial Model Incorporating Dynamic, Endogenous Network Interdependence: A Political Science Application. *Statistical Methodology* 7(3): 406–428.
- Franzese, Robert J., Jude C. Hays and Scott J. Cook 2016. Spatial- and Spatiotemporal-Autoregressive Probit Models of Interdependent Binary Outcomes. *Political Science Research and Methods* 4(1): 151–173.
- Greene, William 2004. The Behaviour of the Maximum Likelihood Estimator of Limited Dependent Variable Models in the Presence of Fixed Effects. *The Econometrics Journal* 7: 98–119.
- Hausman, Jerry A. 1978. Specification Tests in Econometrics. *Econometrica* 46(6): 1251–71.
- Hausman, Jerry A. and William E. Taylor 1981. Panel Data and Unobservable Individual Effects. *Econometrica* 49: 1377–1398.
- Hsiao, C., M.H. Pesaran and A.K. Tahmiscioglu 2002. Maximum Likelihood Estimation of

- Fixed Effects Dynamic Panel Data Models Covering Short Time Periods. *Journal of Econometrics* 109:107–150.
- Inoue, Atsushi and Gary Solon 2006. A Portmanteau Test for Serially Correlated Errors in Fixed Effects Models. *Econometric Theory* 22(5): 835–851.
- Katz, Ethan 2001. Bias in Conditional and Unconditional Fixed Effects Logit Estimation. *Political Analysis* 9(4): 379–384.
- Keele, L., Linn, S. and C.M. Webb 2016. Treating Time with all Due Seriousness. *Political Analysis* 24(1): 31–41.
- Kiviet, J.F. 1995. On Bias, Inconsistency, and Efficiency of Various Estimators in Dynamic Panel Data Models. *Journal of Econometrics* 68: 53–78.
- Kmenta, J. 1986. *Elements of Econometrics*. New York: Macmillan Publishing Company.
- Lancaster, T. 2000. The Incidental Parameter Problem Since 1948. *Journal of Econometrics* 95: 391–413.
- Mundlak, Yair 1978. On the Pooling of Time Series and Cross Section Data. *Econometrica* 46: 69–85.
- Nickell, Stephen 1981. Biases in Dynamic Models with Fixed Effects. *Econometrica* 49(6): 1417–1426.
- Parks, Richard W. 1967. Efficient Estimation of a System of Regression Equations when Disturbances are Both Serially and Contemporaneously Correlated. *Journal of the American Statistical Association* 62(318): 500–509.
- Pickup, Mark 2018. *A General-to-Specific Approach to Dynamic Panel Models with a Very Small T*. Presented to the 2017 Meeting of the Midwest Political Science Association, Chicago, IL.
- Pickup, Mark and Vera E. Troeger 2019. *Specifying Dynamic Processes in Panel Data*. Presented to the 2018 Meeting of the American Political Science Association, Boston, MA.
- Plümper, Thomas and Vera E. Troeger 2007. Efficient Estimation of Time-Invariant and Rarely Changing Variables in Finite Sample Panel Analyses with Unit Fixed Effects. *Political Analysis* 15: 124–139.
- Plümper, Thomas and Vera E. Troeger 2011. Fixed-effects vector decomposition: properties, reliability, and instruments. *Political Analysis* 19(2): 147–164.
- Plümper, Thomas and Vera E. Troeger 2019. Not so Harmless After All: The Fixed-Effects Model. *Political Analysis* 27(1): 21–45.
- Plümper, Thomas, Troeger, Vera E. and Philip Manow 2005. Panel Data Analysis in Comparative Politics. Linking Method to Theory. *European Journal of Political Research* 44: 327–354.
- Wilkins, Arjun S. 2018. To Lag or Not to Lag?: Re-Evaluating the Use of Lagged Dependent Variables in Regression Analysis. *Political Science Research and Methods* 6(2): 393–411.
- Williams, Laron K. and Guy D. Whitten 2011. Dynamic Simulation of Autoregressive Relationships. *The Stata Journal* 11(4): 577–588.
- Wooldridge, Jeffrey M. 2010. *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.





# Dynamic Systems of Equations

Mark Pickup

This chapter explores dynamic systems of equations in three parts. First, the topic is motivated by a discussion of why a researcher might use a dynamic model and might go beyond a single-equation model to a system of equations. This discussion is structured in terms of the violations of exogeneity that such models are designed to resolve. Second, the chapter describes common multiple-equation dynamic models: structural vector autoregression (SVAR), reduced form vector autoregression (VAR) and vector error-correction (VEC). Third, the chapter describes the state-space approach to dynamic modelling, showing that it encompasses these models and additional models, such as dynamic-factor and dynamic seemingly unrelated regression (SUR) equations.

## **Dynamics**

Giving the parameter estimates of any model a causal interpretation requires some form of

exogeneity assumption. For example, if the model were:

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t,$$

a strict exogeneity assumption would be:

$$E(\varepsilon_t | x_{t+h}) = 0, \forall h$$

and a contemporaneous exogeneity assumption would be:

$$E(\varepsilon_t | x_t) = 0.$$

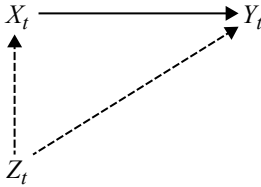
These are examples of the zero-conditional-mean assumption, and when it is violated we say that  $x_t$  is endogenous for the estimation of  $\beta_1$  in the model. Most readers will be familiar with the particular violation of this assumption that causes omitted variable bias. Let us say we are interested in testing the causal effect of  $X_t$  on  $Y_t$ :

$$X_t \longrightarrow Y_t$$

and our data model is:

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t,$$

but the data generating process is:

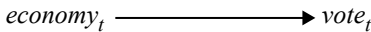


$Z_t$  is an omitted variable if it is causally prior to and correlated with current values of  $X_t$  and  $Y_t$ . (Note:  $Z_t$  does not have to cause  $X_t$ ; it is enough that it is correlated.) If this is the case,  $E(\varepsilon_t | X_t) \neq 0$ , and the consequence is that the standard ordinary least squares (OLS) estimate of  $\beta_1$  will be a biased estimate of the causal effect of  $X_t$  on  $Y_t$ <sup>1</sup>:

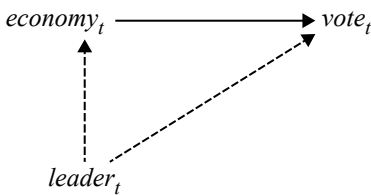
$$E(\hat{\beta}_1 | X_t) = \beta_1 + X_t' X_t^{-1} X_t' \eta \neq \beta_1$$

where  $\eta = E(\varepsilon_t | X_t)$ .

An omitted variable  $Z_t$  can be of different types, resulting from different data-generating processes (DGP). As an example, say that  $Y_t$  is an aggregate measure of government vote intention and that  $X_t$  is an aggregate measure of subjective evaluations of how the economy has performed:



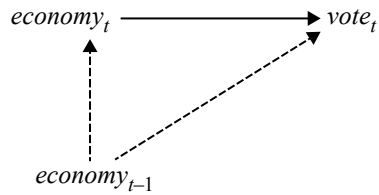
but both current government vote intention and current economic evaluations are a function of current leadership evaluations:



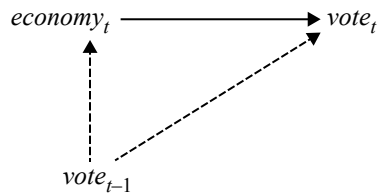
The omission of leadership evaluations from our data model will result in a biased estimate of

the relationship between current economic evaluations and current government vote intention.

Another possibility is that the omitted variable is a past value of ( $X_{t-1}$ ) that is causally prior to and correlated with current values of  $X_t$  and  $Y_t$ . For example, both current government vote intention and current economic evaluations are a function of past economic evaluations, resulting in a biased estimate of the relationship between current economic evaluations and current government approval:



A third possibility is that  $Z_t$  is a past value of  $Y_t$  that is causally prior to and correlated with current values of  $X_t$  and  $Y_t$ . For example, both current government vote intention and current economic evaluations are a function of past government vote intention:



The omission of past vote intention will result in a biased estimate of the causal relationship between current economic evaluations and current government approval.

In practice, we can resolve these endogeneity problems by including the omitted variable.<sup>2</sup> In addition to specifying which variables to include, it also means specifying which lags of  $X_t$  and  $Y_t$  to include (the lag structure).

Any model that includes a lag of  $Y_t$  (i.e., a lagged dependent variable) is known as a dynamic model. A common dynamic model is the autoregressive distributed lag (ADL) model. With a single  $x_t$ , this model is:

$$y_t = \alpha_0 + \sum_{j=1}^p \alpha_j y_{t-j} + \beta_1 x_t + \sum_{i=1}^m \beta_{i+1} x_{t-i} + \varepsilon_t \tag{1}$$

with  $\varepsilon_t \sim NID(0, \sigma_\varepsilon^2)$

The ADL(p,m) includes  $p$  lags of the dependent variable (included as independent variables) and  $m$  lags of the other independent variable(s), in addition to the non-lagged independent variable(s) (see Troeger, Chapter 33, this *Handbook*; Cook et al., Chapter 39, this *Handbook*; Neumayer and Plümper, Chapter 38, this *Handbook*).

An alternative to the ADL is the autoregressive moving-average model (ARMA). It models dynamics in  $Y_t$  by combining an autoregressive AR( $P$ ) process with a moving-average MA( $Q$ ) process:

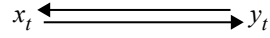
$$y_t = \alpha_0 + \frac{\sum_{i=1}^p \alpha_i y_{t-i}}{AR(P)} + \frac{\sum_{j=0}^q \phi_j \varepsilon_{t-j}}{MA(Q)} \tag{2}$$

An ARMA model with  $P$  lags of  $Y_t$  and a  $Q$  lag of  $\varepsilon_t$  is denoted ARMA( $P, Q$ ). Box et al. (1994) developed an approach to determine the optimal values for  $P$  and  $Q$ . The lag structure of the  $X_t$  to be included in the model are then determined by a second procedure developed by Box et al. (1994). The procedures these researchers developed have minimal prediction errors and parsimonious models as their goals. Meanwhile, ADL models are more common when estimating causal effects are the goal and it is believed that a data model can be specified that has a lag structure that encompasses that of the DGP.<sup>3</sup> Either way, the use of these and other dynamic models is motivated by the need to address violations of exogeneity caused by not accounting for the dynamics in the DGP.

**Multiple Equations**

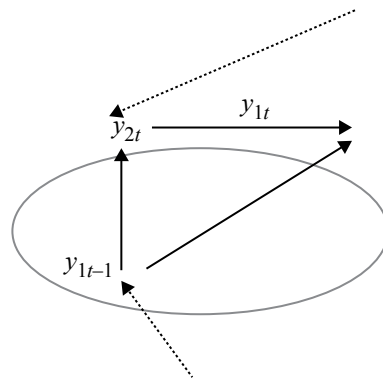
Another way the zero-conditional-mean assumption (i.e., exogeneity) may be violated is through simultaneity. This occurs when  $X_t$  and  $Y_t$  concurrently have causal

effects on each other, not because of their relationship with any other variable but simply because they each cause the other at the same point in time:



In terms of our economic voting example, we may expect one’s vote intention to influence one’s evaluation of the economy. A Republican voter under a Republican incumbent president may evaluate the economy differently to someone who intends to vote for the Democratic candidate, because viewing the economy as performing well feels consistent with their intention to vote for the Republican incumbent. Other examples of potentially simultaneous relationships include defence expenditures in country A and defence expenditures in country B; the relationship between foreign policy and presidential popularity; and public perception of a political event and the media’s coverage of the event.

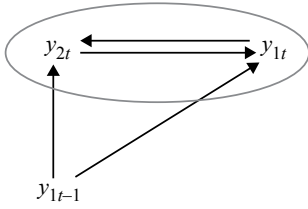
The possibility of simultaneity requires us to consider a model that not only includes an equation for  $Y_t$  but also for  $X_t$ ; in other words, a system of equations, known as a multi-equation model. For example, we have discussed the following endogeneity problem created by omitting a lag of the dependent variable, where we previously denoted  $y_{1t}$  as  $y_t$  and  $y_{2t}$  as  $x_t$ . We are interested in this:



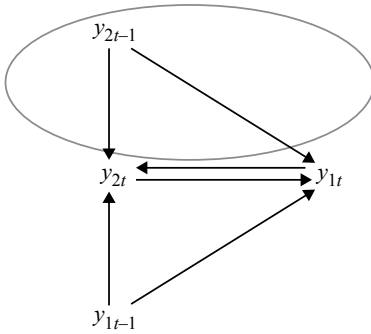
and we need to take this into account.

We have also noted that the following will produce an endogeneity issue. In addition to

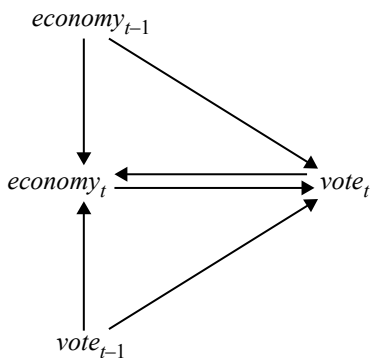
$y_{2t}$  contemporaneously causing  $y_{1t}$ ,  $y_{1t}$  also contemporaneously causes  $y_{2t}$ ;



The following will also produce an endogeneity issue. In addition to the bidirectional, contemporaneous relationship between  $y_{1t}$  and  $y_{2t}$ , past values of  $y_{2t}$  predict  $y_{1t}$  and  $y_{2t}$ ;



Returning to our economic voting example, in addition to past economic evaluations and past vote intentions being causally prior to current vote intentions, we might expect vote intentions (current and past) to have a causal effect on one's evaluation of the economy's performance and past economic performance to be causally prior to current economic performance and vote intention:



These potential relationships are captured by the following system of equations:

$$\begin{aligned} y_{1t} &= \alpha_1 + \alpha_{10}y_{2t} + \gamma_{11}y_{1t-1} + \gamma_{12}y_{2t-1} + \mu_{1t} \\ y_{2t} &= \alpha_2 + \alpha_{20}y_{1t} + \gamma_{21}y_{1t-1} + \gamma_{22}y_{2t-1} + \mu_{2t} \end{aligned} \quad (3)$$

$$\text{where } \mu_{1t} \sim \mu_{2t} \left( 00, \begin{pmatrix} \sigma_{11}^2 & \sigma_{12} \\ \sigma_{12} & \sigma_{22}^2 \end{pmatrix} \right)$$

Note that no assumption is made about the independence of the errors. The covariance between the errors in the two equations is allowed to be non-zero:  $\sigma_{12}$ .

This is a structural model that captures the expected causal effects between the variables.<sup>4</sup> This particular structural model includes an equation for each endogenous variable, and each equation includes lags of each of the endogenous variables. This is known as a SVAR model. Before we discuss SVAR models, we will discuss the reduced form VAR model.

### REDUCED FORM VAR

Because of the contemporaneous relationships between  $y_{1t}$  and  $y_{2t}$  in (3), it cannot be estimated simply. Unless  $\alpha_{20} = 0$ , the OLS estimate of  $\alpha_{10}$  in the first equation will not be consistent because of the simultaneity and violation of the zero-conditional-mean assumption:  $E(\mu_{1t}|y_{2t}) \neq 0$ .<sup>5</sup> Equivalently, unless  $\alpha_{10} = 0$ , the OLS estimate of  $\alpha_{20}$  in the second equation will not be consistent:  $E(\mu_{2t}|y_{1t}) \neq 0$ . We can get consistent estimates if we can apply restrictions, such as  $\alpha_{10} = 0$ , but these restrictions must be valid. This is equivalent to knowing something about the causal relationships between the endogenous variables *a priori*. For example, knowing that economic evaluations have a contemporaneous effect on vote intention but vote intention does not have a contemporaneous effect on economic evaluations (although it

may have a lagged effect). If we do not wish to make such identifying restrictions, we can do the following.

The system of equations in (3) can be transformed through a straightforward use of algebra (Brandt and Williams, 2007: 16–17). We substitute the equation for  $y_{2t}$  into  $y_{1t}$  and solve for  $y_{1t}$ . Conversely, we substitute the equation for  $y_{1t}$  into  $y_{2t}$  and solve for  $y_{2t}$ . The result is the following system of equations, called a *reduced form VAR* model:

$$\begin{aligned} y_{1t} &= \beta_{10} + \beta_{11}y_{1t-1} + \beta_{12}y_{2t-1} + \varepsilon_{1t} \\ y_{2t} &= \beta_{20} + \beta_{21}y_{1t-1} + \beta_{22}y_{2t-1} + \varepsilon_{2t} \end{aligned}$$

$$\begin{aligned} \text{where: } \beta_{11} &= \frac{\alpha_{10}\gamma_{21} + \gamma_{11}}{(1 - \alpha_{10}\alpha_{20})}, \beta_{12} \\ &= \frac{\alpha_{10}\gamma_{22} + \gamma_{12}}{(1 - \alpha_{10}\alpha_{20})}, \beta_{21} \\ &= \frac{\alpha_{20}\gamma_{11} + \gamma_{21}}{(1 - \alpha_{10}\alpha_{20})}, \end{aligned}$$

$$\begin{aligned} \beta_{22} &= \frac{\alpha_{20}\gamma_{12} + \gamma_{22}}{(1 - \alpha_{10}\alpha_{20})}, \varepsilon_{1t} \\ &= \frac{\alpha_{10}\mu_{2t} + \mu_{1t}}{(1 - \alpha_{10}\alpha_{20})}, \text{ and } \varepsilon_{2t} \\ &= \frac{\alpha_{20}\mu_{1t} + \mu_{2t}}{(1 - \alpha_{10}\alpha_{20})}. \end{aligned}$$

In the reduced form VAR model, each variable is regressed on its past values and the past values of the other variables in the system (Freeman et al., 1989). The reduced form VAR model can include additional lags of the variables ( $P$  of them):

$$\begin{aligned} y_{1t} &= \beta_{10} + \sum_{i=1}^P \beta_{1i}y_{1t-i} \\ &\quad + \sum_{i=1}^P \beta_{1(p+i)}y_{2t-i} + \varepsilon_{1t} \\ y_{2t} &= \beta_{20} + \sum_{i=1}^P \beta_{2i}y_{1t-i} \\ &\quad + \sum_{i=1}^P \beta_{2(p+i)}y_{2t-i} + \varepsilon_{2t} \end{aligned} \quad (4)$$

This can be extended even further to allow for additional endogenous variables  $Y_{kt}$ ;  $\Omega$ one additional equation is required for each additional endogenous variable. Exogenous variables can also be included (this does not require additional equations):

$$\begin{aligned} y_{1t} &= \beta_{10} + \sum_{i=1}^P \beta_{1i}y_{1t-i} \\ &\quad + \sum_{i=1}^P \beta_{1(p+i)}y_{2t-i} + \\ &\quad + \sum_{i=1}^K \kappa_{1(k)}x_{k,t} + \varepsilon_{1t} \\ y_{2t} &= \beta_{20} + \sum_{i=1}^P \beta_{2i}y_{1t-i} \\ &\quad + \sum_{i=1}^P \beta_{2(p+i)}y_{2t-i} \\ &\quad + \sum_{i=1}^K \kappa_{2(k)}x_{k,t} + \varepsilon_{2t} \end{aligned} \quad (5)$$

Compared with the SVAR model, an advantage of the reduced form VAR model is that the system of equations can be estimated (by OLS or maximum likelihood estimation) without making any restricting assumptions. The disadvantage is that the parameters in the reduced form model do not tell us anything about the causal contemporaneous relationships between the endogenous variables ( $\alpha_{10}$  and  $\alpha_{20}$ ). In fact, the parameters do not tell us anything *directly* about the non-contemporaneous causal relationships (the  $\gamma_{1t}$ s and  $\gamma_{2t}$ s). These cannot be calculated from the estimated parameters (the  $\beta$ s and  $\varepsilon_t$ s), unless we can make some identifying restriction – more on this to come.

The reduced form VAR does allow us to estimate how one endogenous variable will respond to a random shock to another endogenous variable. This response is not causal but predictive. This is achieved by transforming the reduced form VAR to what is known as the vector moving-average (VMA) representation. For a VAR with  $K$  endogenous variables and  $P$  lags, this is:

$$\begin{aligned}
 y_t - \mu_t &= \sum_{i=0}^{\infty} \Phi_i \varepsilon_{t-i} \\
 y_t &\equiv \begin{pmatrix} y_{1t} \\ \vdots \\ y_{Kt} \end{pmatrix} \quad \mu_t \equiv \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_K \end{pmatrix} \\
 \varepsilon_t &\equiv \begin{pmatrix} \varepsilon_{1t} \\ \vdots \\ \varepsilon_{Kt} \end{pmatrix} \quad \Sigma_{\varepsilon_t} \equiv \begin{pmatrix} \sigma_{11}^2 & \cdots & \sigma_{1K}^2 \\ \vdots & \ddots & \vdots \\ \sigma_{K1}^2 & \cdots & \sigma_{KK}^2 \end{pmatrix}, \\
 \Phi_i &\equiv \begin{cases} I_k & \text{if } i = 0 \\ \sum_{j=1}^i \Phi_{i-j} \beta_j & \text{if } i > 0 \end{cases}
 \end{aligned} \tag{6}$$

where  $\beta_j$  is a  $K \times K$  matrix of parameters for the  $j$ th lag of  $y_t$ , and  $\beta_j = 0 \forall j > p$ .

The  $\Phi_i$  (which is  $K \times K$ ) tells us how the endogenous variables respond to a one-time, unit increase in the innovations – the  $\varepsilon_t$ . Specifically, the  $j, k$  element of  $\Phi_1$  gives the magnitude of the effect of a one-time, unit increase (an impulse) in the innovation of the  $k$ th endogenous variable on the  $j$ th endogenous variable after 1 periods. For example, the 1, 2 element of  $\Phi_1$  gives the magnitude of the effect of an impulse in  $\varepsilon_{2t-1}$  on  $Y_{1t}$  after 1 periods. The plot of the effects (the elements of the  $\Phi_i$ ) over time is called an impulse-response function (IRF). For example, plotting the magnitudes of the 1, 2 elements of  $\Phi_1, \Phi_2, \Phi_3 \dots$  plots the response of  $Y_{1t}$  to a one-unit impulse in  $\varepsilon_{2t}$ . This gives us a visual representation of the response of  $Y_{1t}$  to changes in  $Y_{2t}$ .

Note the variance–covariance matrix for the innovations ( $\Sigma_{\varepsilon_t}$ ) is not diagonal. In other words, the innovations correlate. This means that the elements of the  $\Phi_i$  cannot be given a causal interpretation. To do so, we would have to interpret the elements of the  $\Phi_i$  as the responses to an impulse in the innovation of one endogenous variable *holding all else*

*constant*. The correlation in the innovations means that all else is not necessarily held constant. In (6), the innovation in one endogenous variable may correlate with the innovation in another.

To provide an empirical example of a reduced form VAR model, we examine monthly prime ministerial (PM) satisfaction ratings, unemployment rates and vote intention data from the UK during the period June 1997 to May 2006 (Pickup, 2010). We are interested in how PM satisfaction and unemployment predict vote intention, taking into account that PM satisfaction and vote intention are likely endogenous to each other. To do this we estimate a reduced form VAR(1) model with PM satisfaction and vote intention as endogenous variables and unemployment rates as exogenous.<sup>6</sup> The unemployment variable is the three-month change in the unemployment rate (lagged by two months).<sup>7</sup> To account for electoral cycle effects, we include an independent trend for each electoral cycle and cycle specific intercepts. The parameter estimates from the model are presented in Table 34.1.

The parameter on lagged PM satisfaction in the vote equation is statistically significant (a 0.05 significance level is used throughout the chapter). PM satisfaction predicts vote intention. However, the reverse is not true. We also see that the unemployment-rate parameter is statistically significant in the PM satisfaction equation but not in the vote intention equation. An increase in the unemployment rate of one percentage point is predicted to cause a decrease in PM satisfaction of almost nine percentage points. While there is no evidence that vote intention responds directly to a change in the unemployment rate, the fact that PM satisfaction predicts vote intention suggests an indirect effect. However, with a reduced form model, we cannot make any causal claims regarding this effect.

**Table 34.1 PM satisfaction, vote intention and unemployment, Britain 1997–2006**

Equation		Parameter estimates
PMsat	PMsat <sub>t-1</sub>	0.739 (0.076)**
	Vote <sub>t-1</sub>	0.243 (0.165)
	Trend 1	-0.086 (0.048)
	Trend 2	-0.022 (0.064)
	Trend 3	-0.496 (0.379)
	Cycle 2	-4.054 (1.705)*
	Cycle 3	-2.889 (2.980)
	Unemployment	-8.703 (2.979)**
	Constant	3.836 (6.312)
	Vote	PMsat <sub>t-1</sub>
Vote <sub>t-1</sub>		0.204 (0.110)
Trend 1		-0.048 (0.032)
Trend 2		-0.170 (0.043)**
Trend 3		0.057 (0.253)
Cycle 2		0.659 (1.139)
Cycle 3		-5.385 (1.991)**
Unemployment		0.194 (1.990)
Constant		30.866 (4.217)**
<i>T</i>		101

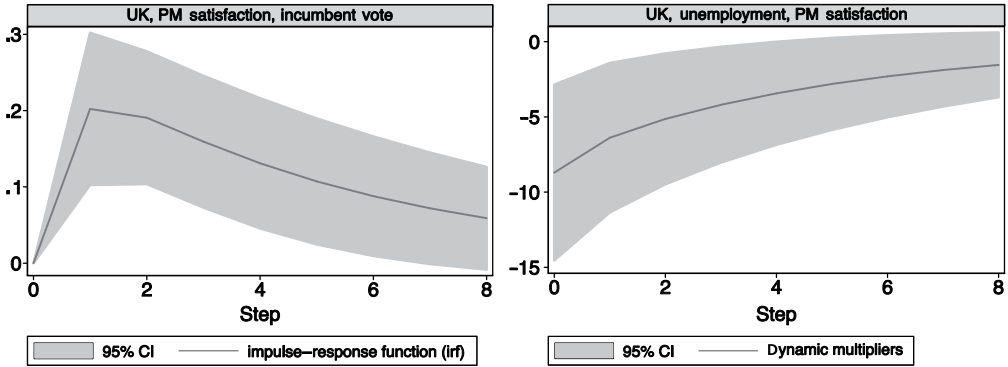
\*  $p < 0.05$ ; \*\*  $p < 0.01$ ; standard errors in parentheses.

With the parameter estimates from the reduced form VAR model in hand, we can plot the impulse–response function. This is a plot of the predicted change in one endogenous variable due to an impulse in the error (innovation) term of the other endogenous variable. This impulse occurs for only one period, after which it returns to zero. As a result, the response to the impulse will decay over time. The left pane of Figure 34.1 shows the impulse–response functions when PM satisfaction is the impulse variable and vote intention is the response variable. Each step is a month, so we see how the response to an impulse in PM satisfaction decays over eight months. Note that the response at time zero is zero. This is because we do not estimate contemporaneous effects/responses in the reduced form model. The right pane of Figure 34.1 is a dynamic multiplier function. This is the effect of a temporary one-unit increase in the unemployment rate on PM satisfaction.

We can see how the effect is initially negative and then decays over an eight-month period.

The advantage of the reduced form VAR is that it is identified, so the parameters can be estimated. The disadvantage is that the parameters of the original structural VAR cannot be calculated from the parameter estimates of the reduced form VAR without making some sort of identifying assumptions. These identifying restrictions amount to making some sort of causal assumptions. Without these assumptions, we are limited to what is known as Granger causality:  $Y_{1t}$  Granger causes  $Y_{2t}$  if past values of  $Y_{1t}$  and  $Y_{2t}$  can better predict the behaviour of  $Y_{2t}$  than past values of  $Y_{2t}$  alone. Put another way, does any individual lag or any combination of lags of  $Y_{1t}$  predict  $Y_{2t}$ , after we control for past values of  $Y_{2t}$ ?

Consider the reduced form VAR as it is expressed in (5). We test Granger causality by defining the null hypothesis that  $Y_{1t}$  does



Graphs by irfname, impulse variable and response variable

Graphs by irfname, impulse variable and response variable

**Figure 34.1 Impulse response and dynamic multiplier functions**

not Granger cause  $Y_{2t}$  as:  $\beta_{2(1)} = \beta_{2(2)} = \dots = \beta_{2(P)} = 0$ . The alternative is that  $Y_{1t}$  does Granger cause  $Y_{2t}$  if:  $\beta_{2(1)} \neq 0, \beta_{2(2)} \neq 0, \dots,$  or  $\beta_{2(P)} \neq 0$ .

If we are willing to make causal assumptions in order to place identifying restrictions, we may be able to ‘back out’ the parameters of the structural form of the model from the reduced form VAR. These SVAR parameters have a causal interpretation. There are two types of assumptions typically made. Depending on the type of assumptions, we call the analysis a short-run SVAR or a long-run SVAR.

**SVAR**

We begin with the short-run SVAR. The reduced form VAR with  $K$  endogenous variables and a lag order of  $P$  can be rewritten in matrix notation as:

$$y_t = b_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \dots + \beta_P y_{t-P} + \varepsilon_t$$

$$y_t = \begin{pmatrix} y_{1t} \\ \vdots \\ y_{Kt} \end{pmatrix}; \quad b_0 = \begin{pmatrix} b_{0,1} \\ \vdots \\ b_{0,K} \end{pmatrix};$$

$$\varepsilon_t = \begin{pmatrix} \varepsilon_{1t} \\ \vdots \\ \varepsilon_{Kt} \end{pmatrix}$$

$$\beta_p = \begin{pmatrix} \beta_{p,1,1} & \dots & \beta_{p,1,K} \\ \vdots & \ddots & \vdots \\ \beta_{p,K,1} & \dots & \beta_{p,K,K} \end{pmatrix}$$

Multiplying both sides by  $K \times K$  matrix  $A$ , this can then be rewritten as:

$$A(y_t - b_0 - \beta_1 y_{t-1} - \beta_2 y_{t-2} - \dots - \beta_P y_{t-P}) = A\varepsilon_t.$$

We can then define a new set of innovations  $e_t$  and  $K \times K$  matrix  $B$ , such that:

$$Be_t \equiv A\varepsilon_t.$$

The VAR can then be written as:

$$A(y_t - b_0 - \beta_1 y_{t-1} - \beta_2 y_{t-2} - \dots - \beta_P y_{t-P}) = Be_t$$

It can be shown that there is a set of values for the elements of  $A$  and  $B$ , such that the  $e_t$  are not correlated across equations. This means  $B^{-1}A$  orthogonalize the innovations  $\varepsilon_t$ . Defining  $P^{-1} = B^{-1}A$ , the VMA



representation of the reduced form VAR can be rewritten:

$$y_t - \mu_t = \sum_{i=0}^{\infty} \Phi_i \varepsilon_t = \sum_{i=0}^{\infty} \Phi_i P P^{-1} \varepsilon_t = \sum_{i=0}^{\infty} \Theta_i e_t$$

$$\Theta_i = \Phi_i P e_t = P^{-1} \varepsilon_t$$

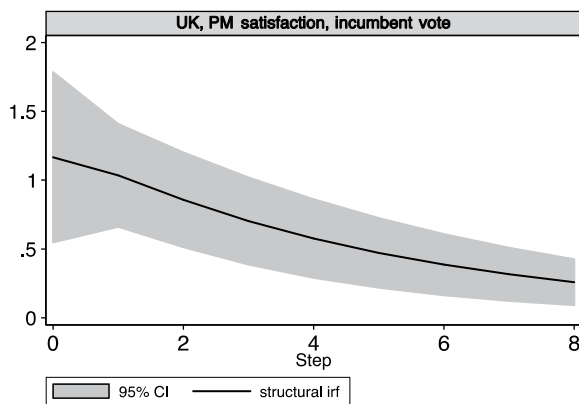
Note that because the innovations are orthogonal, the elements of  $\Theta_i$  give the causal effects of the endogenous variables on each other.

Unfortunately, just as the SVAR is not identified, neither is  $P^{-1}$ . We proceed by placing enough restrictions on  $A$  and  $B$ , so that  $P^{-1}$  is identified. The parameters in  $B$  are the variances of the error terms. The parameters in  $A$  are the (negative of the) contemporaneous effects of the endogenous variables. Placing restrictions on  $A$  then means making assumptions about the contemporaneous causal relationships. The  $i, j$  th element of  $A$  determines how  $Y_{it}$  is contemporaneously caused by  $Y_{jt}$  – i.e., the contemporaneous causal effect of  $Y_{jt}$  on  $Y_{it}$ . A restriction could be that the contemporaneous causal effect of  $Y_{jt}$  on  $Y_{it}$  is 0.<sup>8</sup> It is important to note that any restrictions that identify  $A$  and  $B$  will correspond to restrictions that would have identified the original SVAR model.

Using this procedure means first estimating the reduced form VAR, then estimating causal effects by placing identifying restrictions on  $A$  and  $B$ , so that we can estimate the other elements of these matrices. The estimates of these other elements of  $A$  are estimates of the contemporaneous causal effects not assumed to be known (e.g., to be zero), and the estimated  $A$  and  $B$  give us the  $\Theta_i$ , and therefore the orthogonalized impulse–response functions.

In our previous example, we might assume that the contemporaneous effect of vote intention on PM satisfaction is zero and estimate the contemporaneous effect of PM satisfaction on vote intention. The estimated causal effect is 0.296 (standard error: 0.060). The result is statistically significant and suggests that an increase of one percentage point in PM satisfaction causes an immediate increase of 0.3 percentage points in vote intention for the incumbent government.

With the estimates for  $A$  and  $B$ , we can plot the structural impulse–response function (Figure 34.2). Comparing this to Figure 34.1, we see we now have an estimate for the contemporaneous effect of PM satisfaction on



Graphs by irfname, impulse variable and response variable

**Figure 34.2** Structural impulse response function

vote intention. Unlike the estimates from the reduced form model, this is an estimate of the causal effect. The units on the y axis have also changed. The constraints we placed on  $A$  and  $B$  had the effect of standardizing the variables. The impulse is now a change of one standard deviation in PM satisfaction and the response is measured in standard deviations of vote intention.

An alternative set of identifying restrictions are called long-run restrictions. Recall that the short-run SVAR can be written as follows:

$$A(y_t - b_0 - \beta_1 y_{t-1} - \beta_2 y_{t-2} - \dots - \beta_p y_{t-p}) = B e_t$$

This can also be written using the lag operator  $L^p$  (for example,  $L^p y_t = y_{t-p}$ ):

$$A(I - b_0 L - \beta_1 L^1 - \beta_2 L^2 - \dots - \beta_p L^p) y_t = B e_t$$

This allows us to rewrite the VAR:

$$y_t = C e_t$$

$$C \equiv \frac{A^{-1} B}{(I - b_0 L - \beta_1 L^1 - \beta_2 L^2 - \dots - \beta_p L^p)}$$

The SVAR can now be identified by placing restrictions on the matrix  $C$ . These restrictions can be interpreted as assumptions about the long-run causal relations between endogenous variables. Setting  $C_{ij} = 0$  can be interpreted as assuming the long-running response of variable  $i$  to the shocks driving variable  $j$  is zero.<sup>9</sup> The estimated parameters of  $C$  give us the long-run causal relationships between the endogenous variables. In this form, nothing is assumed about the short-run effects.<sup>10</sup>

The long-run SVAR makes assumptions about the long-run effects of the variables and system of equations. An alternative set of assumptions regarding the long-run relationship between variables motivates the VEC model, to which we turn next.

## VEC MODELS

To motivate vector error-correction models (VECMs), we begin with a review of cointegration. Up until now, we have assumed the variables are  $I(0)$  stationary (see Linn and Webb, Chapter 32, this *Handbook*), at least once we control for deterministic trends and periodicity. Say  $Y_t$  and  $X_t$  are  $I(1)$  non-stationary processes and that there is a  $\beta$  such that the linear combination of  $Y_t$  and  $X_t$ ,  $z_t = y_t - \alpha - \beta x_t$ , is a stationary process. If so, we say that  $Y_t$  and  $X_t$  are cointegrated and call  $\beta$  the cointegration parameter.<sup>11</sup> A common notation to indicate cointegration is  $(Y_t, X_t)' \sim CI(d, b)$ , where  $d$  indicates the order of integration for  $Y_t$  and  $X_t$  and  $d - b$  indicates the order of integration for  $z_t = y_t - \alpha - \beta x_t$ . For example, if  $Y_t$  and  $X_t$  are both  $I(1)$  and are cointegrated, such that  $Z_t$  is  $I(0)$ , then we could denote this as  $(Y_t, X_t)' \sim CI(1, 1)$ . If the two  $I(1)$  processes are cointegrated,  $\beta$  tells us the long-run relationship between  $Y_t$  and  $X_t$  - we will return to this later. If the two  $I(1)$  processes are not cointegrated, then there is no long-run relationship between  $Y_t$  and  $X_t$ , and  $y_t = \hat{\alpha} + \beta x_t + \hat{\varepsilon}_t$  is a spurious regression (although, note we could still estimate the short-run relationship with  $\Delta y_t = \beta \Delta x_t + \Delta \varepsilon_t$ ).

Certain combinations of variables cannot cointegrate. If  $z_t = y_t - \alpha - \beta x_t$  is a cointegrating relationship,  $Y_t$  and  $X_t$  must be of the same order of integration and  $Z_t$  must be of a lower order of integration. It does not make sense to claim that  $Y_t \sim I(0)$ ,  $X_t \sim I(1)$  and  $Z_t \sim I(0)$ ; an  $I(0)$  variable and an  $I(1)$  variable cannot combine to produce a  $I(0)$  variable. Nor does it make sense to claim that  $Y_t \sim I(2)$ ,  $X_t \sim I(1)$  and  $Z_t \sim I(0)$ . However, it is also possible to have the following situation:

$$z_t^* = y_t - \alpha - \beta_1 w_t - \beta_2 x_t$$

$$Y_t \sim I(2), X_t \sim I(1), W_t \sim I(2)$$

If  $Y_t$  and  $W_t$  are cointegrated, such that  $z_t = y_t - \alpha - \beta_1 z_t$  is  $I(1)$  (i.e.,  $(Y_t, W_t)' \sim CI(2,1)$ ) and  $Z_t$  and  $X_t$  are cointegrated, such that  $z_t^* = z_t - \beta_2 x_t$  is  $I(0)$  (i.e.,  $(Z_t, X_t)' \sim CI(1,1)$ ), then:

$$z_t^* = z_t - \beta_2 x_t = y_t - \alpha - \beta_1 w_t - \beta_2 x_t$$

is  $I(0)$ . This is called multi-cointegration. The requirement that the processes on the right-hand side must combine in a logical way to produce the process on the left-hand side is called balance. Balance is a somewhat advanced topic that we will not cover here (Banerjee et al., 2003).

We are now in a position to discuss the VECM:  $I(1)$  data can be modelled by first differencing it. However, this only allows us to look at short-run relationships. Cointegration allows us to expand the type of dynamic models that can be estimated for  $I(1)$  data beyond using first differences. If a cointegrating process  $Z_t$  is stationary, it can be included in time series models. Such models include error-correction models (also known as equilibrium-correction models (Hendry 2003)).

A standard VECM is just a VAR model that has been transformed. Starting with the SVAR with  $K = 2$  endogenous variables and a lag-order of  $P = 1$ :

$$\begin{aligned} y_{1,t} &= \alpha_1 + \alpha_{1,0}y_{2,t} + \gamma_{1,1}y_{1,t-1} \\ &\quad + \gamma_{1,2}y_{2,t-1} + \mu_{1,t} \\ y_{2,t} &= \alpha_2 + \alpha_{2,0}y_{1,t} + \gamma_{2,1}y_{1,t-1} \\ &\quad + \gamma_{2,2}y_{2,t-1} + \mu_{2,t} \end{aligned} \tag{6}$$

Subtract  $y_{1,t-1}$  from each side of the first equation and  $y_{2,t-1}$  from each side of the second equation:

$$\begin{aligned} y_{1,t} - y_{1,t-1} &= \alpha_1 + \alpha_{1,0}y_{2,t} \\ &\quad + \gamma_{1,1}y_{1,t-1} - y_{1,t-1} \\ &\quad + \gamma_{1,2}y_{2,t-1} + \mu_{1,t} \end{aligned}$$

$$\begin{aligned} y_{2,t} - y_{2,t-1} &= \alpha_2 + \alpha_{2,0}y_{1,t} \\ &\quad + \gamma_{2,1}y_{1,t-1} + \gamma_{2,2}y_{2,t-1} \\ &\quad - y_{2,t-1} + \mu_{2,t}, \end{aligned}$$

which can be rewritten as:

$$\begin{aligned} \Delta y_{1,t} &= \alpha_1 + (\gamma_{1,1} - 1)y_{1,t-1} \\ &\quad + \alpha_{1,0}y_{2,t} + \gamma_{1,2}y_{2,t-1} + \mu_{1,t} \\ \Delta y_{2,t} &= \alpha_2 + (\gamma_{2,2} - 1)y_{2,t-1} \\ &\quad + \alpha_{2,0}y_{1,t} + \gamma_{2,1}y_{1,t-1} + \mu_{2,t} \end{aligned}$$

Add and subtract  $\alpha_{1,0}y_{2,t-1}$  on the right-hand side of the first equation and add and subtract  $\alpha_{2,0}y_{1,t-1}$  on the right-hand side of the second equation:

$$\begin{aligned} \Delta y_{1,t} &= \alpha_1 + (\gamma_{1,1} - 1)y_{1,t-1} + \alpha_{1,0}y_{2,t} - \alpha_{1,0}y_{2,t-1} \\ &\quad + \alpha_{1,0}y_{2,t-1} + \gamma_{1,2}y_{2,t-1} + \mu_{1,t} \\ \Delta y_{2,t} &= \alpha_2 + (\gamma_{2,2} - 1)y_{2,t-1} + \alpha_{2,0}y_{1,t} - \alpha_{2,0}y_{1,t-1} \\ &\quad + \alpha_{2,0}y_{1,t-1} + \gamma_{2,1}y_{1,t-1} + \mu_{2,t} \end{aligned}$$

Note

$$\begin{aligned} \alpha_{k,0}y_{k,t} - \alpha_{k,0}y_{k,t-1} &= \alpha_{k,0}\Delta y_{k,t} \\ &\quad \text{and } \alpha_{k,0}y_{k,t-1} + \gamma_{k,1}y_{k,t-1} \\ &= (\alpha_{k,0} + \gamma_{k,1})y_{k,t-1} : \end{aligned}$$

$$\begin{aligned} \Delta y_{1,t} &= \alpha_1 + (\gamma_{1,1} - 1)y_{1,t-1} + \alpha_{1,0}\Delta y_{2,t} \\ &\quad + (\alpha_{1,0} + \gamma_{1,2})y_{2,t-1} + \mu_{1,t} \\ \Delta y_{2,t} &= \alpha_2 + (\gamma_{2,2} - 1)y_{2,t-1} + \alpha_{2,0}\Delta y_{1,t} \\ &\quad + (\alpha_{2,0} + \gamma_{2,1})y_{1,t-1} + \mu_{2,t} \end{aligned}$$

We then use  $(\alpha_{k,0} + \gamma_{k,1})$

$$\begin{aligned} &= \frac{(\gamma_{k,k} - 1)(\alpha_{k,0} + \gamma_{k,1})}{(\gamma_{k,k} - 1)} \\ &= -\frac{(\gamma_{k,k} - 1)(\alpha_{k,0} + \gamma_{k,1})}{(1 - \gamma_{k,k})} : \end{aligned}$$

$$\begin{aligned} \Delta y_{1,t} &= \alpha_1 + (\gamma_{1,1} - 1)y_{1,t-1} + \alpha_{1,0}\Delta y_{2,t} \\ &\quad - \frac{(\gamma_{1,1} - 1)(\alpha_{1,0} + \gamma_{1,2})}{(\gamma_{1,1} - 1)}y_{2,t-1} + \mu_{1,t} \\ \Delta y_{2,t} &= \alpha_2 + (\gamma_{2,2} - 1)y_{2,t-1} + \alpha_{2,0}\Delta y_{1,t} \\ &\quad - \frac{(\gamma_{2,2} - 1)(\alpha_{2,0} + \gamma_{2,1})}{(\gamma_{2,2} - 1)}y_{1,t-1} + \mu_{2,t} \end{aligned}$$

$$\begin{aligned} \Delta y_{1,t} &= \alpha_1 + \gamma_1 [y_{1,t-1} - \kappa_1 y_{2,t-1}] \\ &\quad + \kappa_{1,2,0}\Delta y_{2,t} + \mu_{1,t} \\ \Delta y_{2,t} &= \alpha_2 + \gamma_2 [y_{1,t-1} - \kappa_1 y_{2,t-1}] \\ &\quad + \kappa_{2,1,0}\Delta y_{1,t} + \mu_{2,t} \end{aligned} \tag{8}$$

Collect terms multiplied by  $(\gamma_{kk} - 1)$ :

$$\begin{aligned} \Delta y_{1,t} &= \alpha_1 + (\gamma_{1,1} - 1) \left[ y_{1,t-1} - \frac{(\alpha_{1,0} + \gamma_{1,2})}{(\gamma_{1,1} - 1)} y_{2,t-1} \right] \\ &\quad + \alpha_{1,0}\Delta y_{2,t} + \mu_{1,t} \\ \Delta y_{2,t} &= \alpha_2 + (\gamma_{2,2} - 1) \left[ y_{2,t-1} - \frac{(\alpha_{2,0} + \gamma_{2,1})}{(\gamma_{2,2} - 1)} y_{1,t-1} \right] \\ &\quad + \alpha_{2,0}\Delta y_{1,t} + \mu_{2,t} \end{aligned}$$

This can be written as:

$$\begin{aligned} \Delta y_{1,t} &= \alpha_1 + \gamma_1 [y_{1,t-1} - \kappa_{1,1}y_{2,t-1}] \\ &\quad + \kappa_{1,2,0}\Delta y_{2,t} + \mu_{1,t} \\ \Delta y_{2,t} &= \alpha_2 + \gamma_2 [y_{2,t-1} - \kappa_{2,1}y_{1,t-1}] \\ &\quad + \kappa_{2,1,0}\Delta y_{1,t} + \mu_{2,t} \end{aligned} \tag{7}$$

where  $\gamma_k = (\gamma_{k,k} - 1), \kappa_{k,1}$

$$= \frac{(\alpha_{k,0} + \gamma_{k,1})}{(1 - \gamma_{k,k})} \text{ and } \kappa_{k,2,0} = \alpha_{k,0}$$

The terms in the square brackets represent the cointegrating relationships between the endogenous variables (note that one could assume the  $\alpha_k$  to be equal and bring it into the cointegrating relationship). It can be shown that for any two  $I(1)$  variables, there can be at most one cointegrating relationship. Therefore, if  $Y_t$  and  $X_t$  are  $I(1)$  and cointegrated,  $[y_{1,t-1} - \kappa_{1,1}y_{2,t-1}] = [y_{2,t-1} - \kappa_{2,1}y_{1,t-1}]$  and (7) can be rewritten as:

If  $Y_{1,t}$  and  $Y_{2,t}$  are  $I(1)$  and cointegrated, the cointegrating relationship  $[y_{1,t-1} - \kappa_1 y_{2,t-1}]$  is  $I(0)$  stationary and the remaining terms are  $I(0)$  stationary. Therefore, all terms in the ECM are  $I(0)$  stationary.

It is not necessary that the variables in a VECM are  $I(1)$  and cointegrated: they can also be all  $I(0)$ . If this is the case, then the  $\kappa_{k,1}$  in (7) are the long-run effects of a change in one endogenous variable on the other, and the  $\kappa_{k,1,0}$  are the short-run effects. If the data are  $I(1)$  and cointegrated, the  $\kappa_{k,1,0}$  continue to be the short-run effect but  $\kappa_1$  in (8) is the cointegrating parameter. This still represents a long-run relationship between the variables but it is of a different nature than the long-run relationship between stationary variables.

If  $y_{1,t-1} - \kappa_1 y_{2,t-1}$  is the cointegrating relationship, the estimates of  $\gamma_1$  and  $\gamma_2$  indicate how quickly  $\Delta Y_t$  and  $\Delta X_t$  respectively respond to  $Y_t$  and  $X_t$  being out of equilibrium. These are the rate of adjustment coefficients. If  $Y_t$  and  $X_t$  are out of equilibrium with each other (i.e.,  $y_{1,t-1} - \kappa_1 y_{2,t-1}$  is not equal to zero), then  $Y_t$  will change such that  $|\gamma_1| \times 100\%$  of the adjustment necessary to bring them into equilibrium occurs in the next time point. Then,  $|\gamma_1| \times 100\%$  of the remaining adjustment occurs in the next time point after that and so on. Similarly,  $X_t$  will change such that  $|\gamma_2| \times 100\%$  of the adjustment necessary to bring them into equilibrium occurs in the next time point and so on.<sup>12</sup>

Equation (8) is the VECM of order  $P - 1 = 0$ . A higher order VECM includes additional lags of the first differences of the variables. For example, a first order VECM would be:

$$\begin{aligned} \Delta y_{1,t} &= \alpha_1 + \gamma_1 [y_{1,t-1} - \kappa_1 y_{2,t-1}] \\ &\quad + \kappa_{1,2,0} \Delta y_{2,t} + \kappa_{1,1,1} \Delta y_{1,t-1} \\ &\quad + \kappa_{1,2,1} \Delta y_{2,t-1} + \mu_{1,t} \\ \Delta y_{2,t} &= \alpha_2 + \gamma_2 [y_{1,t-1} - \kappa_1 y_{2,t-1}] \\ &\quad + \kappa_{2,1,0} \Delta y_{1,t} + \kappa_{2,1,1} \Delta y_{1,t-1} \\ &\quad + \kappa_{2,2,1} \Delta y_{2,t-1} + \mu_{2,t} \end{aligned}$$

This is a structural VECM(1) model. The issue of deciding how many lags of the differenced endogenous variables should be included is one of choosing a lag length that ensures the resulting residuals contain no serial correlation. Generalizing to a structural VECM( $P - 1$ ), where  $P$  denotes the order of the corresponding VAR:

$$\begin{aligned} \Delta y_{1,t} &= \gamma_1 [z_{t-1}] + \sum_{i=0}^{P-1} \kappa_{1,1,i} \Delta y_{t-i} \\ &\quad + \sum_{i=0}^{P-1} \kappa_{1,2,i} \Delta x_{t-i} + \varepsilon_{1,t} \\ \Delta y_{2,t} &= \gamma_2 [z_{t-1}] + \sum_{i=0}^{P-1} \kappa_{2,1,i} \Delta y_{t-i} \\ &\quad + \sum_{i=1}^{P-1} \kappa_{2,2,i} \Delta x_{t-i} + \varepsilon_{2,t} \end{aligned} \tag{9}$$

$$z_{t-1} \equiv y_{t-1} - \lambda_1 - \kappa_{1,1} x_{t-1}$$

A VECM is isomorphic with a VAR model. Specifically, the VECM(1) with a single lag of the first difference of the dependent variables is isomorphic to a VAR(2), and just as with any VAR, this can be transformed into the reduced form:

$$\begin{aligned} \Delta y_{1,t} &= \Gamma_{1,0,1} [z_{t-1}] + \Gamma_{1,1,1} \Delta y_{t-1} \\ &\quad + \Gamma_{1,2,1} \Delta x_{t-1} + \varepsilon_{1,t} \\ \Delta y_{2,t} &= \Gamma_{2,0,1} [z_{t-1}] + \Gamma_{2,1,1} \Delta y_{t-1} \\ &\quad + \Gamma_{2,2,1} \Delta x_{t-1} + \varepsilon_{2,t} \end{aligned} \tag{10}$$

$$z_{t-1} \equiv y_{t-1} - \lambda_1 - \kappa_{1,1} x_{t-1} \rightarrow$$

Generalizing to the reduced form VECM( $P - 1$ ):

$$\begin{aligned} \Delta y_{1,t} &= \Gamma_{1,0,1} [z_{t-1}] + \sum_{i=1}^{P-1} \Gamma_{1,1,i} \Delta y_{t-i} \\ &\quad + \sum_{i=1}^{P-1} \Gamma_{1,2,i} \Delta x_{t-i} + \varepsilon_{1,t} \\ \Delta y_{2,t} &= \Gamma_{2,0,1} [z_{t-1}] + \sum_{i=1}^{P-1} \Gamma_{2,1,i} \Delta y_{t-i} \\ &\quad + \sum_{i=1}^{P-1} \Gamma_{2,2,i} \Delta x_{t-i} + \varepsilon_{2,t} \\ z_{t-1} &\equiv y_{t-1} - \lambda_1 - \kappa_{1,1} x_{t-1} \end{aligned}$$

Note that the reduced form does not include contemporaneous values of the endogenous variables on the left-hand side.

The constants in the VECM equations can be included within or outside the cointegrating relationship. Further, a trend can be included inside or outside the cointegrating equation.

$$\begin{aligned} \Delta y_{1,t} &= \Gamma_{1,0,1} [z_{t-1}] + \sum_{i=1}^{P-1} \Gamma_{1,1,i} \Delta y_{t-i} \\ &\quad + \sum_{i=1}^{P-1} \Gamma_{1,2,i} \Delta x_{t-i} + \gamma + \tau t + \varepsilon_{1,t} \\ \Delta y_{2,t} &= \Gamma_{2,0,1} [z_{t-1}] + \sum_{i=1}^{P-1} \Gamma_{2,1,i} \Delta y_{t-i} \\ &\quad + \sum_{i=1}^{P-1} \Gamma_{2,2,i} \Delta x_{t-i} + \gamma + \tau t + \varepsilon_{2,t} \\ z_{t-1} &\equiv y_{t-1} - \lambda_1 - \rho t - \kappa_{1,1} x_{t-1} \end{aligned}$$

The constant inside the cointegrating equation allows the cointegrating variables to be stationary around the constant. The constant outside the cointegrating equation adds a linear time trend in the levels of the data. The trend within the cointegrating equation allows the cointegrating variables to be stationary around a trend. The trend outside the cointegrating equation allows for quadratic trends in the levels of the variables.

The VECM with two endogenous variables examined so far can be extended to include additional endogenous variables. If

these are  $I(1)$ , they can be included in the cointegrating relationship. For example,  $[y_{1t-1} - \kappa_1 y_{2t-1} - \kappa_2 y_{3t-1}]$ . It is important to determine if the endogenous variables are cointegrated, and if we have more than two endogenous variables, we *may* have more than one cointegrating relationship. Both cointegration and the number of cointegrating equations can be tested with the Johansen rank test (Johansen, 1995).

For our next example, we will use data from an article by Moore and Lanoue (2003). They use a vector error-correction model to look at the relationship between hostility sent ( $HS_t$ ) from the United States to other countries, hostility received ( $HR_t$ ) by the United States from other countries and the popularity of the US president ( $PP_t$ ). Hostility sent and received is measured on the basis of coded news reports of foreign policy behaviour – news reports are given a score ranging from cooperative foreign policy behaviour to conflictual foreign policy behaviour. The data used by Moore and Lanoue (2003) are quarterly means of these scores, from 1953–78. The popularity of the US president is based on the proportion of poll respondents indicating they approved of the performance of the president, aggregated quarterly.

We have three potentially endogenous variables. Using the Johansen rank test, Moore and Lanoue (2003) determined that the data are cointegrated with one cointegrating equation and the appropriate model is a VECM(1):

$$\begin{aligned} \Delta HS_t &= \hat{\gamma}_{11}^* [\hat{z}_{1,t-1}] + \hat{\Gamma}_{1,1,1} \Delta HS_{t-1} \\ &\quad + \hat{\Gamma}_{1,2,1} \Delta HR_{t-1} + \hat{\Gamma}_{1,3,1} \Delta PP_{t-1} + \hat{\epsilon}_{1,t} \\ \Delta HR_t &= \hat{\gamma}_{21}^* [\hat{z}_{1,t-1}] + \hat{\Gamma}_{2,1,1} \Delta HS_{t-1} \\ &\quad + \hat{\Gamma}_{2,2,1} \Delta HR_{t-1} + \hat{\Gamma}_{2,3,1} \Delta PP_{t-1} + \hat{\epsilon}_{2,t} \\ \Delta PP_t &= \hat{\gamma}_{31}^* [\hat{z}_{1,t-1}] + \hat{\Gamma}_{3,1,1} \Delta HS_{t-1} \\ &\quad + \hat{\Gamma}_{3,2,1} \Delta HR_{t-1} + \hat{\Gamma}_{3,3,1} \Delta PP_{t-1} + \hat{\epsilon}_{3,t} \\ \hat{z}_{1,t-1} &\equiv \hat{\kappa}_{1,1} HS_{t-1} + \hat{\kappa}_{1,2} HR_{t-1} + \hat{\kappa}_{1,3} PP_{t-1} + \hat{\lambda}_1 \end{aligned}$$

The estimated parameters from this model are presented in Table 34.2. The coefficients on the lag of the first difference of the variables are the short-run responses of the dependent variable to these variables. The only statistically significant short-run response (other than a variable’s response to its own change in a previous quarter) is that of presidential approval to hostility sent to other countries (positive effect). Note that as with the reduced form VAR, these ‘responses’ are not causal in the traditional sense; they are predictive.

The  $CI_{t-1}$  indicates the cointegrating relationship, which is also calculated and reported. The coefficients on  $CI_{t-1}$  are the (rate of) adjustment coefficients. Each endogenous variable does adjust to being out of equilibrium (the adjustment coefficients are statistically significant) but not particularly quickly. Approval responds the quickest, moving to close 13% of the gap each period.

Looking to the cointegrating relationship, we have restricted the coefficient on the first variable (hostility sent) to be one – this is a necessary but innocuous identifying restriction. The effect of presidential popularity in the cointegrating relationship is not statistically significant at the 0.05 significance level. The coefficient on hostility received is statistically significant in the cointegrating relationship.

To understand the cointegrating relationship, consider the impact of a change in hostility received, such that it moves out of equilibrium with hostility sent. The expected value for the cointegrating equation is zero. The coefficient on hostility received is negative and the coefficient on hostility sent is positive (1).

$$CI_{t-1} = HS_t - 5.33HR_t + 0.36PP_t + 44.93$$

This means that when hostility received increases so that it is out of equilibrium with hostility sent, the cointegrating equations will

**Table 34.2 Hostility sent, hostility received and presidential popularity, United States 1953–78**

Equation		Parameter estimates
$\Delta HS_t$	$CI_{t-1}$	-0.100 (0.048)*
	$\Delta HS_{t-1}$	-0.223 (0.098)*
	$\Delta HR_{t-1}$	-0.137 (0.280)
	$\Delta PP_{t-1}$	0.063 (0.107)
	Constant	0.038 (0.571)
$\Delta HR_t$	$CI_{t-1}$	0.053 (0.018)**
	$\Delta HS_{t-1}$	0.014 (0.101)*
	$\Delta HR_{t-1}$	-0.250 (0.101)*
	$\Delta PP_{t-1}$	-0.039 (0.039)
	Constant	-0.074 (0.207)
$\Delta PP_t$	$CI_{t-1}$	-0.129 (0.044)**
	$\Delta HS_{t-1}$	0.188 (0.089)*
	$\Delta HR_{t-1}$	-0.251 (0.254)
	$\Delta PP_{t-1}$	0.142 (0.097)
	Constant	-0.060 (0.519)
	<i>T</i>	102

\*  $p < 0.05$ ; \*\*  $p < 0.01$ ; standard errors in parentheses.

### Cointegrating equation

	Parameter estimates
$HS_t$	1 (Fixed)
$HR_t$	-5.32854 (1.264653)**
$PP_t$	0.356464 (0.211377)
Constant	44.93087

produce negative values (below equilibrium). The adjustment coefficient for the cointegrating equation in the hostility-sent equation is negative, which means that when the cointegrating equation produces a negative value, hostility sent responds by increasing (starting in the next period). Putting all this together, when hostility received increases so that it is too high relative to hostility sent (moving the two out of equilibrium), hostility sent increases to close the gap.

The adjustment coefficient in the hostility-received equation is positive and so if hostility received increases so that it is too high relative to hostility sent (the cointegrating

relationship produces negative values), hostility received decreases (starting in the next period). The adjustment coefficient in the presidential-approval equation is negative and so presidential approval increases when the cointegrating relationship produces negative values – if hostility received is too high relative to hostility sent.

### THE STATE-SPACE APPROACH

We now turn to the state-space approach to dynamic systems of equations. Those that are

familiar with the ARMA or ARIMA (Autoregressive Integrated Moving Average) model will have (perhaps unknowingly) already seen a state-space model. The ARMA model is sometimes written as follows:

$$y_t = \alpha_0 + \sum_{i=1}^p \alpha_i y_{t-i} + \sum_{j=1}^q \theta_j \xi_{t-j} + \xi_t$$

$$\xi_t \sim NID(0, \sigma_\xi^2)$$

But it is also commonly written in two components:

$$y_t = \beta_0 + a_t$$

$$a_t = \sum_{i=1}^p \alpha_i a_{t-i} + \sum_{j=1}^q \theta_j \xi_{t-j} + \xi_t$$

The first equation is called the structural (or observation or measurement) component and the second is the disturbance (or state) component. This is a state-space model and it is the representation often used for the purposes of ARMA/ARIMA model estimation and interpretation of the estimation results.

The model can also be written in matrix form. Take the ARMA(1,1), for example, which can be written as

$$y_t = (1 \quad 1 \quad 0) \begin{pmatrix} \beta_0 \\ a_{1,t} \\ a_{2,t} \end{pmatrix}$$

Observation equation

$$\begin{pmatrix} \beta_0 \\ a_{1,t} \\ a_{2,t} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \alpha_1 & 1 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \beta_0 \\ a_{1,t-1} \\ a_{2,t-1} \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \\ \theta_1 \end{pmatrix} \xi_t$$

State equation

or

$$y_t = \beta_0 + \mathbf{Z}a_t \quad \text{Observation equation}$$

$$a_t = \mathbf{T}a_{t-1} + R\xi_t \quad \text{State equation}$$

$$a_t = \begin{pmatrix} \beta_0 \\ a_{1,t} \\ a_{2,t} \end{pmatrix}, \mathbf{Z} = (1 \quad 1 \quad 0),$$

where

$$\mathbf{T} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \alpha_1 & 1 \\ 0 & 0 & 0 \end{pmatrix} \text{ and } R = \begin{pmatrix} 0 \\ 1 \\ \theta_1 \end{pmatrix}.$$

To understand the etymology of the name ‘state-space’, consider the current ARMA example. The state equation contains two state variables:  $a_{1t}$  and  $a_{2t}$  ( $\beta_0$  can also be thought of as a state that does not vary). These describe the *state* of the system we are observing. If we are observing public opinion, then this is the state of public opinion.

The *space* is a mathematical abstraction; it is the mathematical space in which different ‘positions’ represent different states of the system. For example, in Figure 34.3, the perpendicular axes define a space within which each point represents the values of  $a_{1t}$  and  $a_{2t}$ . Each pair of values, and therefore each point in the space, defines a state. We measure and therefore observe ( $y_t$ ), and it is a function of this state.

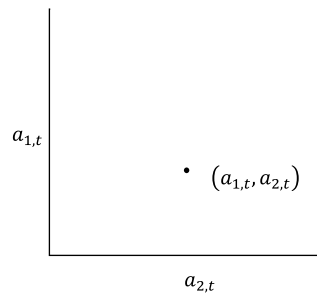


Figure 34.3 Two-dimensional space



The state-space approach to modelling time series is a more flexible way to solve problems of non-stationarity (e.g.,  $I(1)$  processes and deterministic trends) and measurement error than classical approaches (Durbin and Koopman, 2001). It is a powerful method for filtering and smoothing time series and forecasting (Commandeur and Koopman, 2007).<sup>13</sup> Most importantly for our discussion, it is easily extended to the multivariate case and it subsumes almost all classical linear time-series models – i.e., most classical models can be expressed in state-space form.

We begin by generalizing beyond the ARMA model. The state-space approach relates observations on one or more variables to unobserved (read latent) *states* by an *observation* equation:

$$y_t = \mathbf{Z}_t a_t + \mathbf{G}_t \varepsilon_t, \quad \varepsilon_t \sim NID(0, \boldsymbol{\sigma}_\varepsilon^2)$$

Observation equation

where  $y_t$  is an  $P \times 1$  vector of observed endogenous variables, in which  $P$  denotes the number of observed variables;  $\mathbf{Z}_t$  is an  $P \times M$  design matrix;  $a_t$  is an unobserved  $M \times 1$  state vector, in which  $M$  denotes the number of states; and  $\varepsilon_t$  is a  $P \times 1$  vector of observation errors with zero means and variances defined by the  $P \times P$  diagonal matrix  $\boldsymbol{\sigma}_\varepsilon^2$  – meaning the errors from the different observed variables do not correlate. Diagonal elements of  $\mathbf{G}_t$  are used to determine which vectors of  $y_t$  have stochastic error terms (set to one) and which are deterministic (set to zero). The state equation is

$$a_t = \mathbf{T}_t a_{t-1} + \mathbf{R}_t \xi_t, \quad \xi_t \sim NID(0, \boldsymbol{\sigma}_\xi^2)$$

State equation

where  $\mathbf{T}_t$  is an  $M \times M$  transition matrix that determines how values of the state vectors  $a_t$  relate to past values  $a_{t-1}$ ;  $\xi_t$  is a  $M \times 1$  vector of state disturbances, with zero means and

variances defined by the  $M \times M$  matrix  $\boldsymbol{\sigma}_\xi^2$ . The observation and state disturbances are assumed to be serially and mutually independent. Diagonal elements of  $\mathbf{R}_t$  are used to determine which states have stochastic error terms (set to one) and which are deterministic (set to zero). It is straightforward to include exogenous covariates in the observation and/or state equations.

The exogenous coefficients in the model can themselves be modelled by a state equation (allowing them to vary over time). For example, see Matthews and Pickup (2019). It is more common to specify a model with time invariant coefficients:

$$y_t = \mathbf{Z} a_t + \beta x_t + G \varepsilon_t, \quad \varepsilon_t \sim NID(0, \boldsymbol{\sigma}_\varepsilon^2)$$

$$a_t = \mathbf{T} a_{t-1} + \lambda \omega_t + R \xi_t, \quad \xi_t \sim NID(0, \boldsymbol{\sigma}_\xi^2)$$

The  $\beta x_t$  and  $\lambda \omega_t$  are exogenous variables within the observation and state equations, respectively.

The state-space form is used to derive the likelihood of the observed endogenous variables, conditional on their own past values and any exogenous variables. Expressing the model in state-space form allows us to use something called a Kalman filter to derive the prediction error form of the likelihood function. The Kalman filter is a set of *recursive* algorithms applied to the observed data to predict current values of the unobserved states and endogenous variables from past values (and exogenous variables). Recursive means the algorithm is first applied to the data at  $t = 1$  (e.g.,  $y_1$ ), then the algorithm is applied to the output and the data at  $t = 2$ , and the algorithm is then applied to the output and the data at  $t = 3$ , and so on.

The purpose of the filter is to obtain *optimal* predictions of the state vector(s) at  $t$  considering the observations  $\{y_1, y_2, \dots$

$y_{t-1}$  } – denoted  $a_{t|t-1}$ . Optimal is defined as the smallest prediction error:  $y_t - Za_{t|t-1}$ ; or  $y_t - Za_{t|t-1} - \beta x_t$ , if exogenous covariates are included.

Assuming a single state equation, so that the transition matrix  $T$  is a scalar, and a single observation equation, and excluding covariates, the Kalman filter algorithm is:

$$a_{t|t-1} = T \left[ a_{t-1|t-2} + K_{t-1} (y_{t-1} - Za_{t-1|t-2}) \right]$$

Predicted state at  $t$  is equal to the predicted state at  $t - 1$ , adjusted by the prediction error of the predicted state at  $t - 1$ , all multiplied by the transition scalar  $T$

The key to the Kalman filter is the ‘Kalman gain’:  $K_{t-1}$ . It determines how much the past prediction error influences the current predicted state. For simplicity, we shall assume the design matrix  $Z = 1$ :

$$a_{t|t-1} = T \left[ a_{t-1|t-2} + K_{t-1} (y_{t-1} - a_{t-1|t-2}) \right]$$

The Kalman gain is calculated as a compromise between the uncertainty in the state estimate based on  $\{y_1, y_2, \dots, y_{t-2}\}$  (i.e.,  $a_{t-1|t-2}$ ) and the uncertainty in the observation  $y_{t-1}$ . If the uncertainty in the observation  $y_{t-1}$  is very large relative to that in the predicted state  $a_{t-1|t-2}$ , we do not want  $y_{t-1}$  to influence subsequent predicted states very much. This is achieved by having  $K_{t-1} \rightarrow 0$ , such that the predicted state becomes

$$a_{t|t-1} = T \left[ a_{t-1|t-2} \right]$$

If the uncertainty in the predicted state  $a_{t-1|t-2}$  is very large relative to that in the observation  $y_{t-1}$ , we want  $y_{t-1}$  to have a large influence, such that the subsequent predicted state  $a_{t|t-1}$  is entirely informed by  $y_{t-1}$ . This is achieved by having  $K_{t-1} \rightarrow 1$ , such that the predicted state becomes:

$$a_{t|t-1} = T \left[ y_{t-1} \right]$$

To achieve these goals, the Kalman gain is defined as:

$$K_{t-1} = \frac{P_{t-1}}{F_{t-1}}$$

If we define the prediction error,  $v_t = y_t - Za_{t|t-1}$ , then  $F_{t-1}$  is the variance of the prior prediction error,  $var(v_{t-1})$ , and  $P_{t-1}$  is the error variance of the prior predicted state,  $a_{t-1|t-2}$ . Therefore,  $K_{t-1}$  is the proportion of the prior prediction error that can be attributed to uncertainty in the prior predicted state. The remaining error is due to the error in our measurement of  $y_{t-1}$ . If the entire prediction error is a result of uncertainty in the predicted state (at  $t - 1$ ), we do not want the predicted state to influence our subsequent predictions,  $K_{t-1} = 1$ . If no part of the prediction error is a result of uncertainty in the predicted state (at  $t - 1$ ), it must come from uncertainty in the observed value ( $y_{t-1}$ ). We do not want the observed value to influence our subsequent prediction,  $K_{t-1} = 0$ .

The means by which  $F_{t-1}$  and  $P_{t-1}$  are calculated and the proof that they give the optimal value for  $K_{t-1}$  is beyond this chapter, but they are functions of the state-space model parameters (Harvey, 1993).<sup>14</sup> These can be estimated by expressing the log-likelihood function for the observed data in terms of the prediction errors ( $v_t$ ) and

their variances ( $F_t$ ) and maximizing this function:

$$\log(L_d) = -\frac{T}{2} \log(2\pi) - \frac{1}{2} \sum_{t=d+1}^T \left( \log(F_t) + \frac{v_t}{F_t} \right)$$

where  $v_t$  and  $F_t$  are functions of the data and the unknown state-space model parameters and  $d$  is the number of *initial elements* for the state. An initial element is the value of the state prior to the first observation. These are required because the Kalman filter is recursive. The recursive nature means that the estimate of the current state always refers to the previous state. Because we need to start the Kalman algorithm somewhere, some assumptions need to be made about the first value of the state(s) to be entered into the algorithm – the initial element(s).<sup>15</sup>

The maximization of the log likelihood simultaneously minimizes the prediction errors,  $v_t$ , and their variances,  $F_t$ . Therefore, estimation of the state-space parameters through maximization of this log likelihood gives us the optimal predictions. This highlights a difference between the estimation of state-space models and classical time-series models. Classical models (using OLS or maximum likelihood estimation (MLE)) minimize the observation errors,  $\hat{\epsilon}_t$ , and their variances,  $\sigma_{\hat{\epsilon}_t}^2$ .

In addition to allowing us to write a likelihood function, the Kalman filter gives us both filtered and predicted values for the state(s):

$$\frac{a_{t|t-1}}{\text{Predicted state : } a_{t|t-1}} = \frac{T \left[ a_{t-1|t-2} + K_{t-1} (y_{t-1} - Z a_{t-1|t-2}) \right]}{\text{Filtered state } a_{t-1}}$$

The predicted state,  $a_{t|t-1}$ , is the optimal prediction of  $a_t$  based on  $\{y_1, y_2, \dots, y_{t-1}\}$ . The

filtered state,  $a_t$ , is the optimal estimate of  $a_t$  based on  $\{y_1, y_2, \dots, y_t\}$ . Each filtered state is a compromise between a prediction of the current state based on past information and the current observation:

$$a_t = a_{t|t-1} + K_t (y_t - Z a_{t|t-1})$$

The compromise is based on how much (un)certainty exists in each.

We can gain an intuition as to why  $a_t$  is a filtered state if we consider the following simple state-space model:

$$y_t = a_t + \epsilon_t, \quad \epsilon_t \sim NID(0, \sigma_\epsilon^2)$$

$$a_t = T a_{t-1} + \xi_t, \quad \xi_t \sim NID(0, \sigma_\xi^2)$$

We can interpret the  $\xi_t$  component as a random disturbance in the state process that we are interested in ( $a_t$ ) and  $\epsilon_t$  as random measurement error in our indirect observations of that state ( $y_t$ ). Our estimate of the filtered state,  $a_t$ , is our estimate of the observed value of  $y_t$ , with the random measurement error removed (or filtered).

In addition to predicted and filtered states, we can calculate *smoothed states*. Smoothed states are obtained by taking the output of the Kalman filter (the filtered states) and applying a similar algorithm backwards through time. This produces estimates of the states at a given time  $t$  based on all observations:  $\{y_1, y_2, \dots, y_T\}$ .

As suggested by the notation that we have been using for the state-space models, it is relatively easy to apply this approach to multivariate models. As examples, we will look at how to specify two classic models in state-space form: the reduced form VAR and the SVAR. We will then examine two models that are specific to the state-space approach: dynamic seemingly unrelated time-series equation and dynamic-factor models.

The reduced form VAR(1) model with two endogenous variables can be expressed in state-space form as follows:

**Table 34.3 PM satisfaction, vote intention and unemployment, Britain 1997–2006**

Equation		Parameter estimates
<i>PMs</i> $a_{PM,t}$	$a_{PM,t}$	1 (Fixed)
	$a_{PM,t-1}$	0.803 (0.086)**
	$a_{V,t-1}$	0.257 (0.184)
	Trend 1	-0.096 (0.051)
	Trend 2	0.023 (0.070)
	Trend 3	-0.460 (0.397)
	Cycle 2	-4.389 (1.814)*
	Cycle 3	-2.192 (3.177)
	Unemployment	-8.931 (3.152)**
	Constant	0.314 (7.118)
	VAR( $a_{PM,t}$ )	17.019 (2.404)**
	Vote $a_{V,t}$	$a_{V,t}$
$a_{V,t-1}$		0.219 (0.115)
$a_{PM,t-1}$		0.227 (0.051)**
Trend 1		-0.053 (0.033)
Trend 2		-0.148 (0.045)**
Trend 3		0.070 (0.256)
Cycle 2		0.489 (1.163)
Cycle 3		-5.041 (2.034)*
Unemployment		0.063 (2.021)
Constant		29.038 (4.533)**
VAR( $a_{V,t}$ )		7.088 (1.014)**
COV ( $a_{PM,t}$ , $a_{V,t}$ )		5.089 (1.219)**
$T$		101

\*  $p < 0.05$ ; \*\*  $p < 0.01$ ; standard errors in parentheses.

$$\begin{pmatrix} y_{1,t} \\ y_{2,t} \end{pmatrix} = \begin{pmatrix} a_{1,1,t} \\ a_{1,2,t} \end{pmatrix}$$

$$\begin{pmatrix} a_{1,1,t} \\ a_{1,2,t} \end{pmatrix}; = \begin{pmatrix} \alpha_{1,1} & \alpha_{1,2} \\ \alpha_{2,1} & \alpha_{2,2} \end{pmatrix} \begin{pmatrix} a_{2,1,t} \\ a_{2,2,t} \end{pmatrix}$$

$$+ \begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \xi_t,$$

$$\xi_t \sim NID \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{\xi_{1,1}}^2 & \sigma_{\xi_{1,2}}^2 \\ \sigma_{\xi_{2,1}}^2 & \sigma_{\xi_{2,2}}^2 \end{pmatrix} \right)$$

$$\begin{pmatrix} a_{2,1,t} \\ a_{2,2,t} \end{pmatrix} = \begin{pmatrix} a_{1,1,t-1} \\ a_{1,2,t-1} \end{pmatrix}$$

We return to our British vote intention example, in which  $Y_{1t}$  = PM satisfaction and  $Y_{2t}$  = vote intention. We also include unemployment as an exogenous variable in the state equation. Estimation of the state-space model produces the results presented in Table 34.3. The results (both parameter estimates and standard errors) are very similar to those from the reduced form VAR results presented in Table 34.2. Increases in the unemployment rate have a negative effect on PM satisfaction and that, in turn, predicts a reduction in vote intention for the incumbent government.

We can also represent a structural VAR(1) model with two endogenous variables in state-space form.

$$\begin{pmatrix} y_{1,t} \\ y_{2,t} \end{pmatrix} = \begin{pmatrix} \mu_{1,1,t} \\ \mu_{1,2,t} \end{pmatrix}$$

$$\begin{pmatrix} \mu_{1,1,t} \\ \mu_{1,2,t} \end{pmatrix}; = \begin{pmatrix} 0 & \alpha_{1,0} \\ \alpha_{2,0} & 0 \end{pmatrix} \begin{pmatrix} \mu_{1,1,t-1} \\ \mu_{1,2,t-1} \end{pmatrix};$$

$$+ \begin{pmatrix} \alpha_{1,1} & \alpha_{1,2} \\ \alpha_{2,1} & \alpha_{2,2} \end{pmatrix} \begin{pmatrix} \mu_{2,1,t} \\ \mu_{2,2,t} \end{pmatrix}$$

$$+ \begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \xi_t,$$

$$\xi_t \sim NID \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{\xi_{1,1}}^2 & 0 \\ 0 & \sigma_{\xi_{2,2}}^2 \end{pmatrix} \right)$$

$$\begin{pmatrix} \mu_{2,1,t} \\ \mu_{2,2,t} \end{pmatrix} = \begin{pmatrix} \mu_{1,1,t-1} \\ \mu_{1,2,t-1} \end{pmatrix}$$

In order for this to be identified for the purposes of estimation, an assumption must be made. To make the assumption about the contemporaneous causal order that we made previously – that satisfaction contemporaneously causes vote intention but not the reverse – we restrict  $\alpha_{2,0} = 0$ . We also restrict the variance–covariance matrix for  $\xi_t$  to be diagonal. Estimation of this model gives us the contemporaneous causal effect of PM satisfaction on vote  $\alpha_{1,0} = 0.299$  ( $p$  value  $< 0.001$ ). An increase in PM satisfaction of one percentage point causes vote intention for the incumbent government to immediately increase by 0.3 of a percentage point. This is equivalent to the results from the structural VAR(1) model estimated previously. We could also calculate the long-run effect, if we like.

Next we examine a dynamic seemingly unrelated regression equation in state-space form.

$$\begin{pmatrix} y_{1,t} \\ \vdots \\ y_{K,t-1} \end{pmatrix} = \begin{pmatrix} a_{1,t} \\ \vdots \\ a_{K,t-1} \end{pmatrix} + \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} \begin{pmatrix} x_{1,t} \\ \vdots \\ x_{k,t} \end{pmatrix}$$

$$\begin{pmatrix} a_{1,t} \\ \vdots \\ a_{K,t} \end{pmatrix} = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_k \end{pmatrix} \begin{pmatrix} a_{1,t-1} \\ \vdots \\ a_{K,t-1} \end{pmatrix} + \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_k \end{pmatrix}$$

$$+ \begin{pmatrix} 1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 1 \end{pmatrix} \begin{pmatrix} \xi_{1,t} \\ \vdots \\ \xi_{K,t} \end{pmatrix},$$

$$\xi_t \sim NID \left( 0, \begin{pmatrix} \sigma_{\xi_{1,1}}^2 & \dots & \sigma_{\xi_{1,k}}^2 \\ \vdots & \ddots & \vdots \\ \sigma_{\xi_{k,1}}^2 & \dots & \sigma_{\xi_{k,k}}^2 \end{pmatrix} \right)$$

This model includes  $K$  observation equations. They represent autoregressive processes that are independent of each other, except the errors are modelled as correlated. In doing so, this model relates multiple observed time processes through their error term.

As an example, we examine 114 weekly estimates of party support (measured through vote intention polls) for the Conservatives, Labour and Liberal Democrats in Britain after the General Election in 2010. We include the lag of inflation (measured as the year-over-year change in the CPI) as a covariate in the observation equation. Note that by including the exogenous variables in the observation equation, we are assuming that inflation has no effect beyond the immediate short-run effect. To see this, plug the state equation into the observation equation:

$$y_t = Ta_{t-1} + \lambda + R\xi_t + \beta x_t$$

and note that by the definition of the observation equation,

$$a_{t-1} = y_{t-1} - \beta x_{t-1}$$

Plugging this into the above and tidying things up a bit:

$$y_t = Ty_{t-1} + \lambda + \beta x_t - T\beta x_{t-1} + R\xi_t$$

The long-run effect of  $x_t$  is calculated as  $\frac{\beta - T\beta}{1 - T} = \beta$ . The long-run effect is the short-run effect. If we place inflation in the state equation, it would have a long-run effect of:  $\frac{\beta}{1 - T}$ .

The estimated parameters from the model are presented in Table 34.4. Before

interpreting the results, we note that the transition coefficient in the Conservative support state equation ( $a_{C,t}$ ) is indistinguishable from one. The transition coefficients for Labour and Liberal Democrat support are also very close to one. This suggests that these states may be unit root processes. Unlike MLE or OLS estimates of classical models, this is not actually a problem for the estimator. The estimator for a state-space model is not biased by a unit root process. However, it suggests that we could gain efficiency by fixing these parameters to one and removing the constant from the observation equations. The constant can be removed under these circumstances because the initial value for the unit root

**Table 34.4 British party support and inflation, 2010–12**

Equation		Parameter estimates
$y_{C,t}$	$a_{C,t}$	1.000 (fixed)
	$CPI_{t-1}$	-0.811 (0.232)**
	Trend	2.518 (7.155)
	Constant	1,984.580 (11,242.997)
$a_{C,t}$	$a_{C,t-1}$	1.001 (0.003)**
	$VAR(a_{C,t})$	0.129 (0.018)**
$y_{L,t}$	$a_{L,t}$	1.000 (fixed)
	$CPI_{t-1}$	0.670 (0.199)**
	Trend	-0.001 (0.017)
	Constant	37.490 (1.667)**
$a_{L,t}$	$a_{L,t-1}$	0.930 (0.013)**
	$VAR(a_{L,t})$	0.087 (0.013)**
$y_{LD,t}$	$a_{LD,t}$	1.000 (fixed)
	$CPI_{t-1}$	-0.048 (0.203)
	Trend	-0.011 (0.017)
	Constant	9.361 (1.568)**
$a_{LD,t}$	$a_{LD,t-1}$	0.912 (0.016)**
	$VAR(a_{LD,t})$	0.095 (0.013)**
	$COV(a_{C,t}, a_{L,t})$	-0.062 (0.013)**
	$COV(a_{C,t}, a_{LD,t})$	-0.003 (0.012)
	$COV(a_{L,t}, a_{LD,t})$	-0.034 (0.010)**
	$T$	100

\*  $p < 0.05$ ; \*\*  $p < 0.01$ ; standard errors in parentheses.

**Table 34.5 British party support and inflation, 2010–12**

Equation		Parameter estimates
$y_{C,t}$	$a_{C,t}$	1.000 (fixed)
	$CPI_{t-1}$	-0.719 (0.240)**
	Trend	-0.040 (0.037)
$a_{C,t}$	$a_{C,t-1}$	1.000 (fixed)
	$VAR(a_{C,t})$	0.137 (0.019)**
$y_{L,t}$	$a_{L,t}$	1.000 (fixed)
	$CPI_{t-1}$	0.547 (0.211)**
	Trend	0.101 (0.033)**
$a_{L,t}$	$a_{L,t-1}$	1.000 (fixed)
	$VAR(a_{L,t})$	0.106 (0.015)**
$y_{LD,t}$	$a_{LD,t}$	1.000 (fixed)
	$CPI_{t-1}$	-0.009 (0.253)
	Trend	-0.137 (0.039)**
$a_{LD,t}$	$a_{LD,t-1}$	1.000 (fixed)
	$VAR(a_{LD,t})$	0.152 (0.022)**
	$COV(a_{C,t}, a_{L,t})$	-0.041 (0.013)**
	$COV(a_{C,t}, a_{LD,t})$	-0.037 (0.015)*
	$COV(a_{L,t}, a_{LD,t})$	-0.063 (0.014)**
	$T$	100

\*  $p < 0.05$ ; \*\*  $p < 0.01$ ; standard errors in parentheses.

process acts as the constant. This explains the enormous standard error on the constant in the conservative support observation equation. Making these changes, we re-estimate the model parameters (Table 34.5).

We can see that the estimated errors have negative covariance. The increase in support for one party reduces support for another party. This effect is biggest for Labour and the Liberal Democrats: the left and centre (-left) parties. This is as we might expect. Turning to the effects of inflation, we see that increases in inflation reduce support for the Conservative party, increase support for the Labour Party and have no direct effect on support for the Liberal Democrats. A one-percentage point increase in inflation causes a 0.7-percentage point reduction in support for the Conservatives and a 0.6-percentage point increase in support for Labour. While statistically significant, these are not large effects. Because of the

relationship between Labour and Liberal Democrat support and Conservative and Liberal Democrat support, inflation does have a small indirect effect on Liberal Democrat support. Because the Labour–Liberal Democrat relationship is stronger, the indirect effect is negative.

Finally, we explore the dynamic-factor model. This is like a factor analysis where the observed variables are time series. Intuitively, the observed time series are observable measures of one or more underlying latent processes. These latent processes can be modelled as dynamic processes (autoregressive, random walk, trending deterministically, etc). As a state-space model, the latent processes are modelled as states and the observed time series as observations (Jackman, 2005; Pickup and Johnston, 2008). As an example, let us say we wish to model the proportion of individuals intending to vote for the Conservatives

in the next UK general election. The underlying state is the percentage of British voters that would vote for the Conservative party if an election were held tomorrow. Through public-opinion polls, we have multiple observations at multiple time points (weeks). We use 95 weeks of data starting the week of the 2017 UK general election. There are 279 polls from 11 polling houses during this period.

$$\begin{pmatrix} y_{1,t} \\ \vdots \\ y_{K,t-1} \end{pmatrix} = a_t + \delta_k + \varepsilon_t, \quad \varepsilon_t \sim NID(0, \sigma_{\varepsilon,t}^2)$$

$$a_t = \alpha_t a_{t-1} + \lambda_k + \xi_t, \quad \xi_t \sim NID(0, \sigma_{\xi}^2)$$

Each  $y_{kt}$  represents the observed measure of  $a_t$  from polling house  $k$  at time  $t$ . The observed measure is the proportion of individuals in the poll indicating they would vote for the Conservatives. This measure of Conservative-party support from polling house  $k$  at time  $t$  equals the latent support for the Conservative party at time  $t$  ( $a_t$ ) plus the systematic measurement error (bias) for polling house  $k$  ( $\delta_k$ ) and random measurement error ( $\varepsilon_t$ ). This measurement error has a variance that varies over time. Specifically, it is a function of sample size and the observed support:  $\sigma_{\varepsilon,t}^2 = \frac{y_{kt}(1-y_{kt})}{N}$ .

The latent support is modelled as first-order autoregressive.

If we allow all  $k$  observation equations to have a biased term ( $\delta_k$ ), the model is not identified. We can estimate the systematic bias of one house relative to another but we need to choose a house as the reference point for the estimation of the bias. We have chosen the pollster YouGov as the reference. Estimation of the model produces the bias terms (Table 34.6). We see that the house biases range from a systematic one-percentage point overestimation of support

**Table 34.6 Systematic polling house biases**

Poll house	Bias (percentage points)
1	Reference
2	-1.80 (0.37)
3	-1.04 (0.36)
4	-1.62 (0.26)
5	-2.57 (0.27)
6	0.55 (0.73)
7	-0.15 (0.37)
8	0.93 (0.94)
9	-0.76 (0.21)
10	0.66 (0.21)
11	0.0022 (0.55)

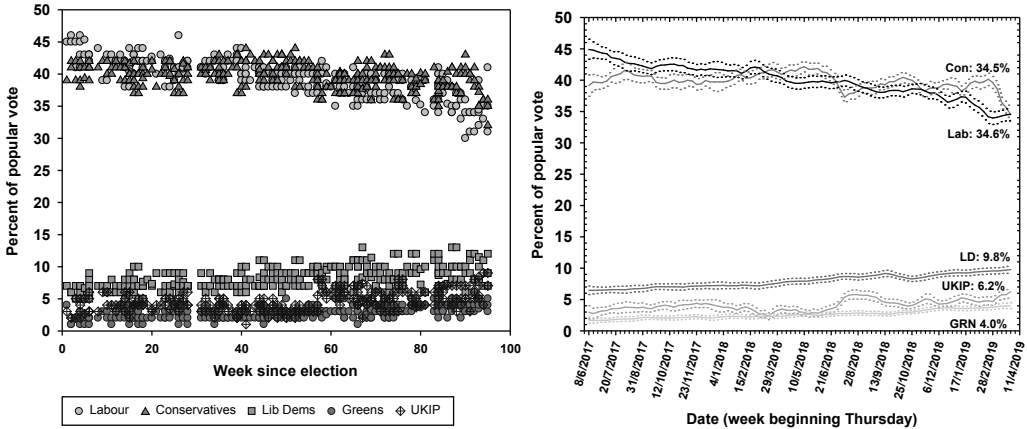
Standard errors in parentheses.

(relative to YouGov) for the Conservatives to a 2.5-percentage point underestimation.

We can next estimate the smoothed state ( $a_t$ ) and its root-mean-squared error. This is the estimate of public opinion with random measurement error filtered out. To produce the estimates, we anchored the systematic bias on the average of all pollsters. We then plotted the estimated smoothed state with 95% confidence interval and compared it to the original poll results. Figure 34.4 does this for the Conservative party, Labour, the Liberal Democrats, UKIP, and the Green party.<sup>16</sup> We can see that the estimated smoothed states contain far less variance than the original poll data.

The estimated root-mean-squared errors are small so it is difficult to see, but these confidence intervals do vary over time. This is an important difference between state-space and most classical time series models. In classical models, there is typically a single root-mean-squared-error term and so the confidence intervals are of a constant width over time. When a state-space model is used for forecasting, the time varying root-mean-squared errors also allow the confidence intervals of the estimated forecast to increase as the forecasts move further away from the





**Figure 34.4 Smooth states for vote intention**

last observed data point. This provides a more accurate estimate of the uncertainty in the forecast. A final distinction between the state-space approach and classical estimators is the way non-stationary processes are modelled. Classical estimators require us to detect  $I(1)$  processes in advance and adjust the model accordingly (typically by first differencing the variables). As  $I(1)$  processes do not bias state-space estimates, we can detect them in the model estimates and model as  $I(1)$  if necessary.

Overall, we can see that state-space models are a very flexible way of expressing and estimating dynamic systems of equations. The greatest drawback currently is that programs/packages/commands for estimating them are still very limited and are often very complex. For example, a missing data problem required the last model to be estimated as a Bayesian model in WinBUGS.

**Notes**

- 1 The estimator will also be inconsistent (Greene, 2012: 225).
- 2 However, it should be noted that modelling the third type of data generating process (including the past value of  $Y_t$ ) will always lead to a violation of strict exogeneity.

- 3 By encompass, we mean that the data model contains at least as many lags of the dependent and independent variables as the DGP.
- 4 It is helpful to keep in mind the difference between a structural equation and simultaneous equation model. In a structural model, the structure is based on prior theoretical ideas/assumptions about how the variables are expected to be causally related. If the prior theoretical assumptions are valid, a structural model allows us to come to conclusions regarding causal relationships. In a simultaneous equation model, where the form is not assumed to reflect the actual causal relationships between the variables, we should not give the estimated parameters a causal interpretation.
- 5 Since  $y_{2t}$  is a function of  $y_{1t}$  and therefore  $\mu_{1t}$ .
- 6 The lag order of the VAR was determined using the following lag-order selection statistics: final prediction error, Akaike's information criterion, Schwarz's Bayesian information criterion (SBIC), Hannan and Quinn information criterion (HQIC) and a likelihood-ratio test.
- 7 This is done because unemployment rates are reported for a three-month period with a two-month lag.
- 8 For a SVAR with  $K$  endogenous variables,  $\frac{2K^2 - K(K+1)}{2}$  restrictions must be placed on the parameters of  $A$  and  $B$ . This is a necessary but not sufficient condition.
- 9 For a SVAR with  $K$  endogenous variables,  $\frac{K^2 - K(K+1)}{2}$  restrictions must be placed on the parameters of  $C$ . This is a necessary but not sufficient condition.

- 10 Caution: the restrictions necessary to identify a SVAR are often difficult to justify theoretically. As social scientists, we very much want to give a causal interpretation to our analysis, BUT the discussion of our results should always include the appropriate caveat: 'The orthogonalized/structural impulse response function can only be given a causal interpretation to the extent that our identifying restrictions are correct'. On the upside, SVARs allow us to make identifying restrictions that are less arbitrary than those implicitly assumed in univariate time-series models. In addition to providing the appropriate caveat, it is also good practice to test different plausible identifying restrictions and note if this results in major differences.
- 11 Assuming we have determined that  $y_t$  and  $x_t$  are each integrated  $I(1)$  processes (e.g., using the Dickey–Fuller test), the Engle–Granger test of cointegration is as follows. As the cointegration parameter  $\beta$  is unknown, we first have to estimate  $\beta$ . We can estimate  $y_t = \alpha + \beta x_t + \mu_t$  (with or without  $\alpha$  or other deterministic terms like a trend) and calculate  $\hat{z}_t = \hat{\mu}_t = y_t - \hat{\alpha} - \hat{\beta}x_t$ . After estimating  $\beta$  we apply the Dickey–Fuller test to  $\hat{z}_t$ , in order to determine if it is  $I(0)$ . This requires different critical values than the usual Dickey–Fuller test; a higher threshold is necessary to take into account that the  $\beta$  is estimated and not known ahead of time (these can be looked up; MacKinnon, 2010). If we can reject a unit-root process for  $\hat{z}_t$ , then  $y_t$  and  $x_t$  are cointegrated.
- 12 Note that when we use a single equation ECM, we assume that only  $y_t$  responds when it and  $x_t$  are out of equilibrium (we are assuming that  $x_t$  is weakly exogenous). The VECM allows us to relax this assumption.
- 13 It can be also used to include time varying coefficients (Matthews and Pickup, 2019) and for non-linear dynamic models (Durbin and Koopman, 2001).
- 14 It can be shown that as  $t \rightarrow \infty$ ,  $F_t$  and  $P_t$  converge to constant values so does  $K_t$ ; when this happens, the filter is said to be in a steady state.
- 15 If the state is a stationary process, there are different techniques for specifying the initial element(s) and the corresponding standard error(s) – for example, we might assume the process was in equilibrium before our first observation of it. If the state is a non-stationary process, it is common to use *diffuse* initial elements. They are diffuse if they are estimated from the observed data and the random distributions used to represent our prior beliefs about them have an infinite variance (we know nothing about it).
- 16 Support for each party is estimated in a separate model but this could have all been done in a single model. This would allow us to relate the parameters for each party to each other (e.g., through the random error term).

## REFERENCES

- Banerjee, A., J. Dolado, J.W. Galbraith and D.F. Hendry. 2003. *Cointegration, Error Correction, and the Econometric Analysis of Nonstationary Data*. Oxford: University of Oxford Press.
- Box, G.E., G.M. Jenkins and G.C. Reinsel. 1994. *Time Series Analysis – Forecasting and Control*. Third edition. Upper Saddle River, New Jersey: Prentice-Hall.
- Brandt, P.T. and J.T. Williams. 2007. *Multiple Time Series Models*. Thousand Oaks, California: Sage.
- Commandeur, J.J.F. and S.J. Koopman. 2007. *An Introduction to State Space Time Series Analysis*. Oxford: Oxford University Press.
- Durbin, J. and S.J. Koopman. 2001. *Time Series Analysis by State Space Methods*. Oxford: Oxford University Press.
- Freeman, J.R., J.T. Williams and T. Lin. 1989. 'Vector Autoregression and the Study of Politics'. *American Journal of Political Science* 33(4):42–877.
- Greene, W. 2012. *Econometric Analysis*. Seventh edition. Upper Saddle River, New Jersey: Prentice-Hall.
- Harvey, A. 1993. *Time Series Models*. Second edition. Cambridge, Massachusetts: The MIT Press.
- Hendry, D.F. 2003. *Dynamic Econometrics*. Oxford: Oxford University Press.
- Jackman, S. 2005. 'Pooling the Polls over an Election Campaign'. *Australian Journal of Political Science* 40(4):499–517.
- Johansen, S. 1995. *Likelihood-Based Inference in Cointegrated Vector Autoregressive Models*. Oxford: Oxford University Press.
- Matthews, J.S. and M. Pickup. 2019. 'Rational Learners or Impervious Partisans? Economic News and Partisan Bias in Economic Perceptions'. *Canadian Journal of Political Science*. 52(2):303–32.

- MacKinnon, James G. 2010. 'Critical Values For Cointegration Tests'. Working Paper 1227, Economics Department, Queen's University.
- Moore, W.H. and D.J. Lanoue. 2003. 'Domestic Politics and US Foreign Policy: A Study of Cold War Conflict Behaviour'. *The Journal of Politics* 65(2):376–96.
- Pickup, M. 2010. 'Better Know Your Dependent Variable: A Multination Analysis of Government Support Measures in Economic Popularity Models'. *British Journal of Political Science* 40(2): 449–68.
- Pickup, M. and R. Johnston. 2008. 'Campaign Trial Heats as Election Forecasts: Measurement Error and Bias in 2004 Presidential Campaign Polls'. *International Journal of Forecasting* 24(2):270–82.

# Duration Analysis

Kentaro Fukumoto

## INTRODUCTION

Political scientists sometimes study time to an event such as the end of a legislator's career, cabinet termination, and a war's conclusion. Duration analysis (sometimes called survival analysis or event history analysis) is a particular class of modeling for time to event.

The first section, which focuses on single duration without considering censoring, emphasizes that this class of analyses models not just *a single time point* when an event happens, but the *duration* where a unit is at risk of the event. By considering this way and comparing with conventional models, it will be more natural to derive seemingly weird distribution functions, in particular by way of hazard and not by way of the mean of event time like usual linear models. I represent non-parametric, parametric, and semi-parametric models in terms of both continuous and discrete time, discuss proportionality, and

explain frailty models. I also translate some technical terms specific to duration analysis into more general statistical terms.

The second section introduces parallel durations and explains competing risks, right censoring, and split population models. I emphasize that *censoring is another event* in which scholars have no interest. I also clarify independence assumptions scholars usually implicitly make and elaborate on how we can relax such assumptions by using frailties, seemingly unrelated regressions, and copula functions.

The third section considers serial durations and deals with left/interval truncation/censoring, time varying covariates, repeated events, and multi-state models. I call attention to whether durations are (conditionally) independent and, if not, how dependence among durations are modeled. The discussion of this section will also underscore how useful it is to analyze not *a point of time* but the *duration*.

This chapter explains duration analysis in unconventionally way, while it aims to shed light on the motivation as to why we model duration as we do and the assumption behind the scene and encourage readers to develop new models, in particular those which take into consideration dependence among multiple durations. Readers who are unfamiliar with duration analysis may refer to textbooks such as Box-Steffensmeier and Jones (2004) for ordinary guidance.

**SINGLE DURATION**

**Non-Parametric Model**

Suppose that  $T_i$  is a real random variable of the time when an event happens to unit  $i \in \mathcal{I} \equiv \{1, 2, \dots, \bar{i}\}$  and the cumulative distribution function (CDF) of  $T_i$  is denoted by  $F_i(t) \equiv Pr(T_i \leq t)$  for a real number  $t$ . We make the following two basic assumptions:

**Assumption 1** (Positive Time).

$$T_i > 0.$$

**Assumption 2** (Independence Across Units).

$$T_i \perp\!\!\!\perp T_{i'} \text{ if } i \neq i'.$$

For identification, we had better assume  $\lim_{t \rightarrow \infty} F_i(t) = 1$ , that is, the event certainly happens someday.

**Discrete time framework**

Suppose that there are fixed cutoff points  $t_j$ 's where  $j$  is a positive integer,  $t_j$  is a positive real number such that  $t_{j-1} < t_j$ , and  $t_0 \equiv 0$ . They might be located at regular intervals (e.g.,  $t_j \equiv j$  in years). Denote the  $j$ -th interval by  $\mathcal{T}_j \equiv (t_{j-1}, t_j]$ .

A conventional model pays attention to a discrete event time variable  $J_i$  which is the positive integer such that  $T_i \in \mathcal{T}_{J_i}$  (e.g.,  $J_i \equiv [T_i]$ ). It follows:

$$\begin{aligned} Pr(J_i = j) &= Pr(T_i \in \mathcal{T}_j) \\ &= F_i(t_j) - F_i(t_{j-1}) \\ &\equiv \Delta F_{ij}. \end{aligned} \tag{1}$$

A duration model focuses not just on a single event time  $J_i$  but on a series of event indicators  $E_i$  where an event indicator is defined as:

$$E_{ij} \equiv \begin{cases} 0 & \text{if } j < J_i \quad \left( \begin{array}{l} \text{the event will} \\ \text{occur after } \mathcal{T}_j \end{array} \right), \\ 1 & \text{if } j = J_i \quad \left( \begin{array}{l} \text{the event does} \\ \text{occur during } \mathcal{T}_j \end{array} \right), \end{cases}$$

for  $j \in \mathcal{J}_i \equiv \{1, 2, \dots, J_i\}$ , and the vector of event indicators is defined as  $E_i \equiv (E_{i1}, E_{i2}, \dots, E_{iJ_i})$ . Define the survivor function as:

$$\begin{aligned} S_i(t) &\equiv Pr(T_i > t) \\ &= 1 - F_i(t), \end{aligned} \tag{2}$$

which is called the complementary cumulative distribution function in the more general statistical literature. For  $j \geq 2$ , denote  $\mathcal{J}_{< j} \equiv \{1, 2, \dots, j - 1\}$  and  $\mathcal{J}_{< 1} \equiv \emptyset$ . It follows:

$$\begin{aligned} Pr(E_{ij} = 1 \mid \forall j' \in \mathcal{J}_{< j}, E_{ij'} = 0) &= Pr(T_i \in \mathcal{T}_j \mid t_{j-1} < T_i) \\ &= \frac{\Delta F_{ij}}{S_i(t_{j-1})} \quad (\because \text{Equations 1 and 2}) \\ &\equiv \Delta H_{ij}, \end{aligned} \tag{3}$$

which is called the discrete hazard rate, and, thus:

$$\begin{aligned} Pr(E_{ij} = 0 \mid \forall j' \in \mathcal{J}_{< j}, E_{ij'} = 0) &= 1 - \Delta H_{ij} \\ &= \frac{S_i(t_j)}{S_i(t_{j-1})}. \end{aligned} \tag{4}$$

Therefore,

$$\begin{aligned}
 \Pr(E_i) &= \Pr(\forall j' \in \mathcal{J}_{<J_i}, E_{ij'} = 0, E_{iJ_i} = 1) \\
 &= \left\{ \prod_{j=1}^{J_i-1} \Pr(E_{ij} = 0 \mid \forall j' \in \mathcal{J}_{<J_i}, E_{ij'} = 0) \right\} \\
 &\quad \Pr(E_{iJ_i} = 1 \mid \forall j' < J_i, E_{ij'} = 0) \\
 &= \left\{ \prod_{j \in \mathcal{J}_i} (1 - \Delta H_{ij}) \right\} \frac{\Delta H_{iJ_i}}{1 - \Delta H_{iJ_i}} \quad (5) \\
 &\quad (\because \text{Equations 3 and 4}) \\
 &= \exp \left\{ \underbrace{\sum_{j \in \mathcal{J}_i} \log(1 - \Delta H_{ij})}_{\text{duration}} \right\} \underbrace{\frac{\Delta H_{iJ_i}}{1 - \Delta H_{iJ_i}}}_{\text{event}},
 \end{aligned}$$

where  $\prod_{j=1}^0 = 1$ .

It is easy to show that Equations 1 and 5 are equal to each other. That is, the conventional and duration models approach the same object from different points of view. On the one hand, Equation 1 expresses the probability that an event happens during  $(t_{J_{i-1}}, t_{J_i}]$ . On the other hand, Equation 5 implies the probability that an event *does not* happen during  $(0, t_{J_{i-1}}]$  and *does* happen during  $(t_{J_{i-1}}, t_{J_i}]$ . Thus, a quantity of natural interest for duration model is not so much  $\Delta F_j$  as  $\Delta H_j$ .

Analysts might have an intrinsic interest in  $\Delta H_j$  rather than  $\Delta F_j$ . For instance, if  $J_i$  is the number of terms lawmaker  $i$  serves,  $\Delta H_{ij}$  represents the probability to be re-elected at time  $j$  given that the lawmaker survived the previous elections up to  $j - 1$ . Theoretically, scholars may expect that senior members are more likely to be reelected than junior ones, that is,  $\Delta H_j > \Delta H_{j'}$  for  $j > j'$ , while  $\Delta F_j = S_{j-1} \Delta H_j$  can be smaller than  $\Delta F_{j'} = S_{j'-1} \Delta H_{j'}$  because of  $S_{j-1} \leq S_{j'-1}$ .

### Continuous time framework

Suppose that  $F_i(t)$  is left differentiable and denote its left derivative by  $f_i(t)$ . A conventional model will analyze

$$p(T_i = t) = f_i(t). \quad (6)$$

A duration model focuses not just on a *point* of time  $T_i$ , but on the *duration* of time  $\mathcal{T}_i \equiv (0, T_i]$ . Define the hazard rate by:

$$\frac{f_i(t)}{S_i(t)} \equiv h_i(t), \quad (7)$$

which is equivalent to the inverse of the Mills ratio in more general statistical literature. Let  $\Delta t_{.j} \equiv t_{.j} - t_{.j-1}$ . It follows that:

$$\begin{aligned}
 &\lim_{\Delta t_{.j} \downarrow 0} \frac{\Delta H_{ij}}{\Delta t_{.j}} \\
 &= \lim_{\Delta t_{.j} \downarrow 0} \frac{F_i(t_{.j}) - F_i(t_{.j} - \Delta t_{.j})}{\Delta t_{.j}} \frac{1}{S_i(t_{.j} - \Delta t_{.j})} \\
 &= \frac{f_i(t_{.j})}{S_i(t_{.j})} \quad (8) \\
 &= h_i(t_{.j}).
 \end{aligned}$$

Denote  $\Delta t \equiv \max_j \Delta t_{.j}$ ; It follows that:

$$\begin{aligned}
 p(\mathcal{T}_i) &\equiv \lim_{\Delta t \downarrow 0} \frac{1}{\Delta t_{J_i}} \Pr(E_i) \\
 &= \lim_{\Delta t \downarrow 0} \exp \left\{ \sum_{j \in \mathcal{J}_i} \frac{\log(1 - \Delta H_{ij})}{\Delta t_{.j}} \Delta t_{.j} \right\} \\
 &\quad \frac{\Delta H_{iJ_i}}{\Delta t_{.J_i} (1 - \Delta H_{iJ_i})} \quad (\because \text{Equation 5}) \\
 &= \exp \left[ \int_0^{T_i} \frac{d \log \{S_i(t)\}}{dt} dt \right] h_i(T_i) \\
 &\quad (\because \text{Equations 4 and 8, } \lim_{\Delta t \downarrow 0} (1 - \Delta H_{iJ_i}) = 1) \\
 &= \underbrace{\exp\{-H_i(\mathcal{T}_i)\}}_{\text{duration}} \underbrace{h_i(T_i)}_{\text{event}},
 \end{aligned} \quad (9)$$

where the cumulative hazard rate is defined as

$$H_i(\mathcal{T}) \equiv \int_{\mathcal{T}} h_i(t) dt, \quad (10)$$

for any set of positive values of  $t$ ,  $\mathcal{T} \subseteq \{t \mid t > 0\}$ .<sup>1</sup>

It is a good exercise to show that Equations 6 and 9 are equal to each other. On the one hand, Equation 6 deals with the probability density that an event happens at  $T_i$ . On the other, Equation 9 implies the probability density that an event *does not* happen during  $(0, T_i)$  and *does* happen at  $T_i$ . Thus, a quantity of natural interest for the duration model is not so much  $f(t)$  as  $h(t)$ .

Obviously, Equation 9 looks similar to its discrete time framework version, Equation 5. Since information on event timing is coarsened in the discrete time framework, estimation is less efficient than in the continuous time framework. Nonetheless, if we can observe not  $T_i$  but  $J_i$ , or if  $F_i(t)$  is not left differentiable (e.g., the event can happen only at finite time points), only the discrete time framework is available.

**Parametric Model**

**Accelerated failure time model**

Conventional models would characterize  $F_i(t)$  only by a vector of sufficient statistics,  $\psi_i \equiv (\psi_i^{(1)}, \psi_i^{(2)}, \dots, \psi_i^{(A)})$ , that is,  $F_i(t) = F(t|\psi_i)$ . A typical example is the log normal distribution:

$$F_{\text{Log Normal}}(t|\psi_i) \equiv \Phi(z_i), \tag{11}$$

where  $\Phi(\cdot)$  is the CDF of the standard normal distribution,  $z_i$  is the standardized logged time such that

$$z_i \equiv z(t|\mu_i, \sigma) \equiv \frac{\log(t) - \mu_i}{\sigma}, \tag{12}$$

$\mu_i$  and  $\sigma > 0$  are location and scale parameters, respectively, and  $\psi_i = (\mu_i, \sigma)$ . Instead, we may use the log logistic distribution  $F_{\text{Log Logistic}}(t|\psi_i) \equiv \Lambda(z_i) \equiv 1/\{1 + \exp(-z_i)\}$ .

Furthermore, scholars are usually interested in the distribution of  $T_i$  conditioned on a row vector of covariates  $x_i \equiv (x_i^{(1)}, x_i^{(2)}, \dots, x_i^{(B)})$  by modeling  $\psi_i^{(a)} = \psi^{(a)}(x_i, \beta) = g^{-1}(x_i\beta)$  where  $\beta$  is a column vector of coefficient parameters not indexed by  $i$  and  $g(\cdot)$  is a link function. (For ease of presentation, I do not annotate transposition of a vector because it

is obvious in the context.) If  $\psi_i = (\mu_i, \sigma)$ , one may model

$$\mu_i = x_i\beta, \tag{13}$$

where the identity link function  $g_{\text{identity}}(\mu) \equiv \mu$  is used. Thus,  $F_i(t) = F(t|x_i, \theta)$  where  $\theta = (\beta, \sigma)$  is a vector of unit invariant extended parameters and variation in  $F_i(t)$  is explained by unit variant  $x_i$  only. If we use Equations 12 and 13, we can rewrite them as

$$\log(T_i) = x_i\beta + \sigma Z_i, \tag{14}$$

where  $\sigma Z_i$  works as an error term and, in the case of log normal model (Equation 11), follows normal distribution with mean 0 and variance  $\sigma^2$ . This representation is called the accelerated failure time model because, if the  $b$ -th coefficient is positive ( $\beta^{(b)} > 0$ ), ‘failure time’  $T$  increases in (or is ‘accelerated’ by) the  $b$ -th covariate,  $x^{(b)}$ . In general, this model is called log linear regression, which most scholars are familiar with.

We can estimate  $\theta$  by maximizing the likelihood: in the case of the discrete time framework,

$$l(\theta|J, x) \propto \prod_{i \in \mathcal{I}} \Delta F_{J_i}(x_i, \theta) \tag{:: Assumption 2 and Equation 1},$$

where  $J \equiv (J_1, J_2, \dots, J_T)$  and  $x \equiv (x_1, x_2, \dots, x_T)$ , and in the case of the continuous time framework,

$$l(\theta|T, x) \propto \prod_{i \in \mathcal{I}} f(T_i|x_i, \theta) \tag{:: Assumption 2 and Equation 6},$$

where  $T \equiv (T_1, T_2, \dots, T_T)$ . Note that, in the discrete time framework, unlike ordered logit or probit models, cutoff points ( $t_j$ 's) are not parameters to be estimated but known values.

**Hazard model**

For a while, we consider the continuous time framework. According to Equations 2, 4, and 9, it holds

$$S_i(t) = \exp\{-H_i(\mathcal{T})\}, \tag{15}$$

where  $\mathcal{T} \equiv (0, t]$ . Thus, thanks to Equation 9, the likelihood of unit  $i$  is proportional to

$$f(T_i | \psi_i) = \exp\{-H(T_i | \psi_i)\}h(T_i | \psi_i).$$

Accordingly, the duration model would specify  $H(\mathcal{T} | \psi_i)$  or  $h(t | \psi_i)$  instead of  $F(t | \psi_i)$ . The most popular distribution is Weibull:

$$\begin{aligned} H_{\text{Weibull}}(T | \psi_i) &\equiv \exp(z_i) \\ &= \mu_i^* t^{\sigma^*}, \end{aligned} \tag{16}$$

where  $\mu_i^* \equiv \exp(-\mu_i / \sigma)$  and  $\sigma^* \equiv 1/\sigma^2$ . Accordingly, the hazard rate is

$$\begin{aligned} H_{\text{Weibull}}(t | \psi_i) &= \frac{dH_{\text{Weibull}}(\mathcal{T} | \psi_i)}{dt} \\ &= \sigma \mu_i^* t^{\sigma^* - 1}, \end{aligned} \tag{17}$$

If  $\sigma^* > 1$  (or  $\sigma^* < 1$ ), the hazard rate  $h(\cdot)$  increases (or decreases) in  $t$ . For example, parliament dissolution becomes more likely as the term approaches expiration. If  $\sigma^* = 1$ , Weibull distribution is reduced to exponential distribution and the hazard rate is constant over  $t$ :

$$h_{\text{exponential}}(t | \psi_i) = \mu_i^*. \tag{18}$$

Usually, we condition on  $x_i$  by way of Equation 13. Thus, we estimate  $\theta$  by maximizing the likelihood:

$$\begin{aligned} l(\theta | \mathcal{T}, x) &\propto \prod_{i \in \mathcal{I}} p(\mathcal{T}_i | x_i, \theta) \\ &\quad (\because \text{Assumption 2}) \\ &= \prod_{i \in \mathcal{I}} \underbrace{\exp\{-h(\mathcal{T}_i | x_i, \theta)\}}_{\text{duration}} \underbrace{h(\mathcal{T}_i | x_i, \theta)}_{\text{event}}, \\ &\quad (\because \text{Equation 9}) \end{aligned}$$

where  $\mathcal{T} \equiv (\mathcal{T}_1, \dots, \mathcal{T}_T)$ . Off-the-shelf statistical software often reports estimates of  $\beta^* \equiv -\beta/\sigma$  and  $\sigma^*$  instead of  $\beta$  and  $\sigma$ . If  $\beta^{*(b)} > 0$ , the hazard rate  $h(\cdot)$  increases in  $x^{(b)}$ , which implies  $T$  decreases in  $x^{(b)}$

( $\because \beta^{(b)} < 0$ ). However, some analysts might be interested in the effect of  $x$  not on  $h(\cdot)$  but on  $T$  and, for them, the package result can be misleading. They would prefer an accelerated failure time representation of the same model (Equation 14) where  $Z_i$  looks like a scaled error term and follows type 1 extreme value (namely, Gumbel maximum) distribution:

$$\begin{aligned} Z_i &\sim \Xi(z_i) \\ &\equiv 1 - \exp\{-\exp(z_i)\} \\ &= F_{\text{Weibull}}(t | \psi_i) \\ &\quad (\because \text{Equations 2, 15, and 16}). \end{aligned} \tag{19}$$

### Cox model

The most popular method of duration analysis is the Cox model because it only has to specify the hazard rate up to the ‘baseline hazard’. In this sense, this model is called semi-parametric.

For  $t \in \mathcal{T}_i$ , an event indicator in the continuous time framework is defined as

$$E_i(t) \equiv \begin{cases} 0 & \text{if } t < T_i \\ \text{(the event will occur after } t), \\ 1 & \text{if } t = T_i \\ \text{(the event does occur at } t). \end{cases}$$

Define the ‘risk set’ as the set of units whose events have not happened before  $t$  (in the continuous time framework) or  $t_{j-1}$  (in the discrete time framework) and, thus, which are still at risk of the event at  $t$  or during  $\mathcal{T}_j$ :

$$\begin{aligned} \mathcal{I}(t) &\equiv \{i | t \in \mathcal{T}_i\}, \\ \mathcal{I}_j &\equiv \{i | j \in \mathcal{J}_i\}. \end{aligned} \tag{20}$$

We make two assumptions.

**Assumption 3** (Proportional Hazard):

$$h_i(t) = h_0(t)\psi_i.$$

We call  $h_0(t)$  the baseline hazard, which is constant across unit  $i$ 's.



**Assumption 4** (No Tied Events):

$$T_i \neq T_{i'} \text{ if } i \neq i'.$$

Thus, for any  $i$ ,  $\sum_{i' \in \mathcal{I}(T_i)} E_{i'}(T_i) = 1$  and, for sufficiently small  $\Delta t_{i,j}$ , it follows  $\sum_{i' \in \mathcal{I}_{i,j}} E_{i'} = 1$ .

If  $i \in \mathcal{I}_j$ , we calculate the probability that unit  $i$  has an event during  $\mathcal{T}_j$  conditioned that one unit in the risk set  $\mathcal{I}_j$  has an event during  $\mathcal{T}_j$ :

$$\begin{aligned} & \Pr(E_{ij} = 1 \mid \sum_{i' \in \mathcal{I}_j} E_{i'} = 1) \\ = & \Pr(E_{ij} = 1, \forall i' \in \mathcal{I}_j \setminus \{i\}, E_{i'} = 0) / \\ & \sum_{i' \in \mathcal{I}_j} \Pr(E_{i'} = 1, \forall i'' \in \mathcal{I}_j \setminus \{i'\}, E_{i''} = 0) \\ & (\because i \in \mathcal{I}_j) \\ = & \left\{ \Pr(E_{ij} = 1 \mid E_{ij} \leq 1) \right. \\ & \left. \prod_{i' \in \mathcal{I}_j \setminus \{i\}} \Pr(E_{i'} = 0 \mid E_{i'} \leq 1) \right\} \\ & / \left\{ \sum_{i' \in \mathcal{I}_j} \Pr(E_{i'} = 1 \mid E_{i'} \leq 1) \right. \\ & \left. \prod_{i'' \in \mathcal{I}_j \setminus \{i'\}} \Pr(E_{i''} = 0 \mid E_{i''} \leq 1) \right\} \\ & (\because \text{Assumption 2 and Equation 20,} \end{aligned}$$

dividing numerator and denominator (21)

$$\begin{aligned} & \text{by } \prod_{i' \in \mathcal{I}_j} \Pr(E_{i'} \leq 1), i \in \mathcal{I}_j \\ = & \left\{ \Delta H_{ij} \prod_{i' \in \mathcal{I}_j \setminus \{i\}} (1 - \Delta H_{i'}) \right\} / \\ & \left\{ \sum_{i' \in \mathcal{I}_j} \Delta H_{i'} \prod_{i'' \in \mathcal{I}_j \setminus \{i'\}} (1 - \Delta H_{i''}) \right\} \\ & (\because \text{Equations 3 and 4}) \\ = & \frac{\Delta H_{ij}}{1 - \Delta H_{ij}} / \sum_{i' \in \mathcal{I}_j} \frac{\Delta H_{i'}}{1 - \Delta H_{i'}} \\ & (\because \text{dividing numerator and denominator by} \\ & \prod_{i' \in \mathcal{I}_j} (1 - \Delta H_{i'}), i \in \mathcal{I}_j). \end{aligned}$$

(Note that  $\Delta H_{ij}/(1 - \Delta H_{ij})$  is odds of an event.) It follows that, if  $i \in \mathcal{I}(t)$ ,  $t \in \mathcal{T}_j$ ,

$$\begin{aligned} & \Pr\{E_i(t) = 1 \mid \sum_{i' \in \mathcal{I}(t)} E_{i'}(t) = 1\} \\ = & \lim_{\Delta t_j \downarrow 0} \Pr(E_{ij} = 1 \mid \sum_{i' \in \mathcal{I}_j} E_{i'} = 1) \quad (\because t \in \mathcal{T}_j) \\ = & \lim_{\Delta t_j \downarrow 0} \frac{\Delta H_{ij}}{\Delta t_j (1 - \Delta H_{ij})} / \\ & \sum_{i' \in \mathcal{I}_j} \frac{\Delta H_{i'}}{\Delta t_j (1 - \Delta H_{i'})} \quad (22) \\ & (\because \text{Equation 21, } i \in \mathcal{I}_j) \\ = & h_i(t) / \sum_{i' \in \mathcal{I}(t)} h_{i'}(t) \\ & (\because \text{Equation 8, } \lim_{\Delta t_j \downarrow 0} (1 - \Delta H_{i'}) = 1) \\ = & \frac{\psi_i}{\sum_{i' \in \mathcal{I}(t)} \psi_{i'}} \quad (\because \text{Assumption 3}). \end{aligned}$$

Thus, we cancel out the baseline hazard and do not have to estimate it. This equation may also remind us of multinomial or conditional logit models derived by assuming that random utility follows the type 1 extreme value distribution (Equation 19). If we model  $\psi_i = \exp(x_i \beta^*)$ , where  $x_i$  does not contain the constant term, we define the partial likelihood function as

$$\begin{aligned} & l^{\text{partial}}(\beta^* \mid \mathcal{I}(T), \mathbf{x}) \\ \propto & \prod_{i \in \mathcal{I}} \Pr \left\{ E_i(T_i) = 1 \mid \sum_{i' \in \mathcal{I}(T_i)} E_{i'}(T_i) = 1 \right\} \\ = & \prod_{i \in \mathcal{I}} \frac{\exp(\mathbf{x}_i \beta^*)}{\sum_{i' \in \mathcal{I}(T_i)} \exp(\mathbf{x}_{i'} \beta^*)} \\ & \text{event} \\ & (\because \text{Equation 22}), \end{aligned} \quad (23)$$

where  $\mathcal{I}(T) \equiv (\mathcal{I}(T_1), \mathcal{I}(T_2), \dots, \mathcal{I}(T_T))$ . Equation 23 uses information of risk sets at event time  $\mathcal{I}(T)$  alone, not sets of duration  $\mathcal{I}$ , and lacks a duration term unlike Equations 5 and 9. Luckily, even though we do not specify the baseline hazard, by maximizing

Equation 23, we obtain an estimate of  $\beta^*$  which has the same properties (such as consistency under the regularity conditions) as a maximum likelihood estimate.

We may relax Assumptions 3 and 4. As for Assumption 4, I refer readers to other textbooks (for example Box-Steffensmeier and Jones, 2004: 53–9). Below, we elaborate on Assumption 3.

**Proportionality**

Suppose that we arbitrarily choose a baseline value such as zero ( $x_i = x_0 \equiv (0, 0, \dots, 0)$ ). The corresponding hazard rate is called the baseline hazard and denoted by  $h_0(t) \equiv h(t|x_0, \theta)$ . If we can factor

$$h(t | x_i, \theta) = h_0(t)\psi(x_i, \theta), \tag{24}$$

it is called a proportional hazard rate. It is easy to see that Assumption 3 is a general case of Equation 24. If we use Equation 13, Weibull distribution is a proportional hazard model where  $h_{0, \text{Weibull}}(t | \theta) = \sigma^* t^{\sigma^* - 1}$  and

$$\begin{aligned} \psi(x_i, \theta) &= \mu_i^* \equiv \exp(-\mu_i / \sigma) \\ &= \exp(-x_i \beta / \sigma) = \exp(x_i \beta^*). \end{aligned} \tag{25}$$

Log normal (Equation 11) and logistic models are not proportional hazard models because  $h(t|x_i, \theta)/h(t|x_0, \theta)$  is not constant over  $t$ .

In fact, though, we sometimes doubt Assumption 3 (or Equation 24). For instance, prior political experience affects a legislator’s reelection hazard at an earlier stage, though it does not at a later stage. There are a few tests to check proportionality (for example, Park and Hendry, 2015). If tests imply non-proportionality, we should replace Equation 24 by

$$h(t | x_i, \theta) = h_0(t)\psi(x_i, t, \theta), \tag{26}$$

and specify  $\psi(x_i, t, \theta)$ . The most popular way is usage of interaction term:  $\psi(x_i, t, \theta) = \exp(x_i \beta + t x_i \gamma)$ .

**Frailty**

Scholars may suspect that  $\mu_i = \mu(x_i, \beta)$  is misspecified and there remains an omitted random variable or ‘frailty’  $W_i$  and, instead, assume that  $\mu_i = \mu(x_i, W_i, \beta)$ .<sup>3</sup> Suppose that we can specify  $h_i(t) = h(t|\psi_i) = h(t|x_i, W_i, \beta, \sigma)$  where  $\psi_i = (\mu_i, \sigma)$ . If  $W_i$  is unobserved, we only have to assume that  $W_i$  follows a certain distribution  $m(w|v)$  in order to integrate out  $w_i$ , obtain the marginal distribution,

$$\begin{aligned} p(\mathcal{T} | x_i, \theta) &= \int \exp\{-H(\mathcal{T} | x_i, w, \beta, \sigma)\} \\ &h(t | x_i, w, \beta, \sigma) m(w | v) dw, \end{aligned} \tag{27}$$

according to Equation 9, and estimate  $\theta = (\beta, \sigma, v)$ . In particular, if we can factor

$$h(t | x_i, W, \beta, \sigma) = W^* h(t | x_i, w_0, \beta, \sigma), \tag{28}$$

where  $W^*$  is a function of  $W$ , it follows that  $H(\mathcal{T}|x_i, W, \beta, \sigma) = W^* H(\mathcal{T}|x_i, w_0, \beta, \sigma)$ . This is called multiplicative frailty. For instance, in the case of the proportional hazard model (Equations 24 and 25), where we replace  $x_i$  by  $x'_i = (x_i, w_i)$  and Equation 13 by  $\mu_i = x_i \beta + W_i$  (where  $W_i$  might be called random effect), if  $W_i^* = \exp(-W_i / \sigma)$  and  $w_0 = 0$ , Equation 28 holds where  $h(t | x_i, w_0, \beta, \sigma) = h_0(t | \sigma) \exp(x_i \beta^*)$ .

Furthermore, if we assume  $W_i^*$  follows the gamma distribution with mean 1 and variance  $v$ ,  $m(w|v) = \Gamma(w^* | 1, v)$ , it follows that

$$S_{\text{gamma}}(t | x_i, \theta) = \{1 + v H(\mathcal{T} | x_i, w_0, \beta, \sigma)\}^{-\frac{1}{v}}. \tag{29}$$

If  $W_i^*$  follows the inverse-Gaussian distribution with mean 1 and variance  $v$ , it follows that

$$\begin{aligned} S_{\text{inv.Gauss}}(t | x_i, \theta) \\ = \exp\left(\frac{1}{v} [1 - \{1 + 2v H(\mathcal{T} | x_i, w_0, \beta, \sigma)\}^{\frac{1}{2}}]\right). \end{aligned} \tag{30}$$

If we give up using duration analysis jargon and apply general statistical terminology,

frailty models are a class of compound probability (or continuous mixture) distributions where  $W$  is the latent random variable or nuisance,  $p(\mathcal{T}|w, \cdot)$  is a conditional distribution,  $m(w|\cdot)$  is a mixing (or weight) distribution, and  $p(\mathcal{T}|\cdot)$  is an unconditional distribution which results by compounding  $p(\mathcal{T}|w, \cdot)$  with  $m(w|\cdot)$ . Thus, frailty models address heterogeneity and/or overdispersion, enabling robust inference.

Another look at frailty models is multi-level or hierarchical models:

$$T_i \sim f(t|x_i, W_i, \beta, \sigma),$$

$$W_i \sim m(w|\nu),$$

where we now replace a single frailty variable  $W_i$  by a vector of multiple frailties  $W_i$ . This representation makes it easy to construct more complicated models in a flexible manner. For instance, we may assume two components of  $W_i$  follow the gamma distribution with mean  $\nu_{\text{mean}}$  and variance  $\nu_{\text{variance}}$ :

$$W_i^{(1)} = W_{y(i)}^{(1)} \sim \Gamma(w^{(1)} | \nu_{\text{mean},y}^{(1)}, \nu_{\text{variance}}^{(1)}),$$

$$W_i^{(2)} \sim \Gamma(w^{(2)} | \nu_{\text{mean},i}^{(2)}, \nu_{\text{variance}}^{(2)}),$$
(31)

where  $y(i)$  indicate the group unit  $i$  belongs to. We call  $W_{y(i)}^{(1)}$  shared frailty. This is a kind of random effect (or coefficient) model. We may nest  $W_i^{(2)}$  in  $W_{y(i)}^{(1)}$  ('nested frailties'):  $\nu_{\text{mean},i}^{(2)} = W_{y(i)}^{(1)}$ . Or we may model  $W_i^{(2)}$  by covariates:  $\nu_{\text{mean},i}^{(2)} = \exp(x_i^{\text{frailty}} \beta^{\text{frailty}})$ . We can estimate parameters by using Markov Chain Monte Carlo (MCMC). Homola and Gill (Nd) elaborate on this topic.

### Discrete time framework

We may model

$$\Delta H_{ij} = \Delta H_{\cdot j}(\mathbf{x}_i, \theta)$$

$$= g^{-1}\{d(j|\tau) + \mathbf{x}_i \beta\},$$
(32)

where  $\theta \equiv (\tau, \beta)$  and  $d(j|\tau)$  is a time dependency function such as  $\bar{\nu}$ -th polynomial  $\left(\sum_{\nu=0}^{\bar{\nu}} \tau^{(\nu)} j^\nu\right)$ ,

and  $x_i$  does not contain the constant term. Since  $0 \leq \Delta H_{ij} \leq 1$ , the canonical link functions are logit  $g_{\text{logit}}(\pi) \equiv \Lambda^{-1}(\pi) = \log\{\pi/(1-\pi)\}$  and probit  $g_{\text{probit}}(\pi) \equiv \Phi^{-1}(\pi)$ . Nonetheless, if we employ the complementary log-log (cloglog) link function  $g_{\text{cloglog}}(\pi) \equiv \Xi^{-1}(\pi) = \log\{-\log(1-\pi)\}$  and specify  $d(j|\tau) = \log\{H_0(\mathcal{T}_j|\tau)\}$ , it is equivalent to assume that the corresponding hazard function in the continuous time framework is  $h_i(t|\theta) = h_0(t|\tau) \exp(x_i \beta)$  (it is a good exercise to prove the equivalence). For instance, if we employ  $d_{\text{Cox}}(j|\tau) \equiv \tau^{(j)} \equiv \log\{H_0(\mathcal{T}_j)\}$ , the model corresponds to the Cox model. However, choice among these link functions does not matter so much (Beck et al., 1998).

The likelihood is proportional to

$$l(\theta | E, \mathbf{x}) \propto \prod_{i \in \mathcal{I}} \Pr(E_i | x_i, \theta)$$

(∵ Assumption 2)

$$= \prod_{i \in \mathcal{I}} \left[ \underbrace{\prod_{j \in \mathcal{J}_i} \{1 - \Delta H_{\cdot j}(\mathbf{x}_i, \theta)\}}_{\text{duration}} \right] \underbrace{\frac{\Delta H_{\cdot j_i}(\mathbf{x}_i, \theta)}{1 - \Delta H_{\cdot j_i}(\mathbf{x}_i, \theta)}}_{\text{event}}$$

(∵ Equation 5),

where  $E \equiv (E_1, E_2, \dots, E_{\bar{T}})$ . This representation reminds us of time series cross section data with a binary dependent variable (BTSCS), where time index is  $j$ , unit of observation is  $ij$ , and the binary dependent variable is  $E_{ij}$  (Beck et al., 1998). That is, analysis of BTSCS data can be interpreted as a duration model in the discrete time framework. As the above argument shows, how to model time dependency function  $d(j|\tau)$  determines how the discrete hazard rate changes over time  $t_j$ 's. If analysts of BTSCS data do not include  $j$  as covariates ( $d_{\text{exponential}}(j|\tau) \equiv \tau^{(0)}$ ), they implicitly assume that  $T_i$  follows exponential distribution – that is, the hazard rate is constant over time  $t$  (Equation 18), which may be inappropriate.

### PARALLEL DURATIONS

Suppose that there are  $\bar{r}$  types of events (and the corresponding durations) or ‘competing risks’  $r \in \mathcal{R} \equiv \{0, 1, 2, \dots, \bar{r} - 1\}$  for every unit and, once an event happens, we cannot observe the other later events. For instance, a cabinet will end with an event among resignation ( $r = 1$ ), vote of non-confidence ( $r = 2$ ), or electoral defeat ( $r = 3$ ). Denote the latent continuous time of event due to risk  $r$  by  $T_i^{(r)}$ . By abusing notation, redefine  $T_i$  as the earliest event time ( $T_i \equiv \min_r T_i^{(r)}$ ) and let  $R_i \in \mathcal{R}$  denote the risk which incurs the earliest event, supposing no tied earliest events ( $\{r \mid r \in \mathcal{R}, T_i^{(r)} = T_i\} = \{R_i\}$  is a singleton set). We observe  $T_i$  and  $R_i$  but not  $T_i^{(r)}$  for  $r \neq R_i$ .

If  $R_i = r$ , duration  $T_i^{(r)} \equiv (0, T_i^{(r)}]$  is observed, and event  $r$  is observed at time  $T_i = T_i^{(r)}$ . If  $R_i \neq r$ ,  $T_i^{(r)}$  is said to be ‘censored’ at time  $T_i = T_i^{(R_i)}$  (or  $T_i^{(r)} > T_i$  is missing in terms of ordinary statistical vocabulary) and we do not observe event  $r$ . Sometimes we say  $T_i^{(r)}$  is ‘right’ censored because we usually locate  $T_i^{(r)}$  to the right of  $T_i$  on the horizontal axis  $t$ . We know that event  $r$  will happen after  $T_i$  (i.e.,  $T_i^{(r)} > T_i$ ), though we do not know exactly when it will happen ( $T_i^{(r)}$ ).

Denote the set of risks in which researchers have no interest by  $\mathcal{R}^0 \subsetneq \mathcal{R}$ . In the example of cabinet termination, the literature does not care about the case where a cabinet technically ends due to the constitutional term of the parliament ( $r = 0 \in \mathcal{R}^0$ ). In the case of  $R_i = r \in \mathcal{R}^0$ , even though duration  $T_i^{(r)}$  and event  $r$  are observed, we call (not just duration  $T_i^{(r)}$  for  $r' \neq r$  but also) unit  $i$  censored at time  $T_i = T_i^{(r)}$ . Let  $\mathcal{R}^+ \equiv \mathcal{R} \setminus \mathcal{R}^0$  (note  $\mathcal{R}^+ \neq \emptyset$ ). In simple duration models, we have substantive interest in a particular event  $r \in \mathcal{R}$  alone ( $\mathcal{R}^+ = \{r\}$ ). In typical competing risks models, however, we define  $\mathcal{R}^0 = \{0\}$ , where I call  $r = 0$  the ‘censoring’ event. In a special case, the censoring event time is constant:  $\Pr(T_i^{(0)} = t_i^{(0)}) = 1$ . The usual situation is that a researcher stops to observe duration at predetermined time  $t_i^{(0)} = t^{(0)}$ , where

the censoring event refers to the end of observation. I emphasize that *censoring is another event* in which scholars have no interest.

### Independent Durations

We make two assumptions.

**Assumption 5** (Stochastic Independence of Parallel Durations):

$$T_i^{(r)} \perp\!\!\!\perp T_i^{(r')} \text{ if } r \neq r'.$$

This is usually called independence of competing risks. In particular, if, for all  $r \in \mathcal{R}^+$  and all  $r' \in \mathcal{R}^0$ , Assumption 5 holds, we call the situation non-informative censoring (or, more generally, missing at random).

**Assumption 6** (Parametric Independence of Parallel Durations):

$$f_i^{(r)}(t) = p(T_i^{(r)} = t \mid x_i, \theta^{(r)}),$$

where  $\theta^{(r)}$  and  $\theta^{(r')}$  share no parameter in the case of  $r \neq r'$ .

Namely, we parametrize the distribution of  $T_i^{(r)}$  across  $r$ 's separately. We redefine  $\theta \equiv (\theta^{(0)}, \theta^{(1)}, \dots, \theta^{(\bar{r}-1)})$ .<sup>4</sup>

### Discrete time framework

Denote the latent discrete time of event due to risk  $r$  by  $J_i^{(r)}$  (where  $T_i^{(r)} \in \mathcal{T}_{J_i^{(r)}}$ ). By abusing notation, redefine  $J_i$  as the earliest event time ( $J_i \equiv \min_r J_i^{(r)}$ ), supposing no tied earliest events ( $J_i^{(r)} > J_i$  for  $r \neq R_i$ ). We observe  $J_i$  but not  $J_i^{(r)}$  for  $r \neq R_i$ . For  $j \in \mathcal{J}_r$ , define the event  $r$  indicator as

$$E_{ij}^{(r)} \equiv \begin{cases} 0 & \text{if } j < J_i^{(r)} & \left( \begin{array}{l} \text{event } r \text{ will} \\ \text{occur after } \mathcal{T}_{j} \end{array} \right), \\ 1 & \text{if } j = J_i^{(r)} & \left( \begin{array}{l} \text{event } r \text{ does} \\ \text{occur during } \mathcal{T}_{j} \end{array} \right), \end{cases}$$

and the vector of event  $r$  indicators as  $E_i^{(r)} \equiv (E_{i1}^{(r)}, E_{i2}^{(r)}, \dots, E_{iJ_i}^{(r)})$ . It follows that:

$$\begin{aligned} \Pr(E_i, R_i) &= \prod_{r \in \mathcal{R}} \Pr(E_i^{(r)}) \quad (\text{Assumption 5}) \\ &= \left\{ \prod_{r \in \mathcal{R}} \prod_{j \in \mathcal{J}_i} (1 - \Delta H_{ij}^{(r)}) \right\} \left( \frac{\Delta H_{i i_i}^{(R_i)}}{1 - \Delta H_{i i_i}^{(R_i)}} \right) \quad (33) \\ & (\because \text{Equation 5, } E_{ij}^{(r)} = 0 \text{ except } E_{i i_i}^{(R_i)} = 1). \end{aligned}$$

Therefore, the likelihood is:

$$\begin{aligned} l(\theta | E, R, x) &\propto \prod_{i \in \mathcal{I}} \Pr(E_i, R_i | x_i, \theta) \quad (\because \text{Assumption 2}) \\ &= \prod_{i \in \mathcal{I}} \left[ \underbrace{\prod_{r \in \mathcal{R}} \prod_{j \in \mathcal{J}_i} \{1 - \Delta H_{ij}^{(r)}(x_i, \theta^{(r)})\}}_{\text{duration}} \right] \quad (34) \\ &\quad \frac{\Delta H_{i i_i}^{(R_i)}(x_i, \theta^{(R_i)})}{\underbrace{1 - \Delta H_{i i_i}^{(R_i)}(x_i, \theta^{(R_i)})}_{\text{event}}} \\ & (\because \text{Equations 32 and 33}), \end{aligned}$$

where  $R \equiv (R_1, R_2, \dots, R_T)$ . Equation 34 indicates the probability that *no* event happens during  $(0, t_{J_i-1}]$  and *only* event  $R_i$  happens during  $(t_{J_i-1}, t_{J_i}]$ . Furthermore,

$$\begin{aligned} l(\theta^{(r)} | E, R, x) &\propto \prod_{i \in \mathcal{I}} \left[ \prod_{j \in \mathcal{J}_i} \{1 - \Delta H_{ij}^{(r)}(x_i, \theta^{(r)})\} \right] \quad (35) \\ &\quad \left\{ \frac{\Delta H_{i i_i}^{(R_i)}(x_i, \theta^{(R_i)})}{1 - \Delta H_{i i_i}^{(R_i)}(x_i, \theta^{(R_i)})} \right\}^{I(R_i=r)}, \end{aligned}$$

where  $I(\cdot)$  is an indicator function of whether the argument holds. Thus, thanks to Assumption 6, we only have to maximize Equation 35 so as to estimate  $\theta^{(r)}$  where  $r \in \mathcal{R}^+$ .

**Continuous time framework**

In the same spirit of Equation 9, we define and derive

$$\begin{aligned} p(\mathcal{T}_i, R_i) &\equiv \lim_{\Delta t \downarrow 0} \frac{1}{\Delta t_{J_i}} \Pr(E_i, R_i) \\ &= \lim_{\Delta t \downarrow 0} \left[ \prod_{r \in \mathcal{R}} \exp \left\{ \sum_{j \in \mathcal{J}_i} \frac{\log(1 - \Delta H_{ij}^{(r)})}{\Delta t_{j_i}} \Delta t_{j_i} \right\} \right] \quad (36) \\ &\quad \frac{\Delta H_{i i_i}^{(R_i)}}{\Delta t_{j_i} (1 - \Delta H_{i i_i}^{(R_i)})} \quad (\because \text{Equation 33}) \\ &= \left[ \prod_{r \in \mathcal{R}} \exp \{-H_i^{(r)}(\mathcal{T}_i)\} \right] h_i^{(R_i)}(\mathcal{T}_i). \end{aligned}$$

Therefore, the likelihood is:

$$\begin{aligned} l(\theta | \mathcal{T}, R, x) &\propto \prod_{i \in \mathcal{I}} p(\mathcal{T}_i, R_i | x_i, \theta) \\ & (\because \text{Assumption 2}) \\ &= \prod_{i \in \mathcal{I}} \left[ \underbrace{\prod_{r \in \mathcal{R}} \exp \{-H_i^{(r)}(\mathcal{T}_i | x_i, \theta^{(r)})\}}_{\text{duration}} \right] \quad (37) \\ &\quad \underbrace{\{h_i^{(R_i)}(\mathcal{T}_i | x_i, \theta^{(R_i)})\}}_{\text{event}} \\ & (\because \text{Equation 36}). \end{aligned}$$

Equation 37 implies the probability density that *no* event happens during  $(0, \mathcal{T}_i)$  and *only* event  $R_i$  happens at  $\mathcal{T}_i$ . Instead, by applying Equations 7 and 15 to Equation 37, we obtain its standard representation:

$$\begin{aligned} l(\theta | \mathcal{T}, R, x) &\propto \prod_{i \in \mathcal{I}} \prod_{r \in \mathcal{R}} \{S^{(r)}(\mathcal{T}_i | x_i, \theta^{(r)})\}^{I(R_i \neq r)} \\ &\quad \{f^{(r)}(\mathcal{T}_i | x_i, \theta^{(r)})\}^{I(R_i=r)}, \quad (38) \end{aligned}$$

which means the probability density that event  $r \neq R_i$  will happen after  $\mathcal{T}_i$  and event  $R_i$  happens at  $\mathcal{T}_i$ .

Furthermore,

$$\begin{aligned} l(\theta^{(r)} | \mathcal{T}, R, x) &\propto \prod_{i \in \mathcal{I}} \exp \{-H_i^{(r)}(\mathcal{T}_i | x_i, \theta^{(r)})\} \\ &\quad \{h_i^{(R_i)}(\mathcal{T}_i | x_i, \theta^{(R_i)})\}^{I(R_i=r)}. \quad (39) \end{aligned}$$

Accordingly, in practice, ready-made statistical software users only have to conduct

duration analysis for each risk  $r \in \mathcal{R}^+$  separately where, in the case of  $R_i \neq r$ , they regard such unit  $i$ 's as censored. This is the most famous advantage of duration analysis. Conventional models would call  $S^{(r)}(t|\cdot) = \int_t^\infty f^{(r)}(t'|\cdot) dt'$  truncated distribution (and call this model Torbit if  $T_i^{(r)}$  follows log normal distribution in the accelerated failure time model), though it is more straight-forward to model the likelihood in terms of hazard.

**Split population**

Some units may never be at risk of interest  $r \in \mathcal{R}^+$  and, thus, duration  $T_i^{(r)}$  does not exist and  $R_i \neq r$ , though observers do not know whether unit  $i$  is such a unit in the case of  $R_i \neq r$ . For instance, if a prime minister has no willingness to dissolve the parliament, all legislators serve till the end of their term – but not vice versa.

Here I introduce an extended split population model. Suppose that there are  $\bar{q}$  (choice) sets of risks at which a unit is,  $\mathcal{R}^{(q)}$ 's, such that  $q \in \mathcal{Q} \equiv \{1, 2, \dots, \bar{q}\}, \mathcal{R}^{(q)} \subseteq \mathcal{R}, \mathcal{R}^{(q)} \neq \emptyset$ , for  $q \neq q', \mathcal{R}^{(q)} \neq \mathcal{R}^{(q')}$ ,<sup>5</sup> let  $\mathcal{R}_i$  denote the set of risks at which unit  $i$  is:  $\mathcal{R}_i \in \{\mathcal{R}^{(q)} \mid q \in \mathcal{Q}\}$ , and denote  $\rho_i^{(q)} \equiv \Pr(\mathcal{R}_i = \mathcal{R}^{(q)})$ . Note that, if  $R_i \notin \mathcal{R}^{(q)}, \rho_i^{(q)} = 0$ . Then, the likelihood is

$$l(\theta \mid \mathcal{T}, R, x) \propto \prod_{i \in \mathcal{I}} p(\mathcal{T}_i, R_i \mid x_i, \theta)$$

$$= \prod_{i \in \mathcal{I}} \left[ \sum_{q \in \mathcal{Q}} \rho_i^{(q)} \underbrace{\prod_{r \in \mathcal{R}^{(q)}} \exp\{-H^{(r)}(\mathcal{T}_i | x_i, \theta^{(r)})\}}_{\text{duration}} \right]$$

$$\underbrace{h^{(R_i)}(\mathcal{T}_i | x_i, \theta^{(R_i)})}_{\text{event}}$$

Since this equation involves summation, in general, we cannot factor it like Equation 37 and estimate each  $\theta^{(r)}$  separately like Equation 39. Nonetheless, if  $r \in \bigcap_{q \in \mathcal{Q}} \mathcal{R}^{(q)}$ , we can factor  $l(\theta \mid \mathcal{T}, R, x) = l(\theta^{(r)} \mid \mathcal{T}, R, x) l(\theta^{(-r)} \mid \mathcal{T}, R, x)$  and estimate  $\theta^{(r)}$  alone, where  $\theta^{(-r)} = (\theta^{(0)}, \theta^{(1)}, \dots, \theta^{(r-1)}, \theta^{(r+1)}, \dots, \theta^{(\bar{r}-1)}, \rho)$ ,  $\rho = (\rho^{(1)}, \rho^{(2)}, \dots,$

$\rho^{(\bar{q})}$ ) and  $\rho^{(q)} = (\rho_1^{(q)}, \rho_2^{(q)}, \dots, \rho_i^{(q)})$ . We may model  $\rho_i^{(q)} = g^{-1}(x_i \beta^{(q)})$  if we have substantive interest in  $\rho_i^{(q)}$  and/or improve estimation (for example, Fukumoto and Masuyama, 2015). To my knowledge, the only setup in practical use is  $\bar{q} = 2, r = 0$  is the ‘censoring’ event,  $\mathcal{R} = \mathcal{R}^{(1)} = \{0, 1\}$  and  $\mathcal{R}^{(2)} = \{0\}$ . Since  $0 \in \bigcap_{q \in \mathcal{Q}} \mathcal{R}^{(q)}$ , we can factor  $l(\theta \mid \mathcal{T}, R, x) = l(\theta^{(0)} \mid \mathcal{T}, R, x) l(\theta^{(1)}, \rho \mid \mathcal{T}, R, x)$  and, thus, estimate either  $\theta^{(0)}$  or  $(\theta^{(1)}, \rho)$  alone.

**Dependent Parallel Durations**

In fact, however, violation of Assumptions 5 and 6 is not uncommon. How to deal with such a situation (e.g., dependent competing risks and informative censoring) is an important research agenda. Below, I sketch the frontier and promising directions. For ease of presentation, I suppose that, for all  $i, \mathcal{R}_i = \mathcal{R}$  unless otherwise noted.

**Stochastic dependence**

Sometimes we have to doubt Assumption 5. In the running example of cabinet duration, if cabinet  $i$  expects a vote of non-confidence ( $r = 2$ ) is likely, it tends to resign ( $r = 1$ ) pre-emptively:  $F(T_i^{(1)} = t \mid T_i^{(2)} = t_{\text{early}}) > F(T_i^{(1)} = t \mid T_i^{(2)} = t_{\text{late}})$  if  $t_{\text{early}} < t_{\text{late}}$ . In this case, we should give up Assumption 5 and instead model the joint survivor function of  $T_i \equiv (T_i^{(0)}, T_i^{(1)}, \dots, T_i^{(\bar{r}-1)})$  which is denoted by  $S_i^{\bar{r}}(t) \equiv \Pr(\forall r \in \mathcal{R}, T_i^{(r)} \geq t)$ . Note that we cannot factor  $p(\mathcal{T}_i, R_i)$  into terms for each risk as in Equation 36 because they are not independent any more. Instead, we will utilize

$$p(\mathcal{T}_i, R_i) = - \frac{\partial S_i^{\bar{r}}(t)}{\partial t^{(R_i)}} \Bigg|_{t=T_i} \tag{40}$$

$$\equiv - \frac{\partial}{\partial t^{(R_i)}} \Pr(\forall r \in \mathcal{R}, T_i^{(r)} \geq t^{(r)}) \Bigg|_{\forall r \in \mathcal{R}, t^{(r)}=T_i}$$

We introduce three approaches to model  $S_i^{\bar{r}}(t \mid x_i, \theta)$  below.

**Frailty**

The first approach is frailty models. We assume stochastic independence of parallel durations conditional on a vector of frailties,  $W_i$ , which follows  $m(w|v)$ :

$$T_i^{(r)} \perp\!\!\!\perp T_i^{(r')} \mid w_i \text{ if } r \neq r'. \quad (41)$$

The likelihood is

$$\begin{aligned} l(\theta \mid \mathcal{T}, R, x) &\propto \prod_{i \in \mathcal{I}} p(T_i, R_i \mid x_i, \theta) \quad (\because \text{Assumption 2}) \\ &\propto \prod_{i \in \mathcal{I}} \int p(T_i, R_i \mid x_i, w, \theta) m(w \mid v) dw \\ &\propto \prod_{i \in \mathcal{I}} \int \underbrace{\left\{ \prod_{r \in \mathcal{R}} S^{(r)}(T_i \mid x_i, w, \theta^{(r)}) \right\}}_{\text{duration}} \underbrace{h^{(R_i)}(T_i \mid x_i, w, \theta^{(R_i)})}_{\text{event}} m(w \mid v) dw \\ &(\because \text{Equations 38 and 41}), \end{aligned} \quad (42)$$

where  $v$  now controls dependency among parallel durations and  $\theta = (\theta^{(0)}, \dots, \theta^{(\bar{r}-1)}, v)$ .

A ‘generalized dependent risks’ model (Gordon, 2002) assumes that  $W_i = (W_i^{(0)}, W_i^{(1)}, \dots, W_i^{(\bar{r}-1)})$  follows a multivariate normal distribution and a multiplicative frailty model (Equation 28) holds for every risk:  $h^{(r)}(t \mid x_i, W_i, \theta^{(r)}) = W_i^{(r)*} h^{(r)}(t \mid x_i, w_0^{(r)}, \theta^{(r)})$ .<sup>6</sup> A drawback of this model is that it has no closed form and, thus, should be calculated by multi-dimensional numerical integration or MCMC. It is computationally intensive and, in particular for not a small number of  $\bar{r}$ , challenging.

A simpler frailty model is that hazard rates of all risks share the same frailty  $W_i^{(r)} = W_i$ . It leads to positive dependency among  $T_i^{(r)}$ ’s across  $r$ ’s within unit  $i$ . Equation 42 leads to

$$\begin{aligned} \prod_{i \in \mathcal{I}} \int \underbrace{\left\{ \prod_{r \in \mathcal{R}} \{S^{(r)}(T_i \mid x_i, w_0, \theta^{(r)})\}^{w^*} \right\}}_{\text{duration}} \underbrace{w^* h^{(R_i)}(T_i \mid x_i, w_0, \theta^{(R_i)})}_{\text{event}} m(w \mid v) dw, \end{aligned} \quad (43)$$

where we only have to integrate out a single frailty  $W_i$ . For instance, if we assume  $m(w|v) = \Gamma(w^*|L, v)$ , according to Equations 29 and 40, Equation 43 can be expressed in closed form:

$$\prod_{i \in \mathcal{I}} \underbrace{\left\{ 1 + v \sum_{r \in \mathcal{R}} H^{(r)}(T_i \mid x_i, w_0, \theta^{(r)}) \right\}^{\frac{1}{v}-1}}_{\text{duration}} \underbrace{h^{(R_i)}(T_i \mid x_i, w_0, \theta^{(R_i)})}_{\text{event}}.$$

**Seemingly unrelated regression**

The second approach is a seemingly unrelated regression which assumes that  $S^{\bar{r}}(t \mid x_i, \theta)$  is equal to a well-defined multivariate distribution function. In the case of  $\bar{r} = 2$ , besides the bivariate normal, one of such distributions is a bivariate Weibull:

$$\begin{aligned} S_{\text{Weibull}}^2(t \mid x_i, \theta) &= \left\{ \prod_{r=0}^1 S_{\text{Weibull}}(t \mid x_i, \theta^{(r)}) \right\} \\ &\left\{ 1 + v \prod_{r=0}^1 F_{\text{Weibull}}(t \mid x_i, \theta^{(r)}) \right\} \end{aligned} \quad (44)$$

(for the univariate Weibull, recall Equation 19). According to Equations 40 and 44, the likelihood is

$$\begin{aligned} l(\theta \mid \mathcal{T}, R, x) &\propto \prod_{i \in \mathcal{I}} \underbrace{\left\{ \prod_{r=0}^1 S_{\text{Weibull}}(T_i \mid x_i, \theta^{(r)}) \right\}}_{\text{duration}} \\ &\underbrace{1 + v F_{\text{Weibull}}(T_i \mid x_i, \theta^{(1-R_i)})}_{\text{duration}} \\ &\underbrace{\left\{ 1 - 2 S_{\text{Weibull}}(T_i \mid x_i, \theta^{(R_i)}) \right\}}_{\text{duration}} \\ &\times \underbrace{h_{\text{Weibull}}(T_i \mid x_i, \theta^{(R_i)})}_{\text{event}}, \end{aligned}$$

which can be expressed in closed form and is easy to calculate.

**Copula**

The third approach, copula modeling, provides a more easy-to-compute, general, and

flexible platform.<sup>7</sup> According to Sklar’s Theorem,  $F^{\bar{r}}(t | x_i, \theta) \equiv \Pr(\forall r \in \mathcal{R}, T_i^{(r)} \leq t)$  and  $S^{\bar{r}}(t | x_i, \theta)$  can be uniquely expressed by a copula  $C(u)$  and a survivor copula  $\tilde{C}(\tilde{u})$ , functions of  $u^{(r)}$ ’s and  $\tilde{u}^{(r)}$ ’s, respectively:

$$\begin{aligned}
 F^{\bar{r}}(t | x_i, \theta) &= C(u), \\
 u &\equiv (u^{(0)}, \dots, u^{(\bar{r}-1)}), \\
 u^{(r)} &\equiv F^{(r)}(t | x_i, \theta^{(r)}), \\
 S^{\bar{r}}(t | x_i, \theta) &= \tilde{C}(\tilde{u}), \\
 \tilde{u} &\equiv (\tilde{u}^{(0)}, \dots, \tilde{u}^{(\bar{r}-1)}), \\
 \tilde{u}^{(r)} &\equiv S^{(r)}(t | x_i, \theta^{(r)}) = 1 - u^{(r)}.
 \end{aligned}$$

Below, we only consider cases of  $\bar{r} = 2$ . It follows  $\tilde{C}(\tilde{u}) = 1 + C(u) - \sum_{r=0}^1 u^{(r)}$ .

There are numerous copula functions developed. Here, we only name a few. An example is the Farlie–Gumbel–Morgenstern (FGM) copula:

$$C_{\text{FGM}}(u | v) \equiv \left\{ \prod_{r=0}^1 u^{(r)} \right\} \left\{ 1 + v \prod_{r=0}^1 (1 - u^{(r)}) \right\}.$$

The most useful class is Archimedean copulas, where there is a generator function  $a(u)$  such that

$$C_{\text{Archi}}\{u | a(\cdot)\} \equiv a^{-1} \left\{ \sum_{r=0}^1 a(u^{(r)}) \right\}.$$

For instance, the generator function of Gumbel copula is  $a_{\text{Gumbel}}(u | v) \equiv \{-\log(u)\}^v$ .

The first merit of copula is ease to compute. Equation 40 turns into

$$\begin{aligned}
 & - \frac{\partial S_i^{\bar{r}}(t)}{\partial t^{(R_i)}} \Big|_{t=T_i} \\
 &= 1 - \underbrace{\frac{\partial \tilde{C}(\tilde{u}_i)}{\partial \tilde{u}^{(R_i)}} \Big|_{\forall r \in \mathcal{R}, \tilde{u}^{(r)} = S_i^{(r)}(T_i)}}_{\text{duration}} S_i^{(R_i)}(T_i) \\
 & \quad \underbrace{h_i^{(R_i)}(T_i)}_{\text{event}}.
 \end{aligned}$$

Seemingly, this is weird, though the first derivative of many copulas has the analytical

solution and, thus, is easy to calculate. For instance,

$$\begin{aligned}
 & \frac{\partial C_{\text{FGM}}(u | v)}{\partial u^{(r)}} \\
 &= u^{(1-r)} \left\{ 1 + v(1 - u^{(1-r)})(1 - 2u^{(r)}) \right\}, \\
 & \frac{\partial C_{\text{Archi}}\{u | a(\cdot)\}}{\partial u^{(r)}} \\
 &= \frac{\partial a(u')}{\partial u'} \Big|_{u'=u^{(r)}} / \frac{\partial a(u'')}{\partial u''} \Big|_{u''=C_{\text{Archi}}\{u | a(\cdot)\}}.
 \end{aligned}$$

The second advantage is that some copulas are general enough to incorporate other approaches. For example, it turns out that one type of bivariate Weibull (Equation 44) is equivalent to FGM copula where the marginal distribution is the univariate Weibull (Equation 19):

$$\begin{aligned}
 & S_{\text{Weibull}}^2(t | x_i, \theta) \\
 &= C_{\text{FGM}} \left\{ \begin{aligned} & S_{\text{Weibull}}(t | x_i, \theta^{(0)}), \\ & S_{\text{Weibull}}(t | x_i, \theta^{(1)}) | v \end{aligned} \right\}. \tag{45}
 \end{aligned}$$

The third strength of copula is flexibility or modularity: any survivor copula  $\tilde{C}(\tilde{u})$  can connect any marginal distributions  $S^{(r)}(t)$ . For instance, scholars may replace FGM copula by Gumbel copula in Equation 45:

$$\begin{aligned}
 & S_{\text{Weibull Alt}}^2(t | x_i, \theta) \\
 &= C_{\text{Archi}} \left\{ \begin{aligned} & S_{\text{Weibull}}(t | x_i, \theta^{(0)}), \\ & S_{\text{Weibull}}(t | x_i, \theta^{(1)}) | a_{\text{Gumbel}}(\cdot | v) \end{aligned} \right\},
 \end{aligned}$$

which is another, often-mentioned type of bivariate Weibull. Or we can substitute log normal (Equation 11) with marginal Weibull in Equation 45:

$$\begin{aligned}
 & S^2(t | x_i, \theta) \\
 &= C_{\text{FGM}} \left\{ \begin{aligned} & S_{\text{Log Normal}}(t | x_i, \theta^{(0)}), \\ & S_{\text{Log Normal}}(t | x_i, \theta^{(1)}) | v \end{aligned} \right\}.
 \end{aligned}$$



**Parametric dependence**

Sometimes, Assumption 6 is violated. For instance, if lawmakers foresee electoral defeat ( $r = 1$ ), they will strategically retire ( $r = 2$ ). Then, scholars may suppose the following systematically dependent competing risks model (Fukumoto, 2009):

$$\begin{aligned} &h^{(1)}(t \mid x_i, \theta^{(1)}) \\ &= h_0^{(1)}(t \mid \sigma^{(1)}) \exp(x_i^{(1)} \beta^{*(1)}), \\ &h^{(2)}(t \mid x_i, \theta^{(2)}) \\ &= h_0^{(2)}(t \mid \sigma^{(2)}) \exp(x_i^{(2)} \beta^{*(2)} + \nu x_i^{(1)} \beta^{*(1)}). \end{aligned}$$

If  $\beta^{*(1)}$ ,  $\nu > 0$ , it follows that  $x_i^{(1)}$  (e.g., prior vote share margin) increases  $h^{(2)}(\cdot)$  as it increases  $h^{(1)}(\cdot)$ . Note that  $\theta^{(1)} = (\sigma^{(1)}, \beta^{*(1)})$  and  $\theta^{(2)} = (\sigma^{(2)}, \beta^{*(1)}, \beta^{*(2)}, \nu)$  share  $\beta^{*(1)}$ . Thus, we cannot estimate  $\theta^{(1)}$  by maximizing Equation 39 because  $\beta^{*(1)}$  is included in  $\theta^{(2)}$  as well. Instead, we have to estimate  $\theta$  by maximizing Equation 37 if Assumption 5 holds. For identification, exclusion restriction is necessary (i.e., at least one covariate in  $x_i^{(1)}$ , which is now a vector, is excluded from  $x_i^{(2)}$ ).

Another approach toward systematic dependence is the simultaneous equations model. For details, see Hays and Kachi (2015).

**SERIAL DURATIONS**

Finally, I consider multiple serial latent durations. I begin with laying out a general format, whose special cases are repeated events and multi-stage models.

**Left Truncation and Time Varying Covariates**

Suppose that each unit has multiple durations at each risk and denote the latent continuous event time of unit  $i$ 's  $k$ -th duration (or, simply, unit  $ik$ ) at risk  $r$  by  $T_{ik}^{(r)}$  where  $k$  is a positive integer and  $T_{ik}^{(r)}$  follows  $F_i^{(rk)}(t)$ . For example, the  $k$ -th duration may refer to the  $k$ -th term of

a legislator in a parliamentary system. The observable event time of unit  $ik$  is denoted by  $T_{ik} \equiv \min_{r \in \mathcal{R}} T_{ik}^{(r)}$ . Let  $T_{i0} \equiv 0$ . Denote the risk which incurs the earliest event at  $T_{ik}$  as  $R_{ik} \in \mathcal{R}$ , supposing no tied earliest events ( $\{r \mid r \in \mathcal{R}, T_{ik}^{(r)} = T_{ik}\} = \{R_{ik}\}$  is a singleton set). We also suppose that, only if  $T_{i,k-1} < T_{ik}$ , we can observe unit  $ik$ 's duration for the period  $\mathcal{T}_{ik} \equiv (T_{i,k-1}, T_{ik}]$ . Unless  $k = 1$ , we call unit  $ik$ 's duration 'left truncated' because we cannot observe the duration for the period  $(0, T_{i,k-1}]$ .

In order to focus on serial, not parallel, durations, we make Assumption 5 within the  $k$ -th durations.

**Assumption 7** (Serial Duration Version of Stochastic Independence of Parallel Durations):

$$T_{ik}^{(r)} \perp\!\!\!\perp T_{ik}^{(r')} \text{ if } r \neq r'.$$

Denote the set of such  $k$ 's for unit  $i$  that we actually observe  $\mathcal{T}_{ik}$  by  $\mathcal{K}_i \subseteq \{k \mid T_{i,k-1} < T_{ik}\}$ . For any  $k$ , denote the history of observed  $k$ 's up to  $k$  by  $\mathcal{K}_{i,<k} \equiv \{k' \mid k' \in \mathcal{K}_i, k' < k\}$ . For now, suppose that  $\mathcal{K}_i$  is a set of consecutive integers from  $\min \mathcal{K}_i \equiv \underline{k}_i$  to  $\max \mathcal{K}_i \equiv \bar{k}_i$ . If  $\underline{k}_i > 1$ , we call unit  $i$  left truncated. For instance, if researchers study lawmakers' careers only after the Second World War, they will overlook the first few terms of old lawmakers.

**Discrete time framework**

Denote the latent discrete event time of unit  $ik$  at risk  $r$  by  $J_{ik}^{(r)}$  (where  $T_{ik}^{(r)} \in \mathcal{T}_{J_{ik}^{(r)}}$ ). The observable event time of unit  $ik$  is denoted by  $J_{ik} \equiv \min_{r \in \mathcal{R}} J_{ik}^{(r)}$ . Let  $J_{i0} \equiv 0$ . We suppose no tied earliest events ( $J_{ik}^{(r)} > J_{ik}$  for  $r \neq R_{ik}$ ). We also suppose that, only if  $J_{i,k-1} < J_{ik}$ , we can observe unit  $ik$ 's duration for the period  $\mathcal{J}_{ik} \equiv \{J_{i,k-1} + 1, J_{i,k-1} + 2, \dots, J_{ik}\}$ . For  $j \in \mathcal{J}_{ik}$ , redefine the event  $r$  indicator as

$$E_{ij}^{(r)} \equiv \begin{cases} 0 & \text{if } j < J_{ik}^{(r)} \\ & \text{(event } r \text{ of the } k\text{-th duration} \\ & \text{will occur after } \mathcal{T}_{.j}), \\ 1 & \text{if } j = J_{ik}^{(r)}. \\ & \text{(event } r \text{ of the } k\text{-th duration} \\ & \text{does occur during } \mathcal{T}_{.j}), \end{cases}$$

and  $E_{ij} \equiv E_{ij}^{(R_{ik})}$ . Denote the vector of unit  $ik$ 's event  $r$  indicators by  $E_{ik}^{(r)} \equiv (E_{i,J_{i,k-1}+1}^{(r)}, E_{i,J_{i,k-1}+2}^{(r)}, \dots, E_{i,k}^{(r)})$  and define  $E_{ik} \equiv (E_{i,J_{i,k-1}+1}, E_{i,J_{i,k-1}+2}, \dots, E_{i,k})$ . Redefine  $E_i \equiv (E_{i,k_1}, E_{i,k_1+1}, \dots, E_{i,k_1}^-)$  and define  $R_i \equiv (R_{i,k_1}, R_{i,k_1+1}, \dots, R_{i,k_1}^-)$ . For any  $k$ , denote the history of observed  $E_{ik}$ 's and  $R_{ik}$ 's up to  $k$  by  $E_{i,<k} \equiv (j_{i,k_1-1}, E_{i,k_1}, E_{i,k_1+1}, \dots, E_{i,k_1}^-)$  and  $R_{i,<k} \equiv (R_{i,k_1}, R_{i,k_1+1}, \dots, R_{i,k_1}^-)$ , where  $k_{i,<k} = \max \mathcal{K}_{i,<k}$ . Note that  $E_{i,<k}$  includes the starting point,  $j_{i,k_1-1}$ .

We derive the probability of  $E_i$  and  $R_i$  like Equation 33, though we now condition it on the starting point,  $j_{i,k_1-1}$ :

$$\begin{aligned} \Pr(E_i, R_i | j_{i,k_1-1}) &= \prod_{k \in \mathcal{K}_i} \Pr(E_{ik}, R_{ik} | E_{i,<k}, R_{i,<k}) \\ (\because \mathcal{K}_i \text{ is a set of consecutive integers}) \\ &= \prod_{k \in \mathcal{K}_i} \prod_{r \in \mathcal{R}} \Pr(E_{ik}^{(r)} | E_{i,<k}, R_{i,<k}) \\ (\because \text{Assumption 7}) \end{aligned} \tag{46}$$

$$\begin{aligned} &= \prod_{k \in \mathcal{K}_i} \left\{ \prod_{r \in \mathcal{R}} \prod_{j \in (J_{ik} \setminus J_{i,k-1})} (1 - \Delta H_{ij}^{(rk)} | E_{i,<k}, R_{i,<k}) \right\} \\ &\quad \frac{(\Delta H_{i,j_{i,k_1}}^{(R_{ik}^k)} | E_{i,<k}, R_{i,<k})}{(1 - \Delta H_{i,j_{i,k_1}}^{(R_{ik}^k)} | E_{i,<k}, R_{i,<k})} \text{ (c.f. Equation 33).} \end{aligned}$$

When we parametrize  $\Delta H_{ij}^{(rk)}$ , we may also introduce time varying covariates  $x_{ik}$  in the sense that it can be  $x_{ik} \neq x_{ik'}$  if  $k \neq k'$ . Thus, we model  $\Delta H_{ij}^{(rk)} = \Delta H_{ij}^{(rk)}(x_{ik}, \theta^{(rk)})$  instead of Equation 32. Redefine  $x_i \equiv (x_{i,k_1}, x_{i,k_1+1}, \dots, x_{i,k_1}^-)$ .

Therefore, the likelihood is

$$\begin{aligned} l(\theta | E, R, x, j_{i,k_1-1}) &\propto \prod_{i \in \mathcal{I}} \Pr(E_i, R_i | x_i, \theta, j_{i,k_1-1}, R_{i,<k}) \\ (\because \text{Assumption 2}) \\ &= \prod_{i \in \mathcal{I}} \prod_{k \in \mathcal{K}_i} \underbrace{\left[ \prod_{r \in \mathcal{R}} \prod_{j \in (J_{ik} \setminus J_{i,k-1})} \{1 - \Delta H_{ij}^{(rk)}\} \right]}_{\text{duration}} \tag{47} \\ &\quad \underbrace{\left[ x_{ik}, \theta^{(rk)} | E_{i,<k}, R_{i,<k} \right]}_{\text{duration}} \\ &\quad \times \underbrace{\frac{\Delta H_{i,j_{i,k_1}}^{(R_{ik}^k)}(x_{ik}, \theta^{(R_{ik}^k)} | E_{i,<k}, R_{i,<k})}{1 - \Delta H_{i,j_{i,k_1}}^{(R_{ik}^k)}(x_{ik}, \theta^{(R_{ik}^k)} | E_{i,<k}, R_{i,<k})}}_{\text{event}} \\ (\because \text{Equation 46}), \end{aligned}$$

where  $j_{i,k-1} \equiv (j_{1,k_1-1}, j_{2,k_2-1}, \dots, j_{\bar{i},k_{\bar{i}}-1})$ ,  $\theta^{(r)} \equiv (\theta^{(r1)}, \theta^{(r2)}, \dots, \theta^{(r\bar{k})})$ , and  $\bar{k} \equiv \max_i \bar{k}_i$ . Equation 47 implies the probability that, for all  $k \in \mathcal{K}_i$ , no event happens during  $(t_{J_{i,k-1}}, t_{J_{ik}}]$  (rather than  $(0, t_{J_{ik}}]$ ) and only event  $R_{ik}$  happens during  $(t_{J_{ik-1}}, t_{J_{ik}}]$ . The difficulty of serial durations arises from the dependence of the  $k$ -th duration ( $E_{ik}$  and  $R_{ik}$ ) on its history ( $E_{i,<k}$  and  $R_{i,<k}$ ). It is also clear that Equation 47 deals with panel data.

### Continuous time framework

Denote  $\mathcal{T}_i \equiv (\mathcal{T}_{i,k_1}, \mathcal{T}_{i,k_1+1}, \dots, \mathcal{T}_{i,k_1}^-)$ . For any  $k$ , denote the history of observed  $\mathcal{T}_{ik}$ 's up to  $k$  by  $\mathcal{T}_{i,<k} \equiv (t_{i,k_1-1}, \mathcal{T}_{i,k_1}, \mathcal{T}_{i,k_1+1}, \dots, \mathcal{T}_{i,k_1}^-)$ . Note that  $\mathcal{T}_{i,<k}$  includes the starting point,  $t_{i,k_1-1}$ . Because  $\mathcal{K}_i$  is a set of consecutive integers, it follows

$$\begin{aligned} &p(\mathcal{T}_i, R_i | t_{i,k_1-1}) \\ &= \prod_{k \in \mathcal{K}_i} p(\mathcal{T}_{ik}, R_{ik} | \mathcal{T}_{i,<k}, R_{i,<k}). \end{aligned} \tag{48}$$

We define  $p(\mathcal{T}_{ik}, R_{ik})$  in the same spirit of Equation 36, though we now condition  $p(\mathcal{T}_{ik}, R_{ik})$  on their history,  $\mathcal{T}_{i,<k}$  and  $R_{i,<k}$ :

$$\begin{aligned} &p(\mathcal{T}_{ik}, R_{ik} | \mathcal{T}_{i,<k}, R_{i,<k}) \\ &\equiv \lim_{\Delta t \downarrow 0} \frac{1}{\Delta t_{J_{ik}}} \Pr(E_{ik}, R_{ik} | E_{i,<k}, R_{i,<k}) \\ &= \lim_{\Delta t \downarrow 0} \left[ \prod_{r \in \mathcal{R}} \exp \left\{ \sum_{j \in (J_{ik} \setminus J_{i,k-1})} \frac{\log(1 - \Delta H_{ij}^{(rk)} | E_{i,<k}, R_{i,<k})}{\Delta t_j} \Delta t_j \right\} \right] \\ &\quad \frac{(\Delta H_{i,j_{i,k_1}}^{(R_{ik}^k)} | E_{i,<k}, R_{i,<k})}{\Delta t_{J_{ik}} (1 - \Delta H_{i,j_{i,k_1}}^{(R_{ik}^k)} | E_{i,<k}, R_{i,<k})} \tag{49} \\ (\because \text{Equation 46}) \\ &= \left[ \prod_{r \in \mathcal{R}} \exp \{ -H_i^{(rk)}(\mathcal{T}_{ik} | \mathcal{T}_{i,<k}, R_{i,<k}) \} \right] \\ &\quad h_i^{(R_{ik}^k)}(\mathcal{T}_{ik} | \mathcal{T}_{i,<k}, R_{i,<k}). \end{aligned}$$

We model  $h_i^{(rk)}(t | \mathcal{T}_{i,<k}, R_{i,<k}) = h^{(rk)}(t | \mathcal{T}_{i,<k}, R_{i,<k}, x_{ik}, \theta^{(rk)})$ . Redefine  $\mathcal{T} \equiv (\mathcal{T}_1, \mathcal{T}_2, \dots, \mathcal{T}_{\bar{i}})$  and  $R \equiv (R_1, R_2, \dots, R_{\bar{i}})$ . Let  $t_{i,k-1} \equiv (t_{1,k_1-1}, t_{2,k_2-1}, \dots, t_{\bar{i},k_{\bar{i}}-1})$ . The likelihood is

$$\begin{aligned}
 l(\theta | \mathcal{T}, \mathbf{R}, \mathbf{x}, \mathbf{t}_{\underline{k}-1}) &\propto \prod_{i \in \mathcal{I}} p(\mathcal{T}_i, \mathbf{R}_i | \mathbf{x}_i, \theta, \mathbf{t}_{i, \underline{k}_i-1}) \\
 &(\because \text{Assumption 2}) \\
 &= \prod_{i \in \mathcal{I}} \prod_{k \in \mathcal{K}_i} p(\mathcal{T}_{ik}, R_{ik} | \mathcal{T}_{i, < k}, \mathbf{R}_{i, < k}, \mathbf{x}_{ik}, \theta) \\
 &(\because \text{Equation 48}) \\
 &= \prod_{i \in \mathcal{I}} \prod_{k \in \mathcal{K}_i} \left[ \prod_{r \in \mathcal{R}} \exp\{-H^{(rk)}\} \right] \\
 &\quad \underbrace{\hspace{10em}}_{\text{duration}} \\
 &\quad \left[ \underbrace{(\mathcal{T}_{ik} | \mathcal{T}_{i, < k}, \mathbf{R}_{i, < k}, \mathbf{x}_{ik}, \theta^{(rk)})}_{\text{duration}} \right] \\
 &\quad \times \underbrace{\{h^{(R_{ik})}(T_{ik} | \mathcal{T}_{i, < k}, \mathbf{R}_{i, < k}, \mathbf{x}_{ik}, \theta^{(R_{ik})})\}}_{\text{event}} \\
 &(\because \text{Equation 49}).
 \end{aligned}
 \tag{50}$$

Equation 50 implies the probability density that *no* event happens during  $(T_{i, k-1}, T_{ik})$  (rather than  $(0, T_{ik})$ ) and *only* event  $R_{ik}$  happens at  $T_{ik}$ . The difficulty of serial durations arises from the dependence of the  $k$ -th duration  $(\mathcal{T}_{ik}, R_{ik})$  on its history  $(\mathcal{T}_{i, < k}, \mathbf{R}_{i, < k})$ . Note that

$$\begin{aligned}
 &\exp\{-H_i^{(rk)}(\mathcal{T}_{ik} | \mathcal{T}_{i, < k}, \mathbf{R}_{i, < k})\} \\
 &= \exp\left\{-\int_{T_{i, k-1}}^{T_{ik}} h_i^{(rk)}(t | \mathcal{T}_{i, < k}, \mathbf{R}_{i, < k}) dt\right\} \\
 &= \frac{S_i^{(rk)}(T_{ik} | \mathcal{T}_{i, < k}, \mathbf{R}_{i, < k})}{S_i^{(rk)}(T_{i, k-1} | \mathcal{T}_{i, < k}, \mathbf{R}_{i, < k})}.
 \end{aligned}$$

If we follow the standard representation like Equation 38 and only pay attention to the fact that event  $r \neq R_{ik}$  will happen after  $T_{ik}$ , it would not be straight-forward to see why we divide  $S_i^{(rk)}(T_{ik} | \cdot)$  by  $S_i^{(rk)}(T_{i, k-1} | \cdot)$ . I also emphasize that we only have to consider the univariate distribution of one random variable,  $T_{ik}$ , conditioned on  $T_{i, k-1}$ , not the bivariate distribution of two random variables,  $T_{ik}$  and  $T_{i, k-1}$ . This view is another merit of my focus on duration rather than a single event time point.

**Interval truncation**

If  $\mathcal{K}_i$  is not a set of consecutive integers, we can partition it into  $\bar{K}$  sets of consecutive

integers:  $\mathcal{K}_i = \bigcup_{K=1}^{\bar{K}} \mathcal{K}_i^{(K)}$  where  $\mathcal{K}_i^{(K)}$  is the  $K$ -th set of consecutive integers from  $\min \mathcal{K}_i^{(K)} \equiv \underline{k}_i^{(K)}$  to  $\max \mathcal{K}_i^{(K)} \equiv \bar{k}_i^{(K)} < \underline{k}_i^{(K+1)} - 1$ . Unit  $i$  is not observed between  $T_{i, \underline{k}_i^{(K)}}$  and  $T_{i, \underline{k}_i^{(K+1)}-1}$  and is said to be interval truncated. For instance, we may observe unit  $i$  for  $k = 1, 2, 4, 5$  but not  $k = 3$ , and this unit is interval truncated between  $T_{i2}$  and  $T_{i3}$ , where  $\mathcal{K}_i^{(1)} = \{1, 2\}$  and  $\mathcal{K}_i^{(2)} = \{4, 5\}$ . Let us redefine  $\mathcal{T}_{i, < k} \equiv (t_{i, \underline{k}-1}, \mathcal{T}_{i, \underline{k}_i^{(1)}}, \mathcal{T}_{i, \underline{k}_i^{(1)}+1}, \dots, \mathcal{T}_{i, \bar{k}_i^{(1)}}, \mathcal{T}_{i, \underline{k}_i^{(2)}}, \mathcal{T}_{i, \underline{k}_i^{(2)}+1}, \dots, \mathcal{T}_{i, \bar{k}_i^{(2)}}, \dots, \mathcal{T}_{i, \underline{k}_i^{(K)}}, \mathcal{T}_{i, \bar{k}_i^{(K)}+1}, \mathcal{T}_{i, \bar{k}_i^{(K)}, \dots, \mathcal{T}_{i, \underline{k}_i^{(K+1)}}, \dots, t_{i, \bar{k}_i^{(K)}})$ , where  $\mathbf{t}_{i, \underline{k}-1} \equiv (t_{i, \underline{k}_i^{(1)}}, t_{i, \underline{k}_i^{(1)}+1}, \dots, t_{i, \bar{k}_i^{(K)}})$  and  $\bar{k}_{i, < k} \in \mathcal{K}_i^{(K)}$ . We replace Equation 48 by

$$\begin{aligned}
 &p(\mathcal{T}_i, \mathbf{R}_i | \mathbf{t}_{i, \underline{k}-1}) \\
 &= \prod_{K=1}^{\bar{K}} \prod_{k \in \mathcal{K}_i^{(K)}} p(\mathcal{T}_{ik}, R_{ik} | \mathcal{T}_{i, < k}, \mathbf{R}_{i, < k}) \\
 &= \prod_{k \in \mathcal{K}_i} p(\mathcal{T}_{ik}, R_{ik} | \mathcal{T}_{i, < k}, \mathbf{R}_{i, < k}).
 \end{aligned}
 \tag{51}$$

**Left or interval censoring**

In some cases, even though  $T_{i, k-1} \geq T_{ik}$  and we do not observe  $T_{ik}$ , we may observe  $T_{i, k-1}$  and know  $T_{i, k-1} \geq T_{ik}$ . For instance, suppose that the  $k - 1 = 1$ -st duration is the time to delivery of a drug to a patient, and the  $k = 2$ -nd duration is the time to the patient’s death. When a doctor gives a drug to a patient at  $T_{i1}$ , the doctor may find the patient has already died ( $T_{i1} \geq T_{i2}$ ) but not know exactly when the patient died ( $T_{i2}$ ). In this case, unit  $ik$ ’s duration is said to be left censored. In particular, if we observe nothing before  $T_{i, k-1}$ , unit  $i$  is said to be left censored. If we observe not  $\mathcal{T}_{i, k-1}$  but  $\mathcal{T}_{ik}$ , for some  $k' < k - 1$ , unit  $i$  is said to be interval censored.

When unit  $ik'$  is left censored, we replace Equation 49 by

$$\begin{aligned}
 &\Pr(T_{ik} \leq T_{i, k-1} | \mathcal{T}_{i, < k}, \mathbf{R}_{i, < k}) \\
 &= 1 - \Pr(T_{ik} > T_{i, k-1} | \mathcal{T}_{i, < k}, \mathbf{R}_{i, < k}) \\
 &= 1 - \Pr(\forall r \in \mathcal{R}, T_{ik}^{(r)} > T_{i, k-1} | \mathcal{T}_{i, < k}, \mathbf{R}_{i, < k}) \\
 &= 1 - \prod_{r \in \mathcal{R}} \Pr(T_{ik}^{(r)} > T_{i, k-1} | \mathcal{T}_{i, < k}, \mathbf{R}_{i, < k}) \\
 &(\because \text{Assumption 7}) \\
 &= 1 - \prod_{r \in \mathcal{R}} S_i^{(rk)}(T_{i, k-1} | \mathcal{T}_{i, < k}, \mathbf{R}_{i, < k}).
 \end{aligned}
 \tag{52}$$

### Independent Serial Durations

In this subsection, we assume stochastic independence of serial durations across duration  $k$ 's within unit  $i$ .

**Assumption 8** (Stochastic Independence of Serial Durations):

$$T_{ik}^{(r)} \perp\!\!\!\perp T_{ik'}^{(r')} \text{ if } k \neq k',$$

where it can be either  $r = r'$  or  $r \neq r'$ .

It is straight-forward to derive  $h_i^{(rk)}(t | \mathcal{T}_{i,<k}, R_{i,<k}) = h_i^{(rk)}(t)$ . Accordingly, we do not have to condition the likelihood on the starting values.

In a simple (and, thus, often used) specification of the continuous time framework, we are interested in a certain risk  $r \in \mathcal{R}^+$  alone. Usually, we make the following assumptions: once an event due to risk  $r$  occurs at the end of  $k$ -th duration, we do not observe the succeeding unit  $i$ 's durations (if  $R_{ik} = r$ , it follows  $\bar{k}_i = k$  and  $R_{ik'} \neq r$  for  $k' < k$ ); hazard rates vary over  $k$  only through  $x_{ik}$  ( $h_i^{(rk)}(t) = h^{(r)}(t | x_{ik}, \theta^{(r)})$ ); and Assumption 6. Equation 50 is reduced to

$$\begin{aligned} & l(\theta^{(r)} | \mathcal{T}, R, x) \\ & \propto \prod_{i \in \mathcal{I}} \prod_{k \in \mathcal{K}_i} \exp\left\{-H^{(r)}(\mathcal{T}_{ik} | x_{ik}, \theta^{(r)})\right\} h^{(R_{ik})} \\ & \quad \left(T_{ik} | x_{ik}, \theta^{(R_{ik})}\right\}^{I(R_{ik}=r)}. \end{aligned}$$

In a popular specification of the discrete time framework, we substitute  $k$  with  $j$ . We also suppose that, for all units, the ‘censoring’ event ( $r = 0$ ) time for the  $k$ -th duration is constantly equal to the  $k$ -th fixed cutoff point:  $T_{ik}^{(0)} = t_{.k}$ . We model  $\Delta H_{ij}^{(rk)} = \Delta H^{(r)}(x_{ik}, \theta^{(r)})$ . Equation 47 is reduced to

$$\begin{aligned} & l(\theta^{(-0)} | E, R, x) \\ & \propto \prod_{i \in \mathcal{I}} \prod_{k \in \mathcal{K}_i} \left[ \prod_{r \in \mathcal{R} \setminus \{0\}} \{1 - \Delta H^{(r)}(x_{ik}, \theta^{(r)})\} \right] \\ & \quad \left\{ \frac{\Delta H^{(R_{ik})}(x_{ik}, \theta^{(R_{ik})})}{1 - \Delta H^{(R_{ik})}(x_{ik}, \theta^{(R_{ik})})} \right\}^{I(R_{ik} \neq 0)}. \end{aligned}$$

Moreover, let us assume that the hazard rate is proportional to the risk invariant baseline,  $h_i^{(rk)}(t): h_i^{(rk)}(t) = h_i^{(k)}(t) \exp(x_{ik} \beta^{*(rk)})$ . We apply the Cox model to  $\bar{r}$  durations in every duration  $k$  for every unit  $i$ . In Equation 22, we replace  $E_i(t)$  by  $E_{ik}^{(r)}$ ,  $i'$  by  $r'$ , and  $\mathcal{I}(t)$  by  $\mathcal{R}$  to obtain

$$\begin{aligned} & \Pr\left(E_{ik}^{(r)} = 1 \mid \sum_{r' \in \mathcal{R}} E_{ik}^{(r')} = 1\right) \\ & = \exp(x_{ik} \beta^{*(rk)}) / \sum_{r' \in \mathcal{R}} \exp(x_{ik} \beta^{*(r'k)}). \end{aligned}$$

By using Equations 23 and 47, it follows

$$\begin{aligned} & l^{\text{partial}}(\beta^* | R, x) \\ & \propto \prod_{i \in \mathcal{I}} \prod_{k \in \mathcal{K}_i} \Pr\left(E_{ik}^{(R_{ik})} = 1 \mid \sum_{r \in \mathcal{R}} E_{ik}^{(r)} = 1\right) \\ & = \prod_{i \in \mathcal{I}} \prod_{k \in \mathcal{K}_i} \exp(x_{ik} \beta^{*(R_{ik}k)}) / \sum_{r \in \mathcal{R}} \exp(x_{ik} \beta^{*(rk)}). \end{aligned}$$

where, conventionally, all elements of  $\beta^{(0k)}$  is set at zero for identification. It turns out that this is multinomial logit model of competing risks, and Assumption 7 is equivalent to the independence of irrelevant alternatives.

### Conditionally Independent Serial Durations

In this subsection, for ease of presentation, we suppose that  $\mathcal{K}_i$  is a set of consecutive integers and  $\underline{k}_i = 1$ .

### Repeated events

In some cases, repeated events are not independent of each other. For instance, once a unit has an event (e.g., electoral defeat), the unit may tend to have the event once more. Denote the number of events  $r$  unit  $i$  has before duration  $k \in \mathcal{K}_i$  by  $N_{ik}^{(r)}$ :

$$N_{ik}^{(r)} \equiv \sum_{k'=1}^{k-1} I(R_{ik'} = r),$$

where  $\sum_{k'=1}^0 \equiv 0$ .

Instead of Assumption 8, we assume stochastic independence of serial durations conditioned on the number of past events and the previous (i.e.,  $N_{ik}^{(r)}$ -th) event time  $T_{i,k_{ik}^{(r)}-1}$  where  $k_{ik}^{(r)} \equiv \min\{k' | N_{ik}^{(r')} = N_{ik}^{(r)}\}$ :

$$T_{ik}^{(r)} \perp\!\!\!\perp T_{ik'}^{(r')} | N_{ik}^{(r)}, T_{i,k_{ik}^{(r)}-1} \text{ if } k > k'. \quad (53)$$

It follows  $h_i^{(rk)}(t | \mathcal{T}_{i,<k}, R_{i,<k}) = h_i^{(rk)}(t | n_{ik}^{(r)}, t_{i,k_{ik}^{(r)}-1})$ . In the gap time (or inter-event time) specification, for  $t > t'$ ,  $h_i^{(rk)}(t | n, t') = h_i^{(rk)}(t - t' | n)$ , while, in the elapsed time (or clock time) specification,  $h_i^{(rk)}(t | n, t') = h_i^{(rk)}(t | n)$ , where we do not have to condition on  $T_{i,k_{ik}^{(r)}-1}$  in Equation 53. Either way, one way to model the hazard rate is  $h_i^{(rk)}(t | n) = h^{(rk)}(t | x_{ik}, \theta^{(rk|n)})$ . For instance, the first and second events have the same hazard rate functions ( $h^{(rk)}(t | x_{ik}, \theta^{(rk|)})$ ), though their parameters may differ ( $\theta^{(rk|0)} \neq \theta^{(rk|1)}$ ).

**Multi-state model**

Scholars may be interested in not just single but multiple risks. Often, the risk which incurred the previous non-censoring event is called ‘state’, ‘stage’, or ‘phase’ in which the unit is currently situated (Metzger and Jones, 2016). We denote it by  $R_{i,k_{ik}^{(+)}}^{(+)}$ , where  $k_{ik}^{(+)} \equiv \min\{k' | N_{ik}^{(+)} = N_{ik}^{(+)}\}$  and

$$N_{ik}^{(+)} \equiv \sum_{k'=1}^{k-1} I(R_{ik'} \in \mathcal{R}^+),$$

and suppose that the initial state is  $R_{i0} \in \mathcal{R}^+$ .

For example, suppose that  $\mathcal{R}^+ = \{‘war’, ‘crisis’, ‘peace’\}$ ,  $\mathcal{R}^0 = \{‘censoring’\}$  and  $R_{i0} = ‘peace’$ . If the  $k = 1$ -st duration ends with crisis ( $R_{i1} = ‘crisis’$ ), the state of the  $k = 2$ -nd duration is crisis ( $k_{i2}^{(+)} = 2, R_{i,k_{i2}^{(+)}-1} = R_{i1} = ‘crisis’$ ). Then, scholars would naturally expect that war in the  $k = 2$ -nd duration is more likely than in peacetime:  $h_i^{(war,2)}(t | R_{i,k_{i2}^{(+)}-1} = ‘crisis’) > h_i^{(war,2)}(t | R_{i,k_{i2}^{(+)}-1} = ‘peace’)$ .

Instead of Assumption 8, we assume stochastic independence of serial durations conditioned on the state:

$$T_{ik}^{(r)} \perp\!\!\!\perp T_{ik'}^{(r')} | R_{i,k_{ik}^{(+)}}-1 \text{ if } k > k'. \quad (54)$$

It follows  $h_i^{(rk)}(t | \mathcal{T}_{i,<k}, R_{i,<k}) = h_i^{(rk)}(t | R_{i,k_{ik}^{(+)}}-1)$ . We may model the hazard rate at state  $r$  by  $h_i^{(rk)}(t | r') = h^{(rk)}(t | x_{ik}, \theta^{(rk|r')})$ , where  $\theta^{(rk|r')}$  can be interpreted as a transition parameter vector from state  $r$  to  $r'$ . Further, we may condition on  $N_{ik}^{(+)}$  (or  $N_{ik}^{(r)}$ ) and  $T_{i,k_{ik}^{(+)}}-1$  as well.

We sometimes assume that a unit is not at risk(s) of the previous event(s). Let  $\mathcal{R}_{ik}$  denotes the (choice) set of risks at which unit  $ik$  is  $\langle \mathcal{R}_{ik} \neq \emptyset$ , and  $\mathcal{R}_{ik} \subseteq \mathcal{R} \setminus \{R_{i,k_{ik}^{(+)}}-1\}$  or  $\mathcal{R}_{ik} \subseteq \mathcal{R} \setminus \{R_{i,k_{ik}^{(+)}}-1 | k' \leq k\}$ . For example, if unit  $ik$  is at war ( $R_{i,k_{ik}^{(+)}}-1 = ‘war’$ ), it is not at risk of war any more ( $\mathcal{R}_{ik} = \{‘censoring’, ‘crisis’, ‘peace’\}$ ).

**Dependent Serial Durations**

We can apply the same methods to dependent serial durations as to dependent parallel durations in the previous section: frailty models, seemingly unrelated regressions, and copula functions. For instance, Chiba et al. (2015) examine the timing of government formation ( $T_{i1}$ ) and survival ( $T_{i2}$ ), respectively, and apply a copula model. Fukumoto (2015) studies the dependence between the duration ( $T_{i1}$ ) and outcome ( $T_{i2}$ ) of civil wars by coarsening  $T_{i2}$  into an ordered event variable and employing copula functions.

**CONCLUSION**

This chapter elaborates on duration analysis with emphasis on its duration nature rather than a point of event time. In particular, this chapter relaxes assumptions of independent multiple durations in tractable ways and explains advanced models in a systematic way. There remains, however, at least two assumptions I have still kept – the first two assumptions I make.

We may doubt even Assumption 1. For instance, if we study democratization or state formation, we may use calendar time as  $T_{ik}$ . It is not easy to nail down in which year  $t = 0$ . Rather, we may cease to assume Assumption 1 and  $t_0 = 0$  and instead may suppose that  $k$  can be a non-positive integer as well.

Another agenda is Assumption 2. An example against it is spatial correlation (Hays and Kachi, 2015, Hays et al., 2015): neighboring countries may introduce a new rule at similar timings. A shared frailty model (Equation 31) also implies that Assumption 2 does not hold. Darmofal (2009) takes spatial correlation into account by using frailty.

My hope is that readers will invent creative duration models by relaxing any assumption in a manageable way.

## Notes

- 1 For derivation of the continuous time framework from the discrete time one, see Alt et al. (2001).
- 2 Admittedly and confusingly, in the literature,  $\mu_i^*$  and  $\sigma^*$  are called 'scale' and shape parameters, respectively – though this chapter calls  $\sigma$  a scale parameter.
- 3 In early applications of this class of models,  $W_i$  refers to a hidden characteristic which makes patients frailer, that is, increases the hazard rate and shortens duration. This is why  $W_i$  is called frailty.
- 4 As for stochastic and parametric independence, see King (1989).
- 5 Admittedly and confusingly, this is different from 'risk set'  $\mathcal{I}(t)$  or  $\mathcal{I}_j$ .
- 6 Gordon (2002) assumes a proportional hazard model (Equations 24 and 25) and  $W_i^{*(r)} = \exp(W_i^{(r)})$ ,  $W_0^{(r)} = 0$  as well, though I don't think these assumptions are essential for this model.
- 7 See Trivedi and Zimmer (2007) for a concise introduction to copula. The appendix of Fukumoto (2015) gives a more handy explanation.

## REFERENCES

Alt, James E., Gary King, and Curtis S. Signorino. 2001. Aggregation among Binary,

- Count, and Duration Models: Estimating the Same Quantities from Different Levels of Data. *Political Analysis* 9(1): 21–44.
- Beck, Nathaniel, Jonathan N. Katz, and Richard Tucker. 1998. Taking Time Seriously: Time-Series-Cross-Section Analysis with a Binary Dependent Variable. *American Journal of Political Science* 42(4): 1260–1288.
- Box-Steffensmeier, Janet M., and Bradford S. Jones. 2004. *Event History Modeling: A Guide For Social Scientists*. Cambridge, UK: Cambridge University Press.
- Chiba, Daina, Lanny W. Martin, and Randolph T. Stevenson. 2015. A Copula Approach to the Problem of Selection Bias in Models of Government Survival. *Political Analysis* 23(1): 42–58.
- Darmofal, David. 2009. Bayesian Spatial Survival Models for Political Event Processes. *American Journal of Political Science* 53(1): 241–257.
- Fukumoto, K. 2009. Systematically Dependent Competing Risks and Strategic Retirement. *American Journal of Political Science* 53(3): 740–754.
- Fukumoto, Kentaro. 2015. What Happens Depends on When It Happens: Copula-Based Ordered Event History Analysis of Civil War Duration and Outcome. *Journal of the American Statistical Association* 110(509): 83–92.
- Fukumoto, Kentaro, and Mikitaka Masuyama. 2015. Measuring Judicial Independence Reconsidered: Survival Analysis, Matching, and Average Treatment Effects. *Japanese Journal of Political Science* 16(1): 33–51.
- Gordon, Sanford C. 2002. Stochastic Dependence in Competing Risks. *American Journal of Political Science* 46(1): 200–217.
- Hays, Jude, and Aya Kachi. 2015. Interdependent Duration Models in Political Science. In Robert Franzese, ed. *Quantitative Research in Political Science*. Vol. 5, pp. 33–61.
- Hays, Jude C., Emily U. Schilling, and Frederick J. Boehmke. 2015. Accounting for Right Censoring in Interdependent Duration Analysis. *Political Analysis* 23(3): 400–414.
- Homola, Jonathan, and Jeff Gill. Nd. A Flexible Class of Bayesian Frailty Models for Political Science Data. *Working Paper*.
- King, Gary. 1989. *Unifying Political Methodology*. *The Likelihood Theory of Statistical*

- Inference*. Cambridge, UK: Cambridge University Press.
- Metzger, Shawna K., and Benjamin T. Jones. 2016. Surviving Phases: Introducing Multistate Survival Models. *Political Analysis* 24(4): 457–477.
- Park, Sunhee, and David J. Hendry. 2015. Reassessing Schoenfeld Residual Tests of Proportional Hazards in Political Science Event History Analyses. *American Journal of Political Science* 59(4): 1072–1087.
- Trivedi, Pravin K., and David M. Zimmer. 2007. *Copula Modeling: An Introduction for Practitioners*. Boston, MA: Now Publishers Inc.

# Multilevel Analysis

Marco Steenbergen

In *Man, the State, and War*, Waltz (1959) outlines three different levels of analysis that help us understand war: individuals, states, and international systems. In international relations, political science, and public administration, it is common that a phenomenon can be approached at different levels. Rather than settling on one of those levels, it would be of considerable interest to bring them together into a single data-analytic framework. This is precisely what multilevel analysis offers.

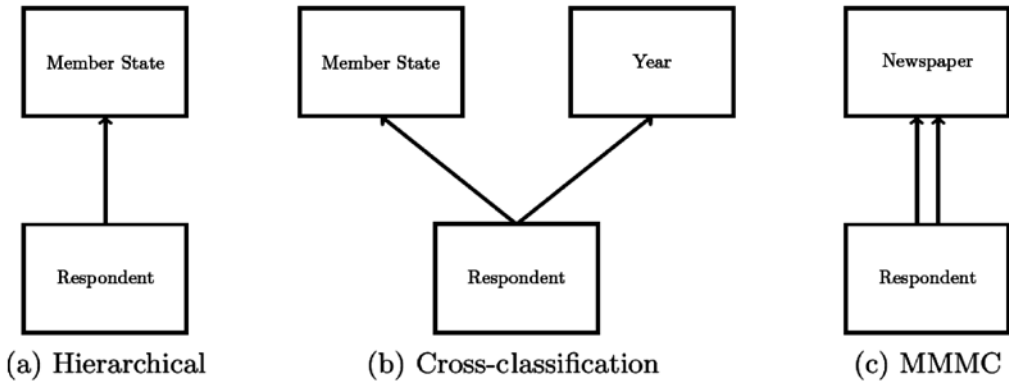
Over the past three decades, statisticians have made major breakthroughs in the analysis of multilevel data structures. Where such analysis was once limited to continuous outcomes and balanced data, it is now possible to analyze unbalanced data for categorical and limited dependent variables, as well as continuous outcomes. This chapter provides an overview of what is all now possible with multilevel data and how this can serve students of international relations, political science, and public administration.

## WHAT ARE MULTILEVEL DATA?

Multilevel data arise when we have multiple units of analysis, which stand in a (partial) hierarchy with each other and where the outcome variable varies across all units. To take the Waltz (1959) example, there are three units that form a clear hierarchy: individuals reside in states, which in turn are part of international systems. The units are called levels and a multilevel data structure consists of at least two levels.

It is possible to identify three canonical multilevel-data structures. To understand those structures, we employ the following example. Imagine we are interested in public support of the EU, analyzing Eurobarometer survey data. At first, we consider a single survey year. This yields a classical **hierarchical data structure**: we have respondents nested in EU member states. The structure is shown in panel (a) of Figure 36.1. This classification diagram (Browne et al., 2001) shows the respondents as the level-one units in the bottom box. The





**Figure 36.1** Three canonical multilevel-data structures

Source: Browne et al. (2001).

EU member states are the level-two units in the top box. An arrow points from the level-one to the level-two units to emphasize that the former is nested in the latter.

A second canonical data structure arises when we take the repeated nature of the Eurobarometer survey into account. In this example, we would bring this aspect in to explore both spatial and temporal variation in EU support. We would still treat respondents as the level-one units. It is less clear what the nesting structure of member states and years should be and, hence, we treat them equivalently. Quite literally, the second level of the data hierarchy comes about by crossing the member states with years, for example, Belgium 2016, Belgium 2017, etc. The resulting **cross-classification** is depicted as a classification diagram in panel (b) of Figure 36.1.

In the hierarchical and cross-classified data structures, it is assumed that a level-one unit may belong to only one level-two unit of a particular type. **Multiple membership and multiple classification (MMMC)** structures relax this assumption (Browne et al., 2001). Imagine that we would like to assess the effect of print media on EU support. The working assumption is that some newspapers are far more supportive of the EU in their discourse than others. We can characterize newspaper readership as a hierarchy between

the respondent and a newspaper, in the sense that a subset of respondents can be identified who read a particular newspaper and are subject to its influence. But what do we do with respondents who read multiple newspapers and can be influenced by all? The MMMC structure allows for this possibility, as is visible from the double arrows in the classification diagram in panel (c) of Figure 36.1.

We can combine the different canonical data structures. For instance, we could combine the MMMC from panel (c) with the cross-classification from panel (b). One way to do this would be to treat newspapers as being nested in member states and years. This would result in a three-level model that could be used inter alia to assess how newspaper influence affects respondent’s EU support and how this varies across space and time.

Table 36.1 shows common choices for level-one units in international relations, political science, and public administration.

**Table 36.1** Common choices of levels

<i>Lowest level</i>	<i>Higher levels</i>
Individuals	Individuals
Measures	Institutions
Time points	Geographic units
	Networks
	Temporal units

At the lowest level, the units are frequently individuals, as in the EU example. This does not have to be the case, however. We could also treat a series of related outcome measures as the lowest level, thinking of these as nested in an individual. Or, we could take a single outcome measured at different occasions for the same person as the lowest level, as we would in panel data. In this case as well, the individual is the higher-level unit. Other common higher-level units are institutions (e.g., bureaucratic agencies), geographic units (e.g., countries), networks (e.g., alliances), or temporal units (e.g., years).

## STATISTICAL PROPERTIES OF MULTILEVEL DATA

A full grasp of the rich potential of multilevel data requires that we understand their statistical properties. The most crucial property is **clustering**, which we also know from *survey research*. Consider the hierarchical data structure from Figure 36.1 and call the EU member states clusters. Clustering is the phenomena that the observations within a particular cluster (i.e., member state) are not independent because they share something in common. The common experience may be as simple as a shared language or as complex as shared norms and institutions. We may be able to capture the common experience through a set of variables. In the context of EU support, for example, we might look at a country's net contributions to the EU in order to explain why the respondents in a particular cluster hold similar views of the EU. Regardless, clustering is a fact of life when it comes to multilevel-data structures.

We capture the degree of clustering through the **intra-class correlation (ICC)**, which ranges between zero and one. When the ICC is zero, then units within the cluster have nothing in common and a blanket independence assumption is reasonable. When the ICC reaches unity, then the units within

a cluster are all alike with respect to some variable.

Clustering may be viewed from two perspectives. On one hand, it could be viewed as a statistical nuisance: it causes a number of issues that affect statistical inferences. Specifically, clustering is associated with a design effect that needs to be considered when computing (deff) standard errors (Kish, 1965). Standard errors computed under the assumption of independence are too small when the ICC exceeds zero. One can also say that we are too optimistic about the amount of data that are available. The effective sample size is  $N/\text{deff}$ , where  $N$  is the total number of observations. This could be as low as the number of clusters when the ICC is unity. In that case, it would literally suffice to study only one unit per cluster; all other units are merely duplicates. If clustering is viewed merely as a nuisance, then it is often not necessary to use multilevel statistical models. One could use fixed effects or use cluster-corrected standard errors, although the feasibility of these strategies depends on the nature of the outcome variable as well as the number of clusters (e.g., Cameron and Miller, 2015; Huang, 2016).

There is another way to approach clustering, namely as something that is of inherent theoretic interest. Clustering speaks to **heterogeneity** – how different are clusters from each other. Data analysis frequently focuses on central tendency in the extreme. However, variation is not only the proverbial spice of life but also frequently of considerable scientific interest. If we can study central tendency and variation in tandem, then we can learn about trends and deviations from those trends. Multilevel analysis is an ideal tool for developing this dual understanding of our data and political phenomena.

## THE HIERARCHICAL LINEAR MODEL

Assume we have a continuous outcome variable  $Y$ , for example, net wealth. We assume

that the variable follows a normal distribution or has been transformed to approximate normality. We observe the variable in clusters  $j = 1, \dots, J$ , for example, countries. In each cluster, we have collected data on units  $i = 1, \dots, n_j$ . Here,  $n_j$  is the cluster size. We refer to the clusters as level-two units and to the units inside, such as individuals, as level-one units. The hierarchical linear model (HLM) is suitable for analyzing these data.

## Model

### Precursors of the HLM

The HLM has several precursors. By far the simplest of these is the **random effects analysis of variance** (ANOVA) model, which has found application in *experimental design* (Searle et al., 1992). Normal ANOVA treats the level-two units as fixed, meaning that these units are the only (relevant) clusters. This results in a fixed-effects approach to the analysis (see Troeger, Chapter 33, this *Handbook*). In the baseline cell specification, this amounts to including  $J - 1$  cluster dummy variables (Jobson, 1999). An F-test on the effects of these variables tests the null hypothesis, which states no mean differences exist across the clusters. The random-effects ANOVA model operates differently: it assumes that the clusters in our data have been sampled from a population, just like the level-one units constitute a sample (Snijders, 2005). The cluster means, then, can be treated as stochastic variables: had we sampled different clusters, we would have obtained different means. Parenthetically, there is some debate in the literature whether the clusters actually constitute a random sample (Searle et al., 1992) or *in principle* are a random sample from some larger population (read ‘super-population’ in Snijders, 2005). In social-scientific practice, both ideas prevail.

We formalize this idea through a set of three model equations. The level-one model contains the outcome variable on the left-hand side:<sup>1</sup>

$$y_{ij} = \beta_{0j} + \varepsilon_{ij} \quad (1)$$

Here,  $\varepsilon_{ij} \sim NID(0, \sigma^2)$  is experimental error and *NID* stands for normally and independently distributed. Equation (1) indicates that the response of level-one unit  $i$  nested in level-two unit  $j$  is made up of a cluster mean,  $\beta_{0j}$ , and stochastic deviations from that mean, captured by the level-one error  $\varepsilon_{ij}$ . Except for its parameterization, this does not deviate from the base cell ANOVA model. The difference arises when we consider the next equation, which is stated in terms of the parameters  $\beta_j$  and is called the level-two model:

$$\beta_{0j} = \gamma_{00} + \delta_{0j}, \quad (2)$$

where  $\delta_{0j} \sim NID(0, \tau_{00})$ . This equation states that the cluster means consist of a grand mean,  $\gamma_{00}$ , and stochastic deviations  $\delta_{0j}$  – the level-two errors – from that mean. We may also write this as  $\beta_{0j} \sim NID(\gamma_{00}, \tau_{00})$ , and the fact that we consider the cluster means to be stochastic is what sets the random-effects ANOVA apart from its traditional cousin. The final model equation combines the level-one and -two models and, as such, gives the mixed model:

$$y_{ij} = \gamma_{00} + \delta_{0j} + \varepsilon_{ij} \quad (3)$$

The *NID* assumption captures the idea of **exchangeability**, which we know from *Bayesian statistics*. The assumption is important as it means that any cluster in the data could have been replaced by any other cluster in the population (Snijders, 2005). This is a key assumption of random-effects models.

We make one additional assumption in the random-effects ANOVA model, to wit  $\mathbb{E}[\delta_{0j}\varepsilon_{ij}] = 0$ . Using this and the *NID* assumptions for the level-one and level-two error terms, we obtain the following expectation and variance functions:

$$\mathbb{E}[y_{ij}] = \gamma_{00} \quad (4)$$

$$\text{Var}[y_{ij}] = \tau_{00} + \sigma^2 \quad (5)$$

It can also be shown that  $\text{Cov}[y_{ij}, y_{mj}] = \tau_{00}$ ; this is the covariance between the outcomes for level-one units belonging to the same level-two unit. The covariance for level-one units belonging to different level-two units is zero. It then follows that

$$ICC = \frac{\tau_{00}}{\tau_{00} + \sigma^2} \tag{6}$$

The particular variance–covariance structure of the random-effects ANOVA model is known as compound symmetry.

The *ICC* captures a great deal of interesting information about the outcome variable. It shows what portion of the variance is due to inter-cluster variation and what portion, namely  $1 - ICC$ , is due to intra-cluster variation. In doing so, it shows how homogeneous the level-two units are. With  $ICC = 1$ , all of the variance is between clusters and none of it is within. By contrast,  $ICC = 0$  means that all of the variance is within clusters: the level-two means do not vary. Cluster homogeneity is one important aspect of multilevel theories (Klein et al., 1999) and the random-effects ANOVA model can be used to shed empirical light on it.

An obvious limitation of the random-effects ANOVA is that it merely shows variation and does not explain it. Enter random-coefficient models (RCMs), which have a long history in econometrics (Longford, 1993; Swamy, 1970; Swamy and Tavlas, 1995). RCMs add level-one covariates to the level-one model. Thus,

$$y_{ij} = \beta_{0j} + \sum_{k=1}^K \beta_{kj} x_{ijk} + \varepsilon_{ij} \tag{7}$$

Here,  $x_{ij1}, \dots, x_{ijK}$  are values on covariates that vary across both levels. The coefficients  $\beta_{0j}, \dots, \beta_{Kj}$  are **random coefficients**. Their behavior is described by the following level-two model equations:

$$\begin{aligned} \beta_{0j} &= \gamma_{00} + \delta_{0j} \\ \beta_{1j} &= \gamma_{10} + \delta_{1j}, \\ &\vdots \\ \beta_{Kj} &= \gamma_{K0} + \delta_{Kj} \end{aligned} \tag{8}$$

where the  $\gamma$ s are known as **fixed effects**. The mixed model is given by

$$y_{ij} = \gamma_{00} + \sum_{k=1}^K \gamma_{k0} x_{ijk} + \delta_{0j} + \sum_{k=1}^K \delta_{kj} x_{ijk} + \varepsilon_{ij} \tag{9}$$

As in the random-effects ANOVA model, we assume  $\varepsilon_{ij} \sim NID(0, \sigma^2)$ . In its most general form, the level-two error structure is given by

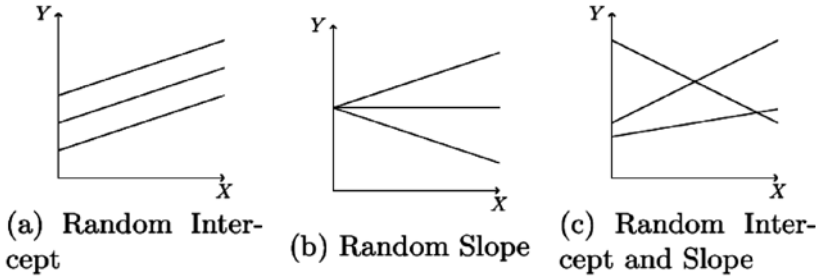
$$\delta \sim NID(0, \mathbf{T}) \tag{10}$$

Here,  $\delta$  is a vector containing all of the level-two errors,  $\mathbf{0}$  is the vector of error means, which are all zero, and  $\mathbf{T}$  is the variance–covariance matrix. This has diagonal elements  $\tau_{pp}$  ( $p = 0, \dots, K$ ), which are **variance components**, and off-diagonal elements  $\tau_{pq}$ , which are covariance components.

The level-two model captures heterogeneity writ large. It shows how the intercepts and slopes vary across level-two units. In this sense, it reflects what Western (1998) calls ‘causal’ heterogeneity. Here, I use the quotation marks deliberately because multilevel models are not inherently causal models. We can learn a great deal from studying the heterogeneity, i.e., the matrix  $\mathbf{T}$  and the coefficients  $\beta_{kj}$ . Specifically, we can ascertain how regression regimes vary across level-two units. Do the regression slopes just vary in intensity, always pointing in the same direction? Or do we have slopes that can be positive, negative, and null?

We can place restrictions on the model. Not all of the coefficients need to be random so that we can remove some of the  $\delta$  terms. This is tantamount to restricting the variances of those terms to zero in the matrix  $\mathbf{T}$ ; all of the covariances involving the error terms are then also zero. Another common strategy is to restrict all of the off-diagonal elements in  $\mathbf{T}$  to zero. As a general rule, however, one should try to avoid this strategy because it is often detrimental to the empirical fit.

Figure 36.2 illustrates a number of different random-coefficient models with a single



**Figure 36.2 Three types of random-coefficient models**

covariate. Panel (a) shows a random intercept model, which is quite common in international relations, political science, and public administration. In this model,  $\beta_{1j} = \gamma_{10}$ , and thus only the intercept is allowed to vary. This results in a series of parallel regression lines. Panel (b) shows a much less common model – the random slope model. Here  $\beta_{0j} = \gamma_{00}$ , but the slopes are allowed to vary across level-two units. Thus, we observe regression lines that fan out from a common intercept. Panel (c) shows a random intercept and slope model, meaning that both the intercepts and slopes vary across level-two units.

In addition to the assumptions of the random-effects ANOVA model, the RCM adds assumptions about the level-one predictors, specifically  $\mathbb{E}[x_{ijk}\epsilon_{ij}] = \mathbb{E}[x_{ijk}\delta_{kj}] = 0$ . This means that weak exogeneity is assumed. Note that this assumption is identical to that of random-effect models in Troeger (Chapter 33, this *Handbook*).

We conclude the discussion of RCMs by considering their implied moment structure. It can be demonstrated that

$$\mathbb{E}[y_{ij} | x_{ij1}, \dots, x_{ijK}] = \gamma_{00} + \sum_{k=1}^K \gamma_{k0} x_{ijk} \quad (11)$$

$$\begin{aligned} \text{Var}[y_{ij} | x_{ij1}, \dots, x_{ijK}] &= \tau_{00} + \sum_{k=1}^K x_{kij}^2 \tau_{kk} \\ &+ 2 \sum_{k=1}^K \tau_{01} x_{ijk} + 2 \sum_{k=1}^K \sum_{l>k}^K \tau_{kl} x_{ijk} x_{ijl} + \sigma^2 \end{aligned} \quad (12)$$

We see that the conditional variance function is now a complex function of the covariates and the (co)variance components. This reflects the heterogeneity inherent in multilevel data. The covariances between outcomes measured in different clusters is zero. When the outcomes pertain to the same level-two unit, the covariance is given by

$$\begin{aligned} \text{Cov}[y_{ij}, y_{mj} | x_{ij1}, \dots, x_{ijk}, x_{mj1}, \dots, x_{mjK}] \\ &= \tau_{00} + \sum_{k=1}^K \tau_{0k} (x_{ijk} + x_{mjk}) \\ &+ \sum_{k=1}^K \tau_{kk} x_{ijk} x_{mjk} + \\ &\sum_{k=1}^K \sum_{l=1}^K \tau_{kl} (x_{ijk} x_{mjl} + x_{ijl} x_{mjk}) \end{aligned} \quad (13)$$

This means that the ICC for  $Y$  is no longer constant, as it was in the random-effects ANOVA, but depends on the values of the covariates: the degree of clustering varies.

**The HLM**

The HLM moves beyond the RCM by allowing the addition of level-two covariates to account for heterogeneity. The RCM allows us to ascertain whether some covariate exerts an effect that varies across level-two units. The HLM allows us to explain this variation by bringing in properties of the level-two units.

Level-two properties come in two different guises (Lazarsfeld and Menzel, 1980). Derived properties are aggregates of level-one variables, whereas integral properties are inherently contextual variables that cannot be reduced to level-one attributes. An example of a derived property is the average skill level in an EU member state. An example of an integral property is the net EU contributions of a member state.

The level-one model of the HLM is identical to that of the RCM (see Equation (7)). The difference arises in the level-two model:

$$\begin{aligned}
 \beta_{0j} &= \gamma_{00} + \sum_{q=1}^Q \gamma_{0q} z_{jq} + \delta_{0j} \\
 \beta_{1j} &= \gamma_{10} + \sum_{q=1}^Q \gamma_{1q} z_{jq} + \delta_{1j} \\
 &\vdots \\
 \beta_{Kj} &= \beta_{K0} + \sum_{q=1}^Q \gamma_{Kq} + \delta_{Kj}
 \end{aligned}
 \tag{14}$$

The level-two error structure is that shown in Equation (10), although the errors should be understood as residual variation in the coefficients that has not been explained by the  $Q$  level-two predictors  $Z_{jq}$ . The mixed model is obtained by substituting Equation (14) into Equation (7):

$$\begin{aligned}
 y_{ij} &= \gamma_{00} + \sum_{q=1}^Q \gamma_{0q} z_{jq} + \sum_{k=1}^K \gamma_{k0} x_{ijk} \\
 &+ \sum_{k=1}^K \sum_{q=1}^Q \gamma_{kq} x_{ijk} \cdot z_{jq} + \\
 &\delta_{0j} + \sum_{k=1}^K \delta_{kj} x_{ijk} + \varepsilon_{ij}
 \end{aligned}
 \tag{15}$$

The variables  $x_{ijk} \cdot z_{jq}$  play an important role in the HLM. They are the **cross-level** interactions and show how the effect of a level-one covariate depends on level-two attributes. As such, they build a model of the heterogeneity

in slopes. The main effects of the predictors  $Z_{jq}$ , by contrast, model the heterogeneity in intercepts.

The matrix  $T$  also plays an important role. The change in this matrix relative to the RCM gives us an impression of how well the level-two predictors account for the variation in intercepts and slopes. If  $T = \mathbf{0}$ , then we would know that the level-two predictors account perfectly for the variation in intercepts and slopes.

This also shows an important contrast with another precursor of the HLM, to wit **contextual analysis** (Boyd and Iversen, 1979). Here, the assumption is that  $T = \mathbf{0}$ , i.e., there are no level-two errors. The HLM turns this claim into a testable proposition but does not a priori impose it as a restriction on the error structure.

Equation (15) is relatively complex. In the literature, it is often re-written in terms of matrices: let  $y$  be the vector of responses across all level-one and level-two units; let  $X$  be a matrix of level-one covariates, including a constant; let  $Z$  be a matrix of level-two covariates, also including a constant; let  $\varepsilon$  be the vector of level-one errors. Then the HLM is given by

$$\begin{aligned}
 y &= XZ\gamma + X\delta + \varepsilon \\
 &= M\gamma + X\delta + \varepsilon
 \end{aligned}
 \tag{16}$$

The model has the following expectation and variance functions:

$$\begin{aligned}
 \mu &= \mathbb{E}[y | X, Z] = XZ\gamma = M\gamma \\
 \Sigma &= \text{Var}[y] = X(I_J \otimes T)X^\top + \sigma^2 I_N
 \end{aligned}
 \tag{17}$$

where  $N$  is the total number of observations.

Ordinarily, restrictions are placed on the elements of  $\gamma$  so that not every random coefficient is accounted for by every level-two predictor. Indeed, theory should guide which predictors are used in a particular level-two equation. As in the RCM, restrictions usually are also placed on  $T$ , reducing the number

of (co)variance components. In terms of the variance–covariance matrix of the level-one errors, the restriction that this should be a diagonal matrix with identical variances can be relaxed. For example, one could allow for heteroskedasticity or serial correlation. I discuss the latter aspect in the section on longitudinal data.

### *Extending the Levels of the Model*

Returning to the EU-support example, we can actually extend the levels beyond respondents and countries. Often, we also have information about the sub-national region (or state or province) in which a respondent resides, and this may be highly relevant for his or her opinions about the EU. For example, the region may depend heavily on agricultural subsidies and this could increase EU support.

Consider level-three units  $k = 1, \dots, K$  (e.g., EU member states). Nested inside those units are level-two units  $j = 1, \dots, J_k$  (e.g., regions). Finally, we have level-one units  $i = 1, \dots, n_{jk}$  (e.g., respondents). A random-effects ANOVA now requires level-one, -two, and -three models. Specifically, the level-one model is

$$y_{ijk} = \alpha_{0jk} + \varepsilon_{ijk} \quad (18)$$

subject to  $\varepsilon_{ijk} \sim \text{NID}(0, \sigma^2)$ . The equation states that the responses consist of a level-two mean  $\alpha_{0jk}$  and a level-one deviation from that mean of  $\varepsilon_{ijk}$ . The level-two model is given by

$$\alpha_{0jk} = \beta_{00k} + \delta_{0jk} \quad (19)$$

subject to  $\delta_{0jk} \sim \text{NID}(0, \tau_\alpha)$ . This equation states that the level-two mean is made up of a level-three mean  $\beta_{00k}$  and a deviation  $\delta_{0jk}$  from that mean. Finally, the level-three model shows how the level-two means relate to the grand mean:

$$\beta_{00k} = \gamma_{000} + \omega_{00k} \quad (20)$$

where  $\omega_{00k} \sim \text{NID}(0, \tau_\beta)$ . The mixed model is

$$y_{ijk} = \gamma_{000} + \omega_{00k} + \delta_{0jk} + \varepsilon_{ijk} \quad (21)$$

We can now define two ICCs:

- 1  $(\tau_\alpha + \tau_\beta)/(\tau_\beta + \tau_\alpha + \sigma^2)$  is the share of the variance among level-two units, for example, inter-regional variation.
- 2  $\tau_\beta/(\tau_\beta + \tau_\alpha + \sigma^2)$  is the share of the variance among level-three units, for example, countries.

It is straightforward to extend the random-effects ANOVA model to an RCM. With a single level-one covariate, for example, we obtain the following random intercept and slope model:

$$y_{ijk} = \alpha_{0jk} + \alpha_{1jk}x_{ijk} + \varepsilon_{ijk} \quad (22)$$

$$\alpha_{0jk} = \beta_{00k} + \delta_{0jk} \quad (23)$$

$$\alpha_{1jk} = \beta_{10k} + \delta_{1jk} \quad (24)$$

$$\beta_{00k} = \gamma_{000} + \omega_{00k} \quad (25)$$

$$\beta_{10k} = \gamma_{100} + \omega_{10k} \quad (26)$$

We can add level-two and -three covariates to obtain a fully fledged HLM. For instance, using the same level-two predictor  $Z_{jk}$  in the equations for  $\alpha$  and the same level-three predictor  $W_k$  in the equations for  $\beta$ , we obtain the following mixed model:

$$\begin{aligned} y_{ijk} = & \gamma_{000} + \gamma_{001}w_k + \gamma_{010}z_{jk} + \gamma_{100}x_{ijk} \\ & + \gamma_{011}z_{jk}w_k + \gamma_{101}x_{ijk}w_k + \gamma_{110}x_{ijk}z_{jk} \\ & + \gamma_{111}x_{ijk}z_{jk}w_k + \omega_{00k} \\ & + z_{jk}\omega_{01k} + x_{ijk}\omega_{10k} \\ & + x_{ijk}z_{jk}\omega_{11k} + \delta_{0jk} + x_{ijk}\delta_{1jk} + \varepsilon_{ijk} \end{aligned} \quad (27)$$

The model contains three two-way cross-level interactions ( $x_{ijk}$  with  $z_{jk}$ ,  $x_{ijk}$  with  $w_k$ , and  $x_{ijk}$  with  $w_k$ ) and a three-way cross-level interaction. The last line of the equation shows the composite error term.

Extensions to more than three levels proceed in an analogous fashion. Whether it is necessary to do so obviously depends on the complexity of the data and the theory. One should not make this decision too lightly, because statistical power may be in short supply when interactions entail ever more levels. In addition, effectively communicating complex interactions can be a challenge.

**Longitudinal Data**

Time can enter the HLM in a number of different places. We could think of it as higher-level unit, as we did in panel (b) of Figure 36.1. We could also think of it as a level-one unit nested in individuals, geographic units, or institutions. This is appropriate when the data constitute a panel.

Imagine we are interested in military spending in countries across time (years). We have countries  $j = 1, \dots, J$  and time points  $i = 1, \dots, n_j$ . Note that  $n_j$  varies; there is no requirement of balanced data. A very simple model for this data is  $y_{ij} = \gamma_{00} + \delta_{0j} + \varepsilon_{ij}$  with  $\delta_{0j} \sim \text{NID}(0, \tau_{00})$  and  $\varepsilon_{ij} \sim \text{NID}(0, \sigma^2)$ . We say that spending across time and countries is expected to be  $\gamma_{00}$ . A particular country deviates from this by  $\delta_{0j}$  and in a particular year the deviation is  $\varepsilon_{ij}$ .

The problem with this basic setup is that we assume the level-one errors to be independent; this is reflected in the last term of Equation (17), which is a diagonal matrix. For time-series data, such an assumption is unrealistic. We would like to specify an alternative level-one error structure, and some of the available options are shown in Table 36.2 (cf. Kincaid, 2005). A simple structure is the (exponential) AR(1) (autoregressive-1) structure, which stipulates

$$\text{Var} \begin{bmatrix} \varepsilon_{1j} \\ \varepsilon_{2j} \\ \vdots \\ \varepsilon_{n_jj} \end{bmatrix} = \begin{bmatrix} \sigma^2 & \rho\sigma^2 & \dots & \rho^{n_j-1}\sigma^2 \\ \rho\sigma^2 & \sigma^2 & \dots & \rho^{n_j-2}\sigma^2 \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{n_j-1}\sigma^2 & \rho^{n_j-2}\sigma^2 & \dots & \sigma^2 \end{bmatrix} \tag{28}$$

Only two parameters need to be estimated:  $\sigma^2$  and  $\rho$ . The Toeplitz structure requires  $n_j$  parameters and takes the form of

$$\text{Var} \begin{bmatrix} \varepsilon_{1j} \\ \varepsilon_{2j} \\ \vdots \\ \varepsilon_{n_jj} \end{bmatrix} = \begin{bmatrix} \sigma_1 & \sigma_2 & \dots & \sigma_{n_j} \\ \sigma_2 & \sigma_1 & \dots & \sigma_{n_j-1} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n_j} & \sigma_{n_j-1} & \dots & \sigma_1 \end{bmatrix} \tag{29}$$

where  $\sigma_1$  is the level-one error variance. We can render the AR(1) and Toeplitz structures more complex by allowing for heteroskedasticity. The most complex structure arises when we allow all of the elements of the variance-covariance matrix to be freely estimated:

$$\text{Var} \begin{bmatrix} \varepsilon_{1j} \\ \varepsilon_{2j} \\ \vdots \\ \varepsilon_{n_jj} \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \sigma_{1,2} & \dots & \sigma_{1,n_j} \\ \sigma_{1,2} & \sigma_2^2 & \dots & \sigma_{2,n_j} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1,n_j} & \sigma_{2,n_j} & \dots & \sigma_{n_j}^2 \end{bmatrix} \tag{30}$$

This requires estimating  $.5n_j(n_j + 1)$  parameters and is usually unpractical.

The random-effects ANOVA model for panel data, that we have considered so far does not model any time trends. We can accommodate such trends by formulating the HLM as a **growth curve model**. The general growth curve model is given by

$$y_{ij} = \beta_{0j} + \sum_{k=1}^K \beta_{kj} t^k + \varepsilon_{ij} \tag{31}$$

$$\beta_{kj} = \gamma_{k0} + \delta_{kj} \tag{32}$$

Here,  $t$  is a time indicator such as years passed since some starting point, and  $K$  determines the order of the time polynomial. Level-two predictors can be added to explain differences in time trends.



**Table 36.2 Level-one error structures for longitudinal data**

Type	No. of parameters	$i, j$ th element
AR(1)	2	$\sigma_{ij} = \sigma^2 \rho^{ i-j }$
Toeplitz	$n_j$	$\sigma_{ij} = \sigma_{ i-j +1}$
Heterogeneous AR(1)	$n_j + 1$	$\sigma_{ij} = \sigma_j \sigma_j \rho^{ i-j }$
Heterogeneous Toeplitz	$2n_j - 1$	$\sigma_{ij} = \sigma_j \sigma_j \rho_{ i-j }$
Unstructured	$.5n_j (n_j + 1)$	$\sigma_{ij} = \sigma_{ij}$

## INFERENCE

### Identification and Sample Size

Identification of the HLM (and other multi-level models) requires that we avoid perfect multicollinearity and micronumerosity, i.e., sample sizes that fall short of the number of parameters to be estimated. This is also true in linear-regression analysis, but the difference is that the identification requirements vary with the level. For the fixed effects associated with level-one covariates, the relevant sample size is  $N$ , i.e., the total number of observations in the data. For the fixed effects associated with the level-two predictors, the relevant sample size is  $J$ , i.e., the number of clusters. The same is true for the matrix  $T$ .

The fact that  $J$  limits the number of variance and covariance components means that one typically cannot leave it to the data to show which slopes are random and which ones are not. In political science, international relations, and public administration,  $J$  is typically small. With  $K$  level-one covariates, the unrestricted  $T$  has  $.5K(K + 1)$  elements. This can be a sizable number, one that easily exceeds  $J$ . Prior restrictions on  $T$  are almost always necessary, requiring the researcher to make theoretical choices upfront.

Even if  $.5K(K + 1) \leq J$  and formal identification is assured, identification issues can still arise in practice. The off-diagonal elements of  $T$  frequently pose difficulties for the estimation. The problem tends to be more severe, the greater the dimensionality of  $T$  is.

Due to identification concerns, researchers sometimes restrict the covariance components to zero, thus turning  $T$  into a diagonal matrix. While this frees up degrees of freedom and may avoid estimation difficulties, the approach is questionable. As an empirical matter, slopes and intercepts tend to correlate. If one rules such correlation out, the model fit suffers. The HLM may also produce far different variance component estimates than those obtained from models allowing covariance components.

If one wants identification to be less of a concern, it pays off to invest in expanding  $J$ . Indeed, the payoff from expanding  $J$  is usually far greater than that of expanding  $n_j$  (Stoker and Bowers, 2002). One should take this into consideration when designing a multilevel study.

Indeed, the topic of how large of a  $J$  is required is hotly debated in the literature. An older literature argued that as many as 100 clusters would be required to draw valid inferences about random effects and (co)variance components. In the meantime, statisticians have become more relaxed on the requirement. Stegmueller (2013) argues that  $J$  can be reduced if one switches to a Bayesian framework. Browne and Draper (2000) claim that valid classical inference is possible with as few as 6–12 groups, provided that one uses restricted maximum likelihood (see below). Maas and Hox (2005) suggest  $J = 30$  with REML and advise against  $J \leq 10$ . Baldwin and Fellingham (2013) show that Bayesian inference is no panacea: when  $J$  is too small, the specification of the prior becomes ever more important, as should be expected. These

results are important because students of international relations, political science, and public administration often have relatively few clusters at their disposal. Especially in cross-national research,  $J$  is often limited and there is usually no way to expand it (we cannot generate new countries for the purposes of our research).

**Estimation**

For estimation, it makes a difference what is actually being estimated: fixed effects, variance-covariance components, or random effects. Fixed effects can be estimated using Bayesian methods, least squares, and maximum likelihood. Variance components are estimated using Bayesian inference of maximum likelihood. Random effects are typically estimated using (empirical) Bayes methods.

**Least Squares**

One of the oldest ideas in multilevel analysis is to estimate fixed effects using least squares. A two-step estimator requires that one first regresses the outcome on the level-one covariates in each cluster using OLS. Next, one stacks the OLS estimates and regresses them onto level-two predictors (Jusko and Shively, 2005; van den Eeden, 1988). The resulting fixed-effects estimator is unbiased but inefficient (de Leeuw and Kreft, 1986). The inefficiencies can be removed using generalized least squares or making adjustments to the standard errors (de Leeuw and Kreft, 1986; Lewis and Linzer, 2005).

**Maximum Likelihood**

While least squares estimation works well for fixed effects, the general consensus is that it performs less than ideally for random effects and variance components. It is because of this problem that researchers propose using maximum-likelihood estimation, which comes in two flavors: full information maximum likelihood (FIML) and restricted maximum likelihood (REML),

which is also known as residual maximum likelihood.

**FIML** takes advantage of the normality assumption we have made throughout the discussion of the HLM. Using Equation (17), we can show that  $y \sim MVN(\mathbf{M}\boldsymbol{\gamma}, \boldsymbol{\Sigma})$ , where *MVN* denotes the multivariate normal distribution. Accordingly, the log-likelihood may be written as

$$\ell_{FIML} = -.5N \ln(2\pi) - .5 \ln \det(\boldsymbol{\Sigma}) - .5(\mathbf{y} - \mathbf{M}\boldsymbol{\gamma})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{y} - \mathbf{M}\boldsymbol{\gamma}) \tag{33}$$

Writing  $\mathbf{v} = \mathbf{y} - \mathbf{M}\boldsymbol{\gamma}$ , the first-order conditions are

$$\frac{\partial \ell_{FIML}}{\partial \boldsymbol{\gamma}} = \mathbf{M}^\top \boldsymbol{\Sigma}^{-1} \mathbf{v} = 0 \tag{34}$$

$$\frac{\partial \ell_{FIML}}{\partial \boldsymbol{\theta}} = -.5 \left\{ \frac{\partial \ln \det \boldsymbol{\Sigma}}{\partial \boldsymbol{\theta}} - \mathbf{v}^\top \boldsymbol{\Sigma}^{-1} \frac{\partial \boldsymbol{\Sigma}}{\partial \boldsymbol{\theta}} \boldsymbol{\Sigma}^{-1} \mathbf{v} \right\} = 0 \tag{35}$$

Here,  $\boldsymbol{\theta}$  is some parameter, i.e., (co)variance component, of  $\boldsymbol{\Sigma}$ .

Various algorithms have been proposed to optimize the likelihood. Goldstein (1986) proposed iterative generalized least squares, whereby the algorithm iterates between estimating the fixed effects and the (co)variance components. Longford (1987) proposed Fisher scoring to accomplish the same task. Under normality, both approaches essentially produce the same results. Other algorithms include BFGS and (penalized) Newton-Raphson.

A significant problem with FIML is that the (co)variance components tend to be underestimated when the number of clusters is small. This is caused by the need to estimate the fixed effects first. That operation consumes degrees of freedom, which the estimation of the matrix  $\mathbf{T}$  and the scalar  $\sigma^2$  does not account for. In very simple terms, a (co)variance estimator divides a sum of

squared deviations or cross-products by a particular denominator and, in FIML, that number is set too high. The result is a downward bias. The problem is more pronounced for  $T$  than for  $\sigma^2$ ; the denominator in the latter case is a function of  $N$ , whereas it is a function of  $J$  for  $T$  and generally  $J \ll N$ .

REML overcomes the problem (Harville, 1974; Patterson and Thompson, 1971). The idea here is to use error contrasts  $A^T y$ . An error contrast has the property that  $A^T M = 0$  and ensures that  $A^T y$  is free from  $\gamma$ . That makes it possible to remove the bias from the (co)variance component estimates. The kernel of the REML log-likelihood function for the (co)variance components is (cf. Longford, 1993)

$$\begin{aligned} \ell_{REML} = & -.5 \ln \det(\Sigma) \\ & -.5 \ln \det(M^T \Sigma^{-1} M) \\ & -.5 (y - M\hat{\gamma})^T \Sigma^{-1} (y - M\hat{\gamma}) \end{aligned} \quad (36)$$

Computational details can be found in de Leeuw and Meijer (2008) and Longford (1993).

### Bayesian Inference

Multilevel models have a strong affinity to Bayesian approaches to statistical inference, since we can think of the higher-level models as priors of sorts. It is no surprise then that Bayesian approaches have made their way into statistical inference about HLMs. Draper (2008) provides a detailed discussion of Bayesian methods. The essence is that we derive the posterior distributions over  $\gamma$  and  $\Sigma$  by introducing priors over both sets of parameters. Different priors have been proposed. Jackman (2011) proposes using a multivariate normal prior over  $\gamma$ , an inverse Gamma prior over  $\sigma^2$ , and an inverse Wishart distribution over  $T$ . Chung et al. (2015) propose uniform priors, with the exception of  $T$ , which receives a weakly informative Wishart prior.

Estimation proceeds using MCMC, either using the Gibbs sampler Draper (2008) or the No U-Turn Sampler (Hoffman and

Gelman, 2014). One should perform the usual diagnostics to ascertain proper convergence (see *Bayesian inference*). While the Bayesian approach is already very useful in the case of hierarchical linear models, it is particularly relevant for generalized linear models, as we shall see below.

### Empirical Bayes Estimation

Inferences about random effects  $\beta_{kj}$  are either part of Bayesian inference or are computed using empirical Bayes' (EB) estimation (Carlin and Louis, 1998). Reconsidering the random-effects ANOVA model (Equations 1–2), one could estimate  $\beta_{0j}$  in two ways. First, one could fit constant-only regression models in each of the clusters. The resulting OLS estimators are  $\hat{\beta}_{0j} = \bar{y}_{.j}$  and have a variance of  $\sigma^2/n_j$ . Alternatively, one could estimate the fixed effect  $\gamma_{00}$ , the rationale being that  $\mathbb{E}[\beta_{0j}] = \gamma_{00}$ . Here, we obtain the estimator  $\hat{\gamma}_{00} = \bar{y}_{..}$  with a variance of  $\tau_{00}$ . The idea behind EB estimation is that we compute a weighted average of the two estimators, shrinking toward the most precise of the two estimators. Let

$$\lambda_j = \frac{\tau_{00}}{\tau_{00} + \frac{\sigma^2}{n_j}} \quad (37)$$

be the reliability. Then, the EB or shrinkage estimator of the cluster means is

$$\tilde{\beta}_{0j} = \lambda_j \bar{y}_{.j} + (1 - \lambda_j) \bar{y}_{..} \quad (38)$$

If  $\sigma^2 = 0$ , i.e., there is no within variance, then  $\lambda_j = 1$  and all of the weight is placed on  $\bar{y}_{.j}$ . By contrast, if  $\tau_{00} = 0$ , i.e., there is no between variance, then  $\lambda_j = 0$  and all the weight is placed on  $\bar{y}_{..}$ .

The idea can be generalized quite easily. Given the level-one regression model  $y_j = X_j \beta_j + \varepsilon_j$  and the level-two regression model  $\beta_j = Z_j \gamma + \delta_j$ , we again can derive two estimators. The within-regression estimators are given by  $\hat{\beta}_j = (X_j^T X_j)^{-1} X_j^T y_j$  and have a variance-covariance matrix of

$\sigma^2(\mathbf{X}_j^\top \mathbf{X}_j)^{-1}$ . The between-regression estimator is  $\hat{\boldsymbol{\gamma}}$  and has a variance of  $\mathbf{T}$ . Define  $\boldsymbol{\Lambda}_j = \mathbf{T}[\mathbf{T} + \sigma^2(\mathbf{X}_j^\top \mathbf{X}_j)^{-1}]^{-1}$ . Then

$$\tilde{\boldsymbol{\beta}}_j = \boldsymbol{\Lambda}_j \hat{\boldsymbol{\beta}}_j + (\mathbf{I} - \boldsymbol{\Lambda}_j) \hat{\boldsymbol{\gamma}} \tag{39}$$

Equation (39) is important because it reveals an extremely useful property of hierarchical linear models, namely the ability to **borrow strength**. Should the within estimator be very imprecise, for example because  $n_j$  is small, then we can compensate for this – i.e., borrow strength – by considering the between estimator,  $\hat{\boldsymbol{\gamma}}$ . One important application of this idea is *multilevel regression with post-stratification*.

**Hypothesis Testing**

Hypothesis tests about the fixed effects are straightforward extensions of what we know from *linear-regression analysis*. Given the null hypothesis  $H_0: \gamma_{pq} = k$ , the test statistic is given by

$$T = \frac{\hat{\gamma}_{pq} - k}{\widehat{SE}[\hat{\gamma}_{pq}]} \stackrel{asy}{\sim} \mathcal{N}(0,1) \tag{40}$$

The reliance on an asymptotic (*asy*) result is often problematic. A safer approach is to rely on the student’s t-distribution with adjusted denominator degrees of freedom (Kenward and Roger, 1997; Satterthwaite, 1941).

Hypothesis testing about the variance components is typically done via a likelihood-ratio test. The typical setup tests  $H_0: \tau_{pp} = 0$ . Such a hypothesis is relevant, for example, if we want to determine whether any variance in the intercepts or slopes remains after we have included level-two covariates. We test the hypothesis by comparing the fit of a model containing the variance component and a model that restricts it to zero. Let  $D = -2\ell$  be the deviance. Then

$$LR = D_r - D_u \stackrel{asy}{\sim} \chi_q^2 \tag{41}$$

Here,  $D_r$  and  $D_u$  are the deviances of the restricted and unrestricted models, respectively, and  $q$  equals the number of restrictions. LaHuis and Ferguson (2009) propose using a one-tailed test, as it provides the best balance between statistical power and Type-I errors.

**Fit**

When we speak of model fit, we speak of the relative performance of successive models. A typical modeling sequence starts by fitting a random-effects ANOVA, which serves the purpose of showing the variance decomposition of the outcome. Next, level-one covariates are added, which might reduce the level-one variance component. In a third step, one or more random effects are specified. A final step is to introduce level-two covariates, which might reduce the level-two variance components (cf. Hox et al., 2017).

Each step adds complexity to the model. At each step, one could compute the AIC or BIC to ascertain whether the added complexity brings a sufficient improvement in the fit. Alternatively, one can compute  $R^2$  values (Hox et al., 2017). For example, when going from the random-effects ANOVA ( $M_1$ ) to a fixed-effects model with level-one covariates ( $M_2$ ), one can compute

$$R_1^2 = \frac{\hat{\sigma}_{M1}^2 - \hat{\sigma}_{M2}^2}{\hat{\sigma}_{M1}^2} \tag{42}$$

as the variance in  $Y$  that is explained by the covariates. In going from a random intercept model ( $M3$ ) to a model that introduces level-two covariates ( $M4$ ), one can compute

$$R_2^2 = \frac{\hat{\tau}_{00}^{M3} - \hat{\tau}_{00}^{M4}}{\hat{\tau}_{00}^{M3}} \tag{43}$$

**Software**

Hierarchical linear models are now so common that all major statistical computing packages, including R, SAS, SPSS, and Stata, cover them. In addition, several stand-alone

packages for multilevel modeling remain popular, including HLM and MLwiN.

**INTERPRETATION**

**Population-Averaged Interpretation**

Most social scientists rely on a population-averaged interpretation of HLMs. Here, we take the expectation over all units, which has the effect of removing all error terms. The result is shown in Equation (17).

A common approach to interpretation is to compute the marginal effect. Specifically, in a level-two model,

$$\frac{\partial \mu}{\partial x_r} = \gamma_{r0} + \sum_{q=1}^Q \gamma_{rq} z_{jq} \tag{44}$$

$$\frac{\partial \mu}{\partial z_s} = \gamma_{0s} z_{js} + \sum_{k=1}^K \gamma_{ks} x_{ijk}$$

Obviously, these equations have to be adjusted in higher-level models. Bauer and Curran (2005) show a general approach to assessing the statistical significance of these marginal effects.

Less common is the use of a discrete-change interpretation, although it is the best approach to addressing the effects of discrete predictors. Let  $X_r$  change from  $x_r$  to  $x_r + \Delta$ . Then

$$\Delta \mu = \gamma_{r0} \Delta + \sum_{q=1}^Q \gamma_{rq} z_{jq} \Delta \tag{45}$$

A similar expression can be derived for the effects due to level-two predictors.

**Subject-Specific Interpretation**

Imagine we have measured EU support in various member states. Rather than interpreting the effect of (say) left–right for the average citizen in the average member state, we would like to see what the effect

is in a specific member state. Now we are in the domain of subject-specific effects. We take the expectation over level-one units – we are still looking at the average citizen – but fixate on a particular level-two unit. Thus, the level-one errors disappear but not the level-two errors.

If we consider the marginal effect of the  $r$ th level-one predictor, we obtain

$$\frac{\partial \mu}{\partial x_{ijr}} = \gamma_{r0} + \sum_{q=1}^Q \gamma_{rq} z_{jq} + \delta_{rj} \tag{46}$$

For estimation purposes, then, we need to add the EB residual,  $\tilde{\delta}_{rj}$ , to the population-averaged marginal effect. Nothing changes for the marginal effect due to a level-two covariate.

Moving to the discrete change, the subject-specific analogue to Equation (45) is

$$\Delta \mu = \gamma_{r0} \Delta + \sum_{q=1}^Q \gamma_{rq} z_{jq} \Delta + \delta_{rj} \Delta \tag{47}$$

Again, this is useful when the level-one covariate is a factor.

**CROSS-CLASSIFICATIONS AND MMMCS**

Imagine we consult Eurobarometer data from several years. As shown in Figure 36.1, this can be considered a cross-classified data structures with respondents  $i$  nested in country–year units  $j_1 j_2$ . An example of the structure is shown in Table 36.3, which is based on the 1970–2002 trend file. A random-effects ANOVA for this structure is given by

$$y_{i(j_1 j_2)} = \gamma_{0(00)} + \delta_{0j_1} + \delta_{j_2} + \epsilon_{i(j_1 j_2)} \tag{48}$$

with  $\mathbb{E}[y_{i(j_1 j_2)}] = \gamma_{0(00)}$  and  $\text{Var}[y_{i(j_1 j_2)}] = \tau_{00_1} + \tau_{00_2} + \sigma^2$ . Units from different rows and columns in the cross-classification are independent. Units within the same row have a covariance of  $\tau_{00_1}$ . In our example, this

**Table 36.3 A cross-classified structure**

Year	Country			
	AT	BE	...	UK
1970	⋮	$Y_{(1,2)} \cdots Y_{1296(1,2)}$	...	⋮
1971	⋮	$Y_{(2,2)} \cdots Y_{1459(2,2)}$	...	⋮
⋮	⋮	⋮	...	⋮
2002	$Y_{1(32,1)} \cdots Y_{4078(32,1)}$	$Y_{1(32,2)} \cdots Y_{4153(32,2)}$	...	$Y_{1(32,18)} \cdots Y_{5255(32,18)}$

covariance is due to the same time point of measurement. Units within the same column have a covariance of  $\tau_{00_2}$ , which in our example is due to the shared country. Units within the same cell have a covariance of  $\tau_{00_1} + \tau_{00_2}$ .

One can expand the model by adding attributes of the respondents in a random intercept and slope specification. In general,

$$\begin{aligned}
 y_{i(j_1j_2)} &= \gamma_{0(00)} \\
 &+ \sum_{k=1}^K \gamma_{k(00)} x_{i(j_1j_2)k} + \delta_{0j_1} \\
 &+ \delta_{0j_2} + \sum_{k=1}^K \delta_{kj_1} x_{i(j_1j_2)k} \\
 &+ \sum_{k=1}^K \delta_{kj_2} x_{i(j_1j_2)k} + \varepsilon_{i(j_1j_2)} \quad (49)
 \end{aligned}$$

This equation can be modified further by adding level-two covariates. In our example, those might be (1) net budgetary intakes, (2) the year that a country joined the EU, and (3) a dummy for the growth of the Dow-Jones in a particular year. The first variable varies across both countries and years, the second across countries, and the last across years.

The MMMC accommodates the possibility that a level-one unit belongs to multiple level-two units. Imagine a newspaper reader who, two-thirds into the year, switched from reading the Times to reading the Guardian. If we want to consider the effect of newspaper slant on the reader’s evaluations of Brexit, for example, then we might want to give

different weights to the two newspapers, say two-thirds to the Times and one-third to the Guardian.

In general, consider a set of weights  $w_{ij}$  such that  $\sum_j w_{ij} = 1$ . The random-effects ANOVA model now takes the form of

$$y_{ij} = \gamma_{00} + \sum_{j \in C(i)} w_{ij} \delta_{0j} + \varepsilon_{ij} \quad (50)$$

where  $C(i) \subset \{1, \dots, J\}$ . The intra-class correlation is  $\tau_{00} \sum_{j \in C(i)} w_{ij}^2 / (\tau_{00} \sum_{j \in C(i)} w_{ij}^2 + \sigma^2)$ . As always, this model can be expanded by adding level-one covariates with or without random slopes. One can also add level-two covariates, but these again will have to be weighted. For example, a measure of the slant of the Times should carry twice as much weight for our hypothetical respondent as that of the Guardian.

## GENERALIZED LINEAR MODELS

### Binary Outcomes

Many outcomes in international relations, political science, and public administration are binary. Think about the absence or presence of (civil) war, voting yay or nay in a legislature, or voting vs abstaining in an election. Such outcomes are ordinarily dealt with using logit, probit, and related models, which can easily be extended to multilevel-data structures.

Consider the decision to vote for or against Brexit in 2016. We know that the outcome of the vote varied across constituencies.

This fact can be captured most simply via an analogue of the random-effects ANOVA:

$$\begin{aligned}
 y_{ij}^* &= \beta_{0j} + \varepsilon_{ij} \\
 \beta_{0j} &= \gamma_{00} + \delta_{0j} \\
 y_{ij} &= \mathbf{1}(y_{ij}^* \geq 0)
 \end{aligned}
 \tag{51}$$

Here,  $y_{ij}^*$  is the latent disposition of voting for Brexit,  $\beta_{0j}$  is the baseline tendency in the  $j$ th constituency,  $\varepsilon_{ij}$  is voter  $i$ 's deviation from this tendency,  $\gamma_{00}$  is the baseline tendency across the UK, and  $\delta_{0j}$  is the constituency's deviation from the overall trend. We assume that a voter in this referendum voted for Brexit if her latent disposition is at least zero. As in the HLM, we assume  $\delta_{0j} \sim \text{NID}(0, \tau_{00})$ . For  $\varepsilon_{ij}$ , we typically choose the standard normal or logistic distributions, although other choices are available. The key is that the error distribution is centered about zero and has a fixed variance to identify the scale of  $y_{ij}^*$ .

The probability of the outcome being one (e.g., a vote for Brexit) is  $\pi_{ij}$ . Assuming a symmetric level-one error distribution, as in the logit and probit variants, it follows that

$$\pi_{ij} = F(\gamma_{00} + \delta_{0j})
 \tag{52}$$

Here,  $F(\cdot)$  is the cumulative distribution function. Equation (52) is the primary quantity of interest, which can be used for purposes of interpretation.

A fully fledged hierarchical model requires the addition of level-one and -two covariates. For symmetric level-one error distributions, it takes on the following form:

$$\begin{aligned}
 \pi_{ij} &= F(\eta_{ij}) \\
 \eta_{ij} &= \gamma_{00} + \sum_{q=1}^Q \gamma_{0q} z_{jq} + \sum_{k=1}^K \gamma_{k0} x_{ijk} \\
 &+ \sum_{k=1}^K \sum_{q=1}^Q \gamma_{kq} x_{ijk} \cdot z_{jq} + \\
 &\delta_{0j} + \sum_{k=1}^K \delta_{kj} x_{ijk}
 \end{aligned}
 \tag{53}$$

Here,  $\eta_{ij}$  is the so-called **linear predictor**. The model is easily adjusted to accommodate cross-classified and MMMC structures. Note that the model has some affinities with heteroskedastic logit and probit models, since the variance of  $y_{ij}^*$  is not constant.

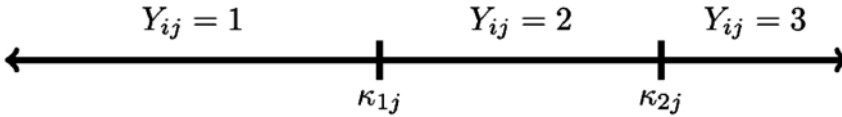
Estimation of HLMs for binary outcomes is complicated by the presence of the level-two stochastic components. Had we full knowledge of those components, then it would suffice to maximize the likelihood  $\mathcal{L}(y, \gamma | \delta_j) = \prod_{i,j} \pi_{i,j}^{y_{ij}} (1 - \pi_{i,j})^{1-y_{ij}}$  with respect to the fixed effects. In actuality, we do not know the level-two errors and therefore will need to optimize

$$\mathcal{L}(y | \gamma, \delta_j) = \int_{-\infty}^{\infty} \phi_{K+1}(\delta_j) \mathcal{L}(y, \gamma | \delta_j) d\delta_j
 \tag{54}$$

Here,  $\phi_{K+1} = \mathcal{N}(\mathbf{0}, T)$  is the  $K + 1$ -variate normal distribution.

The presence of the multivariate integral in the likelihood function makes this a difficult estimation problem, except for the special case of the random intercept model. In practice, two approaches are used. One can approximate the integrals using a Laplace approximation (Shun and McCullagh, 1995) or (adaptive) Gaussian quadrature (Skrondal and Rabe-Hesketh, 2004). One can also bypass the integration altogether and opt for inference via MCMC (Draper, 2008).

Interpretation is another area where complexities arise. Consider a random intercept model with a single level-one covariate. Here,  $\eta_{ij} = \gamma_{00} + \gamma_{10} x_{ij} + \delta_{0j}$ . For the population-averaged predicted probabilities, one might be tempted to set  $\delta_{0j} = 0$  and then take the inverse of the logit or probit link function over  $\gamma_{00} + \gamma_{10} x_{ij}$ , but this is not correct. Instead, we need to integrate out the level-two error term:  $\Pr(y_{ij} = 1 | x_{ij}) = \int \Pr(y_{ij} = 1 | x_{ij}, \delta_{0j}) \phi(\delta_{0j}) d\delta_{0j}$ . The term  $\Pr(y_{ij} = 1 | x_{ij}, \delta_{0j})$  is the subject-specific predicted probability. The integral can be evaluated numerically and some packages will do this automatically.



**Figure 36.3** Conceptualizing ordinal outcomes

One way to bypass the integral in logit models is to perform the interpretation in terms of odds ratios. The expected logit is  $\gamma_{00} + \gamma_{10}x_{ij}$ . A change of  $\Delta$  units in the level-one covariate generates a change in the logit of  $\gamma_{10}\Delta$ . This translates into a change in the odds ratio of  $\exp(\gamma_{10}\Delta)$ . A point estimate can be obtained by substituting  $\hat{\gamma}_{10}$ . For an interval estimate, one should apply endpoint transformation or the delta method to  $\hat{\gamma}_{10}$ .

**Ordinal Outcomes**

Imagine we have asked a question about whether a person feels (1) exclusively national, (2) both national and European, or (3) exclusively European. We treat this as an ordinal measure of the extent to which a person possesses a European identity ( $y_{ij}^*$ , where  $j$  is a European country and  $i$  is a respondent). The relationship between this latent variable and the observed responses is shown in Figure 36.3. If  $y_{ij}^* > \kappa_{1j}$ , then the respondent indicates a mixed identity, and if  $y_{ij}^* > \kappa_{2j}$ , she indicates an exclusive European identity. Note that we let the thresholds,  $\kappa$ , vary across level-2 units.

Define  $\pi_{ij(m)} = \Pr(Y_{ij} > m)$ , where  $m$  is a response category, as the cumulative response probability. For instance,  $\pi_{ij(1)}$  is the probability of indicating a mixed or European identity. If we do not include any predictors,  $\pi_{ij(m)} = \Pr(y_{ij}^* > \kappa_{mj}) = \Pr(\epsilon_{ij} > \kappa_{mj})$ . For a symmetric level-one error distribution, this becomes  $F(-\kappa_{mj})$ .

We can also parameterize this differently. Let  $\beta_{0j} = -\kappa_{1j}$  and let  $\theta_{mj} = \kappa_m - \beta_{0j}$  for  $m > 1$ . Thus, we introduce a random intercept and define the thresholds at fixed distances from it. Now,

$$\pi_{ij(m)} = F(\beta_{0j} - \theta_m) \tag{55}$$

This is the ordinal equivalent of the random-effects ANOVA, where  $\beta_{0j} \sim \mathcal{N}(\gamma_{00}, \tau_{00})$ . The effect of this specification is that the distance between the thresholds remains the same while their location depends on  $\beta_{0j}$ .

Table 36.4 shows how different values of  $\delta_{0j}$  affect the cumulative response probabilities for a scenario in which we have set  $\gamma_{00} = 0$  and  $\theta_2 = 1$ . We observe that probability of scoring one decreases as  $\delta_{0j}$  increases. Conversely, the probability of scoring three increases.

We can add level-one and level-two covariates to the model. Specifically,

$$\begin{aligned} \pi_{ij(m)} &= F(\eta_{ij} - \theta_m) \\ \eta_{ij} &= \gamma_{00} + \sum_{q=1}^Q \gamma_{0q} z_{qj} + \sum_{k=1}^K \gamma_{k0} x_{ijk} \\ &+ \sum_{k=1}^K \sum_{q=1}^Q \gamma_{kq} x_{ijk} \cdot z_{qj} + \\ &\delta_{0j} + \sum_{k=1}^K \delta_{kj} x_{ijk} \end{aligned} \tag{56}$$

After substitution, we obtain  $\delta_{0j} + \sum_{k=1}^K \delta_{kj} x_{ijk} - \theta_m$ , which may be viewed as context- and covariate-adjusted thresholds.

The estimation of Equation (56) proceeds analogous to the binary model. From a statistical viewpoint, the easiest interpretation is when the focus is on cumulative odds ratios. For example, in the model  $\pi_{ij(m)} = F(\gamma_{00} + \gamma_{10}x_{ij} + \delta_{0j} + \delta_{1j}x_{ij} - \theta_m)$ , the population-averaged odds ratio is  $\exp(\gamma_{10}\Delta)$  for an increase in  $X$  of  $\Delta$  units.



**Table 36.4 Cumulative response probabilities with  $\gamma_{00} = 0$  and  $\theta_2 = 1$**

	$\pi_{ij(1)}$	$\pi_{ij(2)}$	$\pi_{ij(3)}$
$\delta_{0j} = -1$	0.731	0.881	1.000
$\delta_{0j} = 0$	0.500	0.731	1.000
$\delta_{0j} = 1$	0.269	0.500	1.000

**Choice Models**

Choice variables occur in every context in which decisions are being made. Consider, for instance, voting in the UK: voters *i* decide which party, if any, receives their vote. At the same time, such decisions may be shaped by the particulars of the constituency.

In statistics, choice is often analyzed using random utility models. Let  $U_{ijm}$  be the utility that voter *i* in constituency *j* adheres to party *m*. We assume that utility consists of a systematic (*V*) and stochastic ( $\epsilon$ ) component:

$$U_{ijm} = V_{ijm} + \epsilon_{ijm} \tag{57}$$

Under utility maximization, the voter is expected to vote for party *p* if  $U_{ijp} > U_{ijr}$  for all  $r \neq p$ .

The probability of voting for party *p* is given by  $\pi_{ijp}$ . The functional form for this probability depends crucially on the assumptions about the joint distribution over the stochastic components. A simple assumption – and the only one we shall consider here – is to assume the errors to be independent draws from identical, generalized, extreme value distributions. In this case,

$$\pi_{ijp} = \sum_{m=1}^M \exp(V_{ijm}) \tag{58}$$

Here, *M* is the number of parties, which may depend on the constituency, in which case we write  $M_j$ . The model in Equation (58) is known as the Luce model.

The specifics of the model depend on how we formalize  $V_{ijm}$ . In a multinomial logit (MNL) model, we make  $V_{ijm}$  a function of

attributes of the decision maker. In the most general form, a hierarchical version of MNL takes the following form:

$$V_{ijm} = \gamma_{00m} + \sum_{q=1}^Q \gamma_{0qm} z_{qj} + \sum_{k=1}^K \gamma_{k0m} x_{ijk} + \sum_{k=1}^K \sum_{q=1}^Q \gamma_{kqm} x_{ijk} \cdot z_{qj} + \delta_{0jm} + \sum_{k=1}^K \delta_{kjm} x_{ijk} \tag{59}$$

For identification purposes, the fixed effects and level-two errors are fixed to zero for one of the alternatives. The specification in Equation (59) shows the utility for a party as a function of voter attributes, features of the constituency, the interaction between those, and stochastic elements.

If  $V_{ijm}$  depends on attributes of the alternatives, then we obtain the conditional logit model. Here, the attributes *X* and *Z* vary across the alternatives, but the coefficients do not. For identification purposes, the intercept is set to zero so that

$$V_{ijm} = \sum_{q=1}^Q \gamma_{0q} z_{qjm} + \sum_{k=1}^K \gamma_{k0} x_{ijkm} + \sum_{k=1}^K \sum_{q=1}^Q \gamma_{kq} x_{ijkm} \cdot z_{qjm} + \sum_{k=1}^K \delta_{kj} x_{ijkm} \tag{60}$$

This model can be extended by adding attributes of the voters and/or allowing for varying effects for attributes of the alternatives.

We can think of the models presented here as three-level models. Party alternatives are nested in voters, which in turn are nested in constituencies. When viewed from this perspective, it is conceptually straightforward to impose different level-one error structures to accommodate correlated errors between parties. In practice, this can create considerable estimation difficulties, so care has to be taken.

**Count Models**

A final example of a GLM is a count model. For the sake of simplicity, we restrict ourselves to the Poisson regression model. Count models are useful, for example, to study social protest activities such as strikes and demonstrations. The intensity of those activities is frequently recorded as an integer from zero onwards.

The Poisson distribution lends itself to modeling counts. Let  $Y_{ij}$  denote the number of demonstrations against the government city  $j$  during month  $i$ . The Poisson distribution gives

$$\Pr(Y_{ij} = y_{ij}) = \frac{\mu_{ij}^{y_{ij}} \exp(-\mu_{ij})}{y_{ij}!} \quad (61)$$

where  $\mu_{ij} > 0$  is the parameter. A simple model stipulates

$$\mu_{ij} = \exp(\beta_{0j}) \quad (62)$$

with  $\beta_{0j} \sim \mathcal{N}(\gamma_{00}, \tau_{00})$ . By exponentiating, we ensure that the inequality constraint on the parameter holds. The model states that the expected count,  $\mu_{ij}$ , depends on a common tendency for protest,  $\gamma_{00}$ , and a city deviation,  $\delta_{0j}$ .

A simple random intercept Poisson regression model is given by

$$\mu_{ij} = \exp(\gamma_{00} + \gamma_{10}x_{ij} + \delta_{0j}) \quad (63)$$

This is an interesting model because it has built-in overdispersion:

$$\text{Var}[y_{ij}] = \mu_{ij} + \mu_{ij}^2 (\exp(\tau_{00}) - 1) \geq \mu_{ij} \quad (64)$$

In an ordinary Poisson regression model,  $\text{Var}[y_{ij}] = \mu_{ij}$ . This is often an unrealistic constraint, especially when data are collected from different contexts.

**CONCLUSION**

Multilevel models have come a long way since they first arrived on the scene. Considerable progress has been made in the areas of inference and interpretation. Moreover, the range of outcomes that can be modeled continues to grow. In this respect, the current review has only managed to scratch the surface. We have had nothing to say, for example, about event duration, structural equations, and latent-variable models.

As strong as the potential of multilevel analysis is, we should also note that current practice in international relations, political science, and public administration realizes only a fraction of what is possible. Most of the research focuses on hierarchical data structures, uses random intercepts only, and limits itself to population-averaged interpretation. We hope that our review has shown that much more is possible and will inspire scholars to explore the many possibilities that we have sketched here.

**Note**

- 1 My notation relies heavily on Bryk and Raudenbush (1992) but deviates by following the standard econometric practice of representing error terms by a Greek symbol.

**REFERENCES**

Baldwin, Scott A. and Gilbert W. Fellingham. 2013. 'Bayesian Methods for the Analysis of Small Sample Multilevel Data with a Complex Variance Structure'. *Psychological*

- Methods* 18(2):151–64. URL: <http://doi.apa.org/getdoi.cfm?doi=10.1037/a0030642>
- Bauer, Daniel J. and Patrick J. Curran. 2005. 'Probing Interactions in Fixed and Multilevel Regression: Inferential and Graphical Techniques'. *Multivariate Behavioral Research* 40(3):373–400. URL: [http://www.tandfonline.com/doi/abs/10.1207/s15327906mbr4003\\_5](http://www.tandfonline.com/doi/abs/10.1207/s15327906mbr4003_5)
- Boyd, Lawrence H. and Gudmund R. Iversen. 1979. *Contextual Analysis: Concepts and Statistical Techniques*. Belmont, CA: Wadsworth.
- Browne, William J. and David Draper. 2000. 'Implementation and Performance Issues in the Bayesian and Likelihood Fitting of Multilevel Models'. *Computational Statistics* 15(3):391–420. URL: <http://link.springer.com/10.1007/s001800000041>
- Browne, William J., Harvey Goldstein and Jon Rasbash. 2001. 'Multiple Membership Multiple Classification (MMMC) Models'. *Statistical Modelling* 1(2):103–124. URL: <http://openurl.ingenta.com/content/xref?genre=article&issn=1471-082X&volume=1&issue=2&page=103>
- Bryk, Anthony S. and Stephen W. Raudenbush. 1992. *Hierarchical Linear Models: Applications and Data Analysis Methods*. Newbury Park, CA: Sage.
- Cameron, A. Colin and Douglas L. Miller. 2015. 'A Practitioner's Guide to Cluster-Robust Inference'. *Journal of Human Resources* 50(2):317–372. URL: <http://muse.jhu.edu/article/581178>
- Carlin, Bradley P. and Thomas A. Louis. 1998. *Bayes and Empirical Bayes Methods for Data Analysis*. Boca Raton, FL: Chapman & Hall.
- Chung, Yeojin, Andrew Gelman, Sophia Rabe-Hesketh, Jingchen Liu and Vincent Dorie. 2015. 'Weakly Informative Prior for Point Estimation of Covariance Matrices in Hierarchical Models'. *Journal of Educational and Behavioral Statistics* 40(2):136–157. URL: <http://journals.sagepub.com/doi/10.3102/1076998615570945>
- de Leeuw, Jan and Erik Meijer. 2008. Introduction to Multilevel Analysis. In *Handbook of Multilevel Analysis*, ed. Jan de Leeuw and Erik Meijer. New York, NY: Springer-Verlag pp. 1–75.
- de Leeuw, Jan and Ita Kreft. 1986. 'Random Coefficient Models for Multilevel Analysis'. *Journal of Educational Statistics* 11(1):57. URL: <https://www.jstor.org/stable/1164848?origin=crossref>
- Draper, David. 2008. Bayesian Multilevel Analysis and MCMC. In *Handbook of Multilevel Analysis*, ed. Jan de Leeuw and Erik Meijer. New York, NY: Springer pp. 77–139.
- Goldstein, Harvey. 1986. 'Multilevel Mixed Linear Model Analysis Using Iterative Generalized Least Squares'. *Biometrika* 73(1):43–56. URL: <https://academic.oup.com/biomet/article-lookup/doi/10.1093/biomet/73.1.43>
- Harville, David A. 1974. 'Bayesian Inference for Variance Components Using Only Error Contrasts'. *Biometrika* 61(2):383–385. URL: <https://academic.oup.com/biomet/article-lookup/doi/10.1093/biomet/61.2.383>
- Hoffman, Matthew D. and Andrew Gelman. 2014. 'The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo'. *Journal of Machine Learning Research* 15(April):1593–1623.
- Hox, Joop J., Mirjam Moerbeek and Rens van de Schoot. 2017. *Multilevel Analysis: Techniques and Applications*. 3rd ed. London: Routledge.
- Huang, Francis L. 2016. 'Alternatives to Multilevel Modeling for the Analysis of Clustered Data'. *The Journal of Experimental Education* 84(1):175–196. URL: <http://www.tandfonline.com/doi/full/10.1080/00220973.2014.952397>
- Jackman, Simon. 2011. *Bayesian Analysis for the Social Sciences*. Chichester: Wiley.
- Jobson, John D. 1999. *Applied Multivariate Data Analysis, Vol. 1: Regression and Experimental Design*. New York, NY: Springer-Verlag.
- Jusko, Karen Long and W. Phillips Shively. 2005. 'Applying a Two-Step Strategy to the Analysis of Cross-National Public Opinion Data'. *Political Analysis* 13(04):327–344. URL: [https://www.cambridge.org/core/product/identifier/S104719870001170/type/journal\\_article](https://www.cambridge.org/core/product/identifier/S104719870001170/type/journal_article)
- Kenward, Michael G. and James H. Roger. 1997. 'Small Sample Inference for Fixed Effects from Restricted Maximum Likelihood'. *Biometrics* 53(3):983–997. URL: <https://www.jstor.org/stable/2533558?origin=crossref>

- Kincaid, Chuck. 2005. 'Guidelines for Selecting the Covariance Structure in Mixed Model Analysis'. URL: <https://support.sas.com/resources/papers/proceedings/proceedings/sugi30/198-30.pdf>
- Kish, Leslie. 1965. *Survey Sampling*. New York, NY: John Wiley.
- Klein, Katherine J., Henry Tosi and Albert A. Cannella. 1999. 'Multilevel Theory Building: Benefits, Barriers, and New Developments'. *Academy of Management Review* 24(2): 248–253. URL: <http://journals.aom.org/doi/10.5465/amr.1999.1893934>
- LaHuis, David M. and Matthew W. Ferguson. 2009. 'The Accuracy of Significance Tests for Slope Variance Components in Multilevel Random Coefficient Models'. *Organizational Research Methods* 12(3):418–435. URL: <http://journals.sagepub.com/doi/10.1177/1094428107308984>
- Lazarsfeld, Paul F. and Herbert Menzel. 1980. On the Relation Between Individual and Collective Properties. In *A Sociological Reader on Complex Organisations*, ed. Amitai Etzioni and Edward W. Lehman. 3rd ed. New York, NY: Holt, Rinehart and Winston pp. 508–521.
- Lewis, Jeffrey B. and Drew A. Linzer. 2005. 'Estimating Regression Models in Which the Dependent Variable is Based on Estimates'. *Political Analysis* 13(04):345–364. URL: [https://www.cambridge.org/core/product/identifier/S1047198700001182/type/journal\\_article](https://www.cambridge.org/core/product/identifier/S1047198700001182/type/journal_article)
- Longford, Nicholas T. 1987. 'A Fast Scoring Algorithm for Maximum Likelihood Estimation in Unbalanced Mixed Models with Nested Random Effects'. *Biometrika* 74(4):817–827. URL: <https://www.jstor.org/stable/2336476?origin=crossref>
- Longford, Nicholas T. 1993. *Random Coefficient Models*. Oxford: Clarendon Press.
- Maas, Cora J. M. and Joop J. Hox. 2005. 'Sufficient Sample Sizes for Multilevel Modeling'. *Methodology* 1(3):86–92. URL: <https://econtent.hogrefe.com/doi/10.1027/1614-2241.1.3.86>
- Patterson, H. D. and R. Thompson. 1971. 'Recovery of Inter-Block Information when Block Sizes are Unequal'. *Biometrika* 58(3):545–554. URL: <https://www.jstor.org/stable/2334389?origin=crossref>
- Satterthwaite, Franklin E. 1941. 'Synthesis of Variance'. *Psychometrika* 6(5):309–316. URL: <http://link.springer.com/10.1007/BF02288586>
- Searle, Shayle R., George Casella and Charles E. McCulloch. 1992. *Variance Components*. New York, NY: John Wiley.
- Shun, Zhenming and Peter McCullagh. 1995. 'Laplace Approximation of High Dimensional Integrals'. *Journal of the Royal Statistical Society: Series B (Methodological)* 57(4):749–760. URL: <http://doi.wiley.com/10.1111/j.2517-6161.1995.tb02060.x>
- Skrondal, Anders and Sophia Rabe-Hesketh. 2004. *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*. Boca Raton, FL: Chapman & Hall.
- Snijders, Tom A. B. 2005. Fixed and Random Effects. In *Encyclopedia of Statistics in Behavioral Science*, ed. Brian S. Everitt and David C. Howell. New York, NY: John Wiley pp. 664–665.
- Stegmueller, Daniel. 2013. 'How Many Countries for Multilevel Modeling? A Comparison of Frequentist and Bayesian Approaches'. *American Journal of Political Science* 57(3):748–761. URL: <http://doi.wiley.com/10.1111/ajps.12001>
- Stoker, Laura and Jake Bowers. 2002. 'Designing Multi-Level Studies: Sampling Voters and Electoral Contexts'. *Electoral Studies* 21(2):235–267. URL: <https://www.sciencedirect.com/science/article/pii/S026137940100021X>
- Swamy, Paravastu A. V. B. 1970. 'Efficient Inference in a Random Coefficient Regression Model'. *Econometrica* 38(2):311. URL: <https://www.jstor.org/stable/1913012?origin=crossref>
- Swamy, Paravastu A. V. B. and George S. Tavlak. 1995. 'Random Coefficient Models: Theory and Applications'. *Journal of Economic Surveys* 9(2):165–196. URL: <http://doi.wiley.com/10.1111/j.1467-6419.1995.tb00113.x>
- van den Eeden, P. 1988. A Two-Step Procedure for Analysing Multi-Level Structural Data. In *Sociometric Research, Vol. 2: Data Analysis*, ed. Willem E. Saris and Irmtraud N. Gallhofer. London: Palgrave Macmillan pp. 180–199.

Waltz, Kenneth N. 1959. *Man, the State, and War: A Theoretical Analysis*. New York, NY: Columbia University Press.

Western, Bruce. 1998. 'Causal Heterogeneity in Comparative Research: A Bayesian

Hierarchical Modelling Approach'. *American Journal of Political Science* 42(4):1233. URL: <https://www.jstor.org/stable/2991856?origin=crossref>

# Selection Bias in Political Science and International Relations Applications

Tobias Böhmelt and Gabriele Spilker

## INTRODUCTION

Empirical scholars in international relations and, more generally, political science regularly face challenges stemming from selection bias, which arise when researchers are confronted with non-random samples. The study of the effectiveness of international institutions, for example, sees countries mainly joining those organizations whose obligations they find rather unproblematic to adhere to. This non-random self-selection, in turn, has probably severe consequences for how international institutions perform. Likewise, the decision to go to war often-times hinges on the same factors that afterwards determine their duration or outcome. Similar problems occur in almost any sub-field of political science from the study of electoral behavior to the examination of protests, public opinion, political economy or environmental politics. After a short introduction of what selection issues are in more detail, this chapter provides an overview of

some of the most commonly used estimation methods to deal with selection problems. It also illustrates several of those methods using examples of recent scholarship in the field.

A selection problem arises if measured or unmeasured factors affect both the selection into the sample, which is used to analyze a specific phenomenon, as well as the outcome variable of interest (Sartori, 2003). Imagine we would like to estimate whether countries that are part of preferential trade agreements (PTAs) including hard, i.e., enforceable, human-rights clauses actually protect human rights in their territory better than those states that are not bound by such strict clauses in their treaties (Hafner-Burton, 2005). To this end, we may consider estimating a simple linear regression model:

$$y_i = x_i \beta + \varepsilon_i$$

where  $y_i$  is some measure of human-rights protection and  $x_i$  the number of hard human-rights clauses a respective country  $i$  is bound

by in its PTAs. However, such an approach would face the problem that the countries that select themselves into PTAs with enforceable human-rights clauses might differ from those that decide to ratify PTAs with only 'soft', i.e., non-enforceable, or no human-rights clauses at all. Hence, in this case, a classic confounding problem,<sup>1</sup> there exist *unmeasured* factors that influence both the decision to ratify a PTA with hard human-rights standards and actual human-rights protection levels. One such factor, for example, is how likely the specific country is to protect human rights even in the absence of international commitments, i.e., the domestic costs of human-rights protection. This implies that one could think of a so-called selection equation that might consist of

$$P_i = w_i \gamma + u_i$$

where  $P_i$  is the probability of country  $i$  to enter into a PTA with hard human-rights clauses and  $w_i$  is some measure of how costly it is for country  $i$  to protect human rights in its territory.

The selection problem now arises due to two reasons (Achen, 1986; Sartori, 2003). First, countries finding it unproblematic due to whatever reason to protect human rights themselves, i.e., those countries having low values of  $w_i$  – the cost of human rights protection – should be more likely to enter into PTAs with hard and enforceable human rights clauses. This is simply driven by the fact that they should encounter few costs in ratifying such agreements, but mostly benefit in terms of, for example, pleasing specific domestic audiences. States with gross human-rights violations, however, i.e., those with high levels of  $w_i$ , should find it more costly to ratify such agreements to begin with and, thus, should be more likely to abstain from doing so. The second source of the selection effect comes exactly from the latter countries if they ratify PTAs with strict human-rights clauses. If these countries do so, this is likely driven by their large error terms  $u_i$ . As a consequence, our sample of analysis consists of

countries that find it easy to protect human rights and have a 'more well-behaved' range of errors, and of states that find it difficult to protect human rights and, hence, have large error terms. In turn, though, the two variables, hard human-rights clauses in PTAs and the cost of protecting human rights, will be correlated leading to biased estimation.

As a consequence of this discussion, one can posit in short that a selection problem occurs if the error term of the selection equation and the error term of the outcome equation are correlated. This leads to inconsistent estimates because the error term in the outcome equation does not have mean 0 and is correlated with the explanatory variables (Heckman, 1979). The literature proposes several ways to address selection bias in quantitative analysis. In the following, we outline five such methods. In the third section of this chapter, we discuss the application of these methods in the context of two empirical examples, the PTA and human rights example and an example of EU enlargement. We also provide R and Stata code for the straightforward replication of these examples.

## EMPIRICAL METHODS TO ADDRESS SELECTION BIAS IN QUANTITATIVE APPLICATIONS

### *The Heckman (1976) Model*

We start our discussion with the 'classical' Heckman (1976) Selection Model, in which the estimated mean function in the outcome stage is conditioned on the first stage selection process and, thereby, provides a consistent estimate for the truncated distribution of the second stage sample (Heckman 1979). It consists of a selection equation

$$s_i^* = w_i \gamma + u_i$$

$$\text{where } s^* = \begin{cases} 1 & \text{if } s^* > 0 \\ 0 & \text{if } s^* \leq 0 \end{cases}$$

and an outcome equation:

$$y_i = \begin{cases} x_i\beta + \varepsilon_i & \text{if } s^* > 0 \\ - & \text{if } s^* \leq 0 \end{cases}$$

If a researcher estimated a simple OLS regression on the outcome equation, the following quantity would be estimated given observation  $i$  is in the sample:

$$E(y_i) = x_i\beta + \theta \left[ \frac{\phi(\gamma'w)}{\Phi(\gamma'w)} \right]$$

where  $\phi$  is the standard normal distribution and  $\Phi$  is the cumulative standard normal distribution. This second part of the equation is thus the omitted variable bias due to selection therefore rendering estimation of  $\beta$  in the outcome equation inconsistent.

However, given the following additional assumptions,

$$u_i \sim N(0,1)$$

$$\varepsilon_i \sim N(0,\delta^2)$$

$$\text{corr}(u_i, \varepsilon_i) \sim \rho$$

researchers can estimate the selection and outcome variable simultaneously. The correlation of the error terms in the two stages,  $\rho$ , and its significance can then be interpreted in line of how important selection in the particular context really is. Its estimation, however, is sensitive to model specifications.

In practice, the Heckman Selection Model is commonly implemented as a two-step model, in which step one consists of estimating a probit model for the selection equation and to obtain the estimated  $\gamma$  to calculate  $\left[ \frac{\phi(\gamma'w)}{\Phi(\gamma'w)} \right]$ , which is the inverse Mills ratio, or sometimes called the 'non-selection hazard', for each observation. More precisely, the inverse Mills ratio indicates the probability for each observation to enter our sample. In the example mentioned in the introduction, it would be a function of the

probability of a specific country to enter into a PTA with hard human rights clauses. Step two implies estimating a corrected version of the outcome equation using OLS. This corrected version of the outcome equation includes as an additional regressor the inverse Mills ratio estimated from the first stage. The coefficient of the inverse Mills ratio indicates the direction of the bias that would have occurred without estimating a selection model. In particular, a positive estimate of the coefficient implies that our actual coefficient of interest, for example, PTAs with hard human-rights clauses, would have been biased upward while a negative coefficient would imply a downward bias. Important for the identification of the Heckman model is that at least one variable influences only the selection into the sample but not the outcome of interest. If this is not the case, the outcome equation will result in imprecise estimates (Sartori, 2003). This implies that researchers need to find an extra exogenous variable for selection, which must have the same properties as an instrument in instrumental-variables estimation and needs to be defended with similar care. In practice, however, such a variable is often not available and violations of the exclusion restriction are just as biasing as they are in instrumental-variables estimation (Achen, 1986).

### **Bivariate Probit Model**

The Heckman model is appropriate in situations where the selection equation is binary, but the outcome variable is continuous. In cases in which the outcome variable is also dichotomous, the Bivariate Probit Model is applicable (Greene, 2003). The Bivariate Probit Model is similar to the Heckman Selection Model:

$$y_{1i}^* = x_{1i} \beta_1 + \varepsilon_{1i}$$

$$y_{2i}^* = x_{2i} \beta_2 + \varepsilon_{2i}$$

$y_1^*$  and  $y_2^*$  are two latent variables where



$$y_{ji} = \begin{cases} 1 & \text{if } y_{ji}^* > 0 \\ 0 & \text{if } y_{ji}^* \leq 0 \end{cases}$$

for  $j=1, 2$ .

In the case of a selection problem, the errors of the two probit models will be

$$\begin{aligned} \varepsilon_{1i} &= v_i + u_{1i} \\ \varepsilon_{2i} &= v_i + u_{2i} \end{aligned}$$

This implies that each error term consists of a unique part,  $u_{1i}$  and  $u_{2i}$ , and of a component that is common to both equations,  $v_i$ . Hence, the two error terms are no longer independent. To estimate the joint probability for non-independent events, the Bivariate Probit Model assumes that the joint distribution of the two dichotomous dependent variables, one in the selection and one in the outcome equation, is a bivariate normal distribution. This implies that the Bivariate Probit Model assumes the errors of the selection and the outcome equation are independent and identically distributed as a standard bivariate normal with correlation  $\rho$ . Their joint pdf will therefore look like this:

$$\begin{aligned} \phi_2 &= \phi_{(\varepsilon_1, \varepsilon_2)} \\ &= \frac{1}{2\pi\sigma_{\varepsilon_1}\sigma_{\varepsilon_2}\sqrt{1-\rho^2}} \\ &\quad \exp\left[-\frac{1}{2}\left(\frac{\varepsilon_1^2 + \varepsilon_2^2 - 2\rho\varepsilon_1\varepsilon_2}{1-\rho^2}\right)\right] \end{aligned}$$

As in the case of the Heckman Selection Model,  $\rho$  stands for the correlation coefficient between the two error terms. However, in case of the Bivariate Probit Model, identification does not hinge upon the inclusion of an additional extra variable to the selection equation, but on a distributional assumption concerning the error terms.

**Sartori (2003) Selection Model**

Sartori (2003) starts the derivation of her selection estimator by pointing out

the difficulty of finding a truly exogenous variable needed for identification in the selection equation in Heckman-type models. She proposes a maximum likelihood estimator for binary outcomes *without* exclusion restriction. The additional assumption she introduces to identify the model is that for any observation in the sample, the error terms in the two equations are the same. She argues that while this assumption might not be true in all settings, it is a reasonable assumption if both the selection into the sample and the outcome of interest follow a similar goal or involve a similar decision, selection and outcome have the same causes, and the two aspects either cluster in time or in space (Sartori, 2003). To illustrate when these conditions are potentially met, Sartori (2003) suggests the analysis of rivalry and war as the same conditions should matter for states' decision to go to war with each other, which have resulted in them being rivals in the first place. In particular, the model looks like this:

$$\begin{aligned} U_{1i} &= x_i \beta_1 + u_i \\ U_{2i} &= x_i \beta_2 + u_i \end{aligned}$$

As in the case of the bivariate probit,  $U$  stands for a latent continuous variable for which

$$Z_{ji} = \begin{cases} 1 & \text{if } U_{ji} > 0 \\ 0 & \text{if } U_{ji} \leq 0 \end{cases}$$

for  $j = 1, 2$ .

Given the assumption that the error terms,  $u_i$ , are identical for both equations the following likelihood function is derived

$$L^* \equiv \ln L \propto \sum_{i=1}^n \sum_{j=0}^2 Y_{ij} \ln P_{ij}$$

where

$$P_{ij} \equiv \text{prob}(Y_{ij} = 1)$$

and

$$Y_{i0} = \begin{cases} 1 & \text{if } Z_1 = 0 \\ 0 & \text{otherwise} \end{cases}$$

$$Y_{i1} = \begin{cases} 1 & \text{if } Z_1 = 1 \text{ and } Z_2 = 0 \\ 0 & \text{otherwise} \end{cases}$$

$$Y_{i2} = \begin{cases} 1 & \text{if } Z_1 = 1 \text{ and } Z_2 = 1 \\ 0 & \text{otherwise} \end{cases}$$

This model can be estimated via maximum likelihood with the following estimator

$$\hat{\beta}_1, \hat{\beta}_2 := \max_{\beta_1, \beta_2 \in \Theta} L^*$$

where  $\theta$  is the vector consisting of all parameters and  $\Theta$  the parameter space.

Using Monte-Carlo simulations, Sartori (2003) shows that this estimator outperforms the Heckman Selection Model in cases in which no exogenous variable for identification exists and the assumption of identical errors for the two equations is plausible, as discussed above.

### Two-Part Model (Vance and Ritter, 2014)

While Heckman-type models can be effective in addressing selection bias, they do rest on certain assumptions, which researchers must be aware of and should be clearly spelled out. If these assumptions are not met, alternative estimators are likely more appropriate. First, Heckman-based models treat censored observations as missing data. This can be appropriate, but it may be equally plausible to treat them as 0s. As Vance and Ritter (2014: 529) state,

[w]ith respect to the modeling of foreign aid, for example, a missing value would indicate that there is some latent level of foreign aid that is unobservable to the analyst, while a zero value would indicate that the level of foreign aid is just that, zero. This distinction has far-reaching implications for both the type of model applied to the data and the conclusions drawn from it.

Second, Heckman-type models focus on *potential* rather than *actual* outcomes: the coefficient estimates capture the effect of an

explanatory variable on the outcome variable, but this is done irrespective of whether values of the dependent variable are expended.

With respect to arms exports, for example, the question arises as to whether interest really centers on modeling the latent expected value of arms exports that might have occurred under different circumstances for countries that export no arms, or on the actual observed level of exports for countries that do export arms. (Vance and Ritter, 2014: 531)

Vance and Ritter (2014), following Cragg (1971), advocate instead the use of a two-part model (2-PM). The general setup of the 2-PM is similar to that of Heckman-type estimators consisting of a selection equation

$$S_i = \begin{cases} 1 & \text{if } S_i^* = x_{1i} \beta_1 + \varepsilon_i \\ 0 & \text{if } S_i^* \leq 0 \end{cases}$$

where  $S_i^*$  is, again, a latent variable and where  $S_i = 0$  is equivalent to observing a zero on the dependent variable of the outcome equation,  $y_i$ , and  $S_i = 1$  implies that  $y_i > 0$ . The 2-PM model estimates this first selection equation using a probit specification. In a second step, 2-PM estimates an OLS regression conditional on  $S_i = 1$ :

$$E[y_i | S_i = 1, x_{2i}] = E[y_i | y_i > 0, x_{2i}] = x_{2i} \beta_2 E[\varepsilon_{2i} | y_i > 0, x_{2i}]$$

In this set-up, the prediction of the outcome variable in the second stage consists of two parts. It includes the outcome of the first stage,  $P(y_i > 0) = \Phi(x_{1i} \beta_1)$ , and it consists of the conditional expectation of  $y_i$ , namely  $E[y_i | y_i > 0]$ . The 2-PM model differs in two important ways from the Heckman Selection Model. On one hand, the inverse Mills ratio is not included in the outcome equation. On the other hand, the 2-PM produces results for actual outcomes, ‘with the coefficients measuring the effect of an

explanatory variable on the actual amount of the outcome variable expended (Vance and Ritter, 2014: 529). The 2-PM has three main advantages over the standard Heckman procedures. First, it is usually the actual outcome researchers and practitioners are interested in, not the potential one. Second, Monte Carlo simulations suggest that the 2-PM is superior. And, finally, the 2-PM is not tied to the Heckman-based identification requirement: the exclusion restriction.

Note, however, that the 2-PM is also based on another (strong) assumption (Vance and Ritter, 2014: 529): it

assumes that both parts of the model are independent conditional on the observed characteristics  $[x]$ . When this assumption is invalid, that is, when unobserved factors that affect the binary outcome are correlated with factors that affect the continuous outcome, then the Heckman Selection model may be more appropriate for corner-solution data.

### **Matching (Selection on Observables)**

Finally, we describe Matching as a technique to deal with selection problems but refer the reader to the more extensive chapter on matching by Richard Nielsen in this volume. In their study on peace-keeping operations, Gilligan and Sergenti (2008) demonstrate that purely parametric strategies, such as the Heckman Selection Model or the Bivariate Probit, can be inaccurate in addressing non-random assignment, since they rely on unverifiable modeling assumptions and are generally unable to deal with the influence of other existent covariates. The authors propose matching as a more effective solution to these problems as it corrects for the non-random assignment while controlling for the existence of confounding factors.

Matching pre-processes the data to form quasi-experimental contrasts by sampling a subset of comparable cases from the overall pool of observations. The observations contained in this subset resemble each other

as closely as possible, i.e., the differences due to confounding factors are reduced to a minimum. The only – and actually crucial – exception is that these ‘most-similar’ cases differ in whether they received the treatment, what would correspond to the outcome of the selection equation in the examples above, or not. After the matching, researchers can then estimate the effect of the treatment by analyzing the matched sample using parametric methods in order to control for any remaining imbalances (Ho et al., 2007; Morgan and Winship, 2007). In practical terms, the goal of matching is to pre-process the data such that

$$\tilde{p}(X | T = 1) = \tilde{p}(X | T = 0)$$

where  $T$  is a dichotomous variable indicating whether a unit has received the treatment ( $=1$ ) or not ( $=0$ ),  $X$  is a vector of independent variables and  $\tilde{p}(\cdot)$  ‘refers to the observed empirical density of the data, rather than a population density’ (Ho et al., 2007: 212). To obtain the matched data, several matching methods exist. First, one could employ one-to-one or exact matching in that each treated unit would be matched to a non-treated unit possessing the exact same characteristics on the other independent variable. In practice, however, this method is rarely used as it requires that exact matches need to be found, which can be difficult in practice. Furthermore, if no exact match can be found for some of the treated units, this can imply extrapolation bias or change the quantity of interest since these observations cannot be used in the later analysis.

However, matching does not require the exact pairing of the observations but only that the distributions of the treated and the non-treated units should be as similar as possible. Two other methods are therefore often used to perform the matching: propensity score and genetic matching. Propensity score matching summarizes all relevant independent variables, which here are represented as  $X$ , in a common propensity score (Rosenbaum and

Rubin, 1983) often estimated via a logistic regression:

$$\pi(X_i) \equiv \Pr(T_i = 1 | X_i) = E(T_i | X_i),$$

given the true propensity score treatment assignment and the observed covariates are conditionally independent

$$X \perp T \mid \pi(X)$$

Yet, since researchers do not know the true propensity score model, misspecification of the model can result in bias. Genetic Matching could address parts of this problem by incorporating an algorithm that iteratively checks and improves the covariate balance (Diamond and Sekhon, 2013). In particular, Genetic Matching minimizes a generalized version of the Mahalanobis distance (GMD)

$$\begin{aligned} GMD(X_i, X_j, W) \\ = \sqrt{(X_i - X_j)^T (S^{-1/2})^T W S^{-1/2} (X_i - X_j)} \end{aligned}$$

where  $W$  is a  $k \times k$  positive definite weight matrix, for which all elements except those on the diagonal are 0, and  $S^{-1/2}$  is the Cholesky decomposition of  $S$ , i.e.,  $S = S^{-1/2} (S^{-1/2})^T$ . It is possible to incorporate, in addition to the independent variables in  $X$ , also the propensity score  $\pi(X)$ .

## EMPIRICAL ILLUSTRATIONS

In light of the theoretical overview of selection problems and ways of addressing them in political science and international relations, we now turn to a series of illustrations. We ultimately opted for three such examples that, in combination, likely cover the majority of selection problems that analysts face in their own research: selection on observables, selection on unobservables and selection issues in light of overly limiting exclusion

restrictions. The first illustration is based on Spilker and Böhmelt (2013) who propose using matching methods to overcome challenges stemming from sample selection. The illustration for the second selection problem follows Plümper et al. (2006) and demonstrates the use of the ‘classical’ Heckman (1976) selection model to address selection on unobservables. Finally, we discuss the work of Vance and Ritter (2014) who have introduced a two-part model (2-PM), which is straightforward in the estimation and requires less difficult identification criteria than the more ‘traditional’ approaches to deal with sample selection.

For each illustration, we first outline the key argument and demonstrate where and how problems stemming from sample selection emerge. Second, we then summarize the results of one ‘naïve’ estimation that does not take sample selection into account. Third, we show the studies’ ways of addressing selection bias, which includes presenting the replication code of the methods used, and discuss the ‘corrected’ results, i.e., model estimates when taking selection bias seriously.

### *Illustration 1: Treaty Design and States’ Self-Selection into International Agreements*

Scholars have long assumed that PTAs can induce domestic policy changes and, most importantly for the discussion in the following, reduce human rights violations. The mechanism of this states that if these treaties, which aim at liberalizing trade between their member countries, comprise ‘hard’, i.e., legally binding and precise human rights standards, trade benefits can be made conditional on treaty members’ compliance with international human rights. In essence, by linking gains from trade to the compliance with human rights, PTAs offer a way to withhold economic benefits or impose economic sanctions in the case of abuse, torture or

repression (Hafner-Burton, 2005). Hence, PTAs comprising a ‘hard-human rights standard’ should lead to an increase in their member countries’ respect for human rights.

Spilker and Böhmelt (2013) question this presumption. They argue that previous work on PTAs and members’ human rights compliance neglects a selection process underlying the formation of these treaties: states take into account what may happen at the succeeding enforcement stage already when they establish a particular regime (von Stein, 2005; Hill, 2010). Put differently, countries are likely to be aware of the ‘shadow of the future’ and, hence, should already consider what may happen at the succeeding enforcement stage when establishing a particular PTA. This implies that countries that agree to include ‘hard’, i.e., binding and precise, human rights standards in their PTAs should differ in important and predictable ways from those countries that do not want to include these standards in their PTAs. Specifically, states should agree on ‘hard’ human rights standards in PTAs only if they have a general propensity to abide by human rights in the first place. If this selection process holds, scholars need to acknowledge countries’ preferences for the establishment of international institutions, including PTAs, when studying their effects in order to avoid biased inferences. It is therefore crucial for studies examining the effectiveness of PTAs in promoting human rights that both theory and empirics acknowledge the factors that motivate countries to include hard human rights standards and those that do not require the inclusion of any human rights clauses. Otherwise, it may well be that the findings we obtain are spurious or biased.

To examine the effect of hard human rights standards in PTAs on countries’ compliance with human rights, Spilker and Böhmelt (2013) have compiled a monadic country-year data set on human rights compliance and PTAs between 1976 and 2009. Ultimately, their time-series cross-section data comprise 4,117 country-years for 174 countries with

249 PTAs. The outcome variable in this study, *Political Repression*, is operationalized by the level of political terror, i.e., data on ‘murder, torture or other cruel, inhuman or degrading treatment or punishment; prolonged detention without charges; disappearance or clandestine detention; and other flagrant violations of the right to life, liberty, and the security of the person’ (Hafner-Burton, 2005: 615–619). The final item follows Gibney et al.’s (2011) 5-point ordinal scale operationalization of the Political Terror Scale with higher values pertaining to more governmental repression. The core explanatory variable in this study, and also the treatment item that helps identifying sample selection, is *PTA Hard Law*, which measures state membership with PTAs supplying hard standards: an observation takes on a value of 1 in a specific year if a state belongs to any PTA with hard law human rights standards.

With these specifications in mind, Spilker and Böhmelt (2013) first estimate a naïve model that ignores the likely selection effect. This model is a replication of Hafner-Burton’s (2005) ordinal logistic regression. Several control variables are included in this model, but we omit them from the presentation in the following. Instead, we focus on the impact of *PTA Hard Law*. As demonstrated in Table 37.1a, this variable exerts a substantial and significantly negative impact on governmental repression (significant at the 0.1 level). However, this finding might be misleading and actually driven by a selection process, since tough human rights standards

**Table 37.1a Naïve estimation**

	<i>Model 1</i>
PTA Hard Law	–0.318 (0.193)
Observations	4,117
Log Pseudolikelihood	–5,243.55
Wald $\chi^2$	142.44

Table entries are coefficients. Robust standard errors clustered on country in parentheses.

should be systematically included in PTAs due to certain kinds of country interests or domestic characteristics.

Spilker and Böhmelt (2013) employ matching to address this problem. Matching corrects for the non-random assignment while controlling for the existence of confounding factors. Matching pre-processes the data to form quasi-experimental contrasts by sampling a subset of comparable cases from the overall pool of observations. The observations contained in this subset resemble each other as closely as possible, i.e., the differences due to confounding factors are reduced to a minimum. The only – and actually crucial – exception is that these ‘most-similar’ cases differ in whether they received the treatment (*PTA Hard Law*) or not. After the matching, one can estimate the effect of the treatment by analyzing the matched sample using parametric methods in order to control for any remaining imbalances. Specifically, Spilker and Böhmelt (2013) use genetic one-to-one matching with replacement (Sekhon, 2007; Diamond and Sekhon, 2013), refraining from matching on all explanatory variables due to two reasons. First, this would not avoid matched datasets with still significant imbalances. Second, and in the words of Ho et al. (2007: 216f):

the theoretical literature emphasizes that including variables only weakly related to treatment assignments usually reduces bias more than it will increase variance, and so most believe that all available control variables should always be included. However, the theoretical literature has focused primarily on the case where the pool of potential control units is considerably larger than the set of treated units. Some researchers seem to have incorrectly generalized this advice to all datasets. If, as is often the case, the pool of potential control units is not much larger than the pool of treated units, then always including all available control variables is bad advice.

The approach in Spilker and Böhmelt (2013), thus, corresponds to the general genetic algorithm used by Sekhon (2007: 12ff), which maximizes the smallest  $p$ -value for  $t$ -tests in

each iteration of the matching procedure. The **R** code for the matching used by Spilker and Böhmelt (2013) is:

```
library("foreign")
library("Matching")
mydata<-read.dta("01_RIO shapefile.
dta")
attach(mydata)

Des<-cbind(ccode, year)

X<-cbind(trade_lag_log, polity2_lag,
hras_lag)
BalanceMatrix<- cbind(trade_lag_log,
gdp_lag, polity2_lag, durable_lag,
density_lag, hras_lag)

gen1<-GenMatch(Tr=hard_lag, X=X,
BalanceMatrix=BalanceMatrix, pop.
size=1000)

mgen1<- Match(Y=repression_final,
Tr=hard_lag, X=X, Weight.matrix=gen1)
balancetest<-MatchBalance(hard_
lag~trade_lag_log+gdp_lag+polity2_
lag+durable_lag+density_lag +hras_
lag, match.out=mgen1, nboots= 1000)
attach(mgen1)

U<-cbind(mgen1$mdata$Y, mgen1$mdata$Tr,
mgen1$mdata$X, mgen1$index.treated,
mgen1$index.control)

V<-d(BalanceMatrix[index.treated,],
BalanceMatrix[index.control,])
X<-rbind(Des[index.treated,], Des
[index.control,])
GenMatch1data<-cbind(U, V, X)
GenMatch1dataset<-data.frame
(GenMatch1data)
summary(mgen1, full=TRUE)
write.dta(GenMatch1dataset, file="02_
Matched data hard PTA")
```

Ho et al. (2007: 211f) suggest using the same parametric estimator for the matched data one would have employed in the first place, i.e., before the matching. Table 37.1b summarizes the main results (control variables omitted) that are based on the matched data. Indeed, unlike in Table 37.1a above or Hafner-Burton (2005), *PTA Hard Law* is insignificant. In other words, the effect of hard human rights standards vanishes once

**Table 37.1b Corrected estimation**

	<i>Model 2</i>
PTA Hard Law	0.191 (0.195)
Observations	2,754
Log Pseudolikelihood	-3,058.422
Wald $\chi^2$	468.19

Table entries are coefficients. Robust standard errors clustered on country in parentheses.

selection is taken seriously: countries agree on including hard law human rights standards in PTAs only if they intend to comply with these standards anyway, i.e., if there is a general tendency to abide by human rights in the first place. Under those circumstances and considering the nature of selection, however, it is unlikely that PTAs can exert a causal impact on states' human rights compliance.

### ***Illustration 2: A Two-Stage Selection Process of EU Enlargement***

Our second illustration focuses on the enlargement process of the European Union (EU). States seeking to join the EU have to adjust legislation prior to accession: laws, regulatory frameworks and administrative practices all have to be brought in line with the *acquis communautaire* (Hillion, 2002; Böhmelt and Freyburg, 2013). Scholars generally agree that the conditional incentive of EU membership was the main force driving the incorporation of the *acquis* by candidate countries (Schimmelfennig and Sedelmeier, 2004). Accession conditionality aims to induce formal and practical compliance with the EU accession criteria as an instrumentally and strategically calculated reaction by the target countries in response to the incentive of EU membership. In other words, governments comply with EU law if the calculated benefits of membership exceed the expected political costs of compliance that

are associated with the accession criteria. Thus, EU conditionality is a (limited) bargaining strategy of 'reinforcement by reward' (Schimmelfennig and Sedelmeier, 2004: 662) used by the EU to make the target countries comply with its conditions.

This perspective treats the EU and applicant countries as two different actors that pursue their own interests in a bargaining environment, but are also closely linked to each other with actions on one side influencing the other. However, the dominant view has been for a long time to study EU enlargement as 'EU-centric', which was driven by the assumption that it is the EU setting the rules and procedures. Plümper et al. (2006) were among the first to move beyond this overly simplistic approach. On one hand, they explicitly sought to model the interaction of the two independent players, i.e., the EU and a potential member state. On the other hand, Plümper et al. (2006) suggested seeing EU accession as a two-stage process: a first stage that pertains to an individual country's decision to apply for membership, and a second stage focusing on the EU as an actor making a decision on whether or not to accept these applicants as new members.

It is precisely this two-stage process that comprises a self-selection process: as Plümper et al. (2006: 17–18) state, 'the EU's repeated declaration to accept only stable democracies with a market-oriented economic system and the clearly stated regulatory conditionality resulted in a self-selection process among transition countries'. This mechanism induced two related, albeit different, outcomes. First, the set of states seeking membership in the EU is not random. Overly autocratic countries, for example, with few possibilities (or incentives) to democratize were excluded ex-ante and, thus, would probably not even submit an application in the first place. Second, the set of countries that eventually applied for EU membership were somewhat similar in their characteristics, for example, some satisfactory degree of democracy to begin with. Applicants thus varied

little in key features that may have motivated them to apply for EU membership in the first stage but would be of little use to the EU in the second stage when deciding to grant membership. As a result, ‘while the factors identified in the enlargement literature may explain the self-selection among the transition countries, they do not contribute much to the EU’s selection amongst applicant countries’ (Plümper et al., 2006: 18).

In sum, EU accession is a strategic, two-stage selection process in which potential members decide whether to apply or not, before, secondly, the EU decides to grant membership to some (or all) of those that submitted an accession application. Both stages are intertwined and ignoring the prior selection stage when evaluating the EU’s actions in the second phase may induce biased estimates. Plümper et al. (2006: 23) eventually argue that despite the strong association between the two levels, there are different factors shaping their outcomes in that the level of

democracy matters only in the first stage, where political leaders in transition countries decide whether to apply or not. In the second stage, the EU primarily uses information on the revealed preferences of political parties. We expect that the strength of parties hostile to the EU and the capability of these countries to implement *acquis communautaire* reforms have the greatest influence on EU’s accession decision.

However, ‘simply excluding non-applicant countries would cause a severe estimation bias that might lead to wrong inferences’ (Plümper et al., 2006: 26).

Plümper et al. (2006) have compiled a monadic, country-year time-series cross-section data set comprising all European transition countries between 1990 and 2001. As above, we start empirically by estimating a naïve model, which does not take into account the outlined selection problem. For that model, we concentrate on Plümper et al.’s (2006) second stage. The dependent variable here is binary, receiving a value of 1 if a country has been selected by the EU

Commission for early accession according to the Copenhagen decision, or 0 otherwise. The explanatory variables pertain to the number of closed chapters and the existence of euroskeptic parties in government. First, applicants have to negotiate and close each of the 30 chapters of the *acquis communautaire* before accession is granted, and the chapter variable comprises scores from 0 to 30 as a result. Second, there is a dichotomous variable on whether a euroskeptic party is part of an executive coalition. It is expected that the chapter variable should be positive correlated with the outcome, while the euroskeptic dummy exerts a negative influence. Table 37.2a is our ‘naïve’ estimation based on a probit model.

When not taking the presumed selection effect into account, *Chapter* is positively signed and statistically significant – as expected. However, the second variable on euroskeptic parties in government is also positively signed and significant at the 10% level, suggesting that euroskeptic parties in government have actually increased the likelihood of the EU granting early accession. Clearly, this finding appears less plausible and, in fact, points to the possibility that we have not taken selection seriously into account. In order to do so, we need to modify the naïve estimation along the lines proposed in Plümper et al. (2006): we must specify variables for the first stage and, in addition, change the estimator. On one hand,

**Table 37.2a Naïve estimation**

	<i>Model 1</i>
Chapter	0.269 (0.048)
Gov. Participation of Euroskeptic Parties	0.809 (0.463)
Constant	-2.240 (0.379)
Observations	120
Log Likelihood	-21.457
Wald $\chi^2$	77.18

Table entries are coefficients. Standard errors in parentheses.



we add a first stage considering the argument of a two-stage process in Plümper et al. (2006), which then captures states' decision whether to apply for EU membership or not. The outcome variable here indicates whether (1) or not (0) a state applies for membership in the EU. Plümper et al. (2006) focus on two explanatory variables for this first stage: a measure from the European Bank for Reconstruction and Development on regulatory quality and a modified version of the *polity2* scale of the Polity IV database. Both variables are expected to positively influence the likelihood of a country applying for EU membership.

On the other hand, we now use a probit version of the classical Heckman (1976) selection model. This 'probit model with sample selection' was originally developed by Van de Ven and Van Praag (1981) and it endogenizes, in this context, the decision to apply for EU membership in the first stage, with the estimated probability of non-application then being used as a regressor in the second stage on early accession (Plümper et al., 2006: 26). In this model,

the estimated mean function in the second stage is conditioned on the selection process of the first stage. [...] It reflects well the self-selection process in the first stage and also assumes that the probability of a country's non-application bears an influence on the likelihood of accession in the second stage. (Plümper et al., 2006: 26)

In the first stage, all possible applicant countries for EU membership are included in 1990–2001, while the second stage only considers the self-selected sample. The code for implementing this in Stata is as follows:

```
heckprob accneg lag_chapter lag_
  oppcoal, select (euapp = lag_regqual
  lag_democ) first nocon
```

The command `heckprob` calls the probit selection model, while `accneg` is the outcome variable in the second stage. The selection equation or first stage is specified in the second part of the code above, with `euapp`

**Table 37.2b Corrected estimation**

	<i>Model 2</i>
<u>Outcome stage</u>	
Chapter	0.128 (0.042)
Gov. participation of euroskeptical parties	–0.599 (0.255)
<u>Selection stage</u>	
Regulatory quality	1.975 (0.381)
Democracy	0.444 (0.152)
Constant	–9.564 (1.456)
Observations	182
Log likelihood	–72.212
Wald $\chi^2$	12.43
LR test $\rho=0$	29.15 (p<0.000)

Table entries are coefficients. Standard errors in parentheses.

being the dependent variable here. The option `first` specifies that the first-stage probit estimates are shown, while `nocon` suppresses the constant term as it is not required (Plümper et al., 2006: 37). Table 37.2b summarizes the findings from the Heckman model. In the first stage, both explanatory variables exert a positive influence, which is expected according to the theory. Countries that were more democratic and had higher regulatory quality were more likely to apply for accession. Coming to the second stage, the number of concluded chapters is, as it was the case in the naïve estimation above, positively correlated with the chances of accession. Strikingly, though, *Gov. Participation of Euroskeptical Parties* is now negatively signed and, hence, has the opposite sign of the coefficient in the naïve model. The corrected model thus suggests that more euroskeptical parties in the government do in fact lower the chances of accession. Finally, the significant estimate of the  $\rho$  parameter indicates that the Heckman model fits the data better than independent estimates of the selection and outcome equations.  $\rho$  is an estimate of the correlation of the error terms in the two stages. The estimate of  $\rho$  is

expected to be negative if unobserved features that increase the likelihood of selection decrease the probability accession, which is given in this case.

### ***Illustration 3: The Self-Selection into Peace and Conflict – A Two-Part Model Application***

The selection problem underlying our third illustration is similar to the ones discussed above, i.e., the analysis of non-random data that induces bias as the error term is correlated with the explanatory items. However, we discuss an estimator that allows researchers to circumvent some of the problems associated with the ‘classical’ Heckman selection models. Recall that Heckman-type models jointly estimate two stages: a selection equation that is usually based on a probit model, which captures the determinants of censoring; and an outcome equation, usually based on OLS, which focuses on the non-censored observations. The key component of such models is that the latter stage comprises the inverse Mills ratio, i.e., the ratio of the density function of the standard normal distribution to its cumulative density function, estimated from the first stage as an additional regressor (Vance and Ritter, 2014: 529).

While Heckman-type models can be effective in addressing selection bias, they do rest on certain assumptions, which might be problematic in practice as discussed in Section 2 of the chapter above. In the following, we there for illustrate the usefulness of the 2-PM as proposed by Vance and Ritter (2014). We present a basic analysis based on Vance and Ritter (2014) who employ Sweeney’s (2003) data on militarized interstate dispute intensity. The selection process is based on incidence and severity of interstate disputes: the outcome variable is censored at 0 for cases in which severity is low or a conflict has not occurred. In what follows, we focus on *Contiguity*, which is a binary item capturing

whether two states in a dyad are contiguous or not. But there are also controls for power (*Capability Ratio*), regime type, trade, alliance, IGO links and countries’ major power status. As before, we present a naïve estimation, which is a simple OLS model using intensity as the outcome variable, and a corrected model, which is based on the 2-PM.

Note the statistical insignificance of *Contiguity* in the naïve model. The item becomes statistically significant, however, when using a specification based on the 2-PM. The replication code used for Table 37.3b is depicted below:

```
probit disputex ln_capratio smlpmat
smldep smigoabi allies contigkb log-
dstab majpower
estimates store v
regress brl2 contigkb if disputex==1
estimates store z
suest v z, robust
```

In a first step, a probit model on dispute incidence (1 = incidence; 0 = no incidence) is estimated, while the estimates are stored in a

**Table 37.3a Naïve estimation**

	<i>Model 1</i>
Contiguity	6.298 (4.083)
Capability Ratio	-12.258 (8.737)
Democracy	0.357 (0.337)
Dependence	-1058.279 (345.705)
Common IGOs	-0.515 (0.114)
Allies	12.610 (4.125)
Distance	-2.581 (1.916)
Major Power	1.507 (3.895)
Constant	78.763 (5.924)
Observations	972

Table entries are coefficients. Standard errors in parentheses.

**Table 37.3b Corrected estimation**

	<i>Model 2</i>
<u>Outcome Stage</u>	
Contiguity	9.670 (3.513)
Constant	64.105 (2.874)
<u>Selection Stage</u>	
Contiguity	1.034 (0.038)
Capability Ratio	-0.670 (0.076)
Democracy	-0.026 (0.003)
Dependence	-19.731 (3.301)
Common IGOs	0.003 (0.001)
Allies	-0.180 (0.040)
Distance	-0.176 (0.015)
Major Power	0.681 (0.035)
Constant	-3.167 (0.046)
Observations	149,004

Table entries are coefficients. Standard errors in parentheses.

second step (*estimates*). Third, we run an OLS model using dispute intensity as the outcome, employing contiguity as an explanatory variable and calculating this conditional on a dispute having occurred. The estimates are stored again, while the final step in the 2-PM is the estimation of a seemingly unrelated regression using the *suest* command in *Stata*. When running these commands, *Contiguity* now becomes statistically significant in Table 37.3b

While the 2-PM can outperform the Heckman-based estimation procedures, there are no ‘hard and fast rules that point to the superiority of one model over the other in any given situation’ (Vance and Ritter, 2014: 536). Instead, theoretical, practical,

and statistical considerations should drive the choice. However, given the right assumptions and if researchers want to focus on actual rather than potential outcome values while circumventing the sometimes overly restrictive identification requirement of any Heckman procedure, the 2-PM can be a powerful estimation tool.

## CONCLUSION

Sample selection and selection bias can pose serious challenges to inferences drawn from systematic data analysis. If neglected, as we have sought to demonstrate theoretically and via three different illustrations, it can bias results quite severely, rendering any conclusions for practitioners and policymakers useless. Fortunately, methods in the social sciences and data analysis have advanced, and there now exist various methodological approaches – albeit with sometimes very strong assumptions that are untestable – for effectively addressing problems stemming from sample-selection bias.

The main motivation of this chapter was to define what selection bias is, what consequences it can have and from an application-based point of view to show where and how this is given in actual empirical work. We have sought to focus on the most important challenges and solutions in this context, but space limitations clearly prevent us from reviewing the broad range of all possible solutions to selection bias that social science methodology has identified. Leemann’s (2014) Strategic Selection Estimator is one of the approaches we could not fully discuss in this chapter, although such strategic estimators are likely to gain importance in the future.

## Note

- 1 Heckman (1976, 1979) introduces his estimator not in the context of a confounding problem, such as our PTA and human-rights example, but

for a situation in which the dependent variable is only observable for specific observations. While in our case it is not random which country enters into which PTA, in the classic Heckman example of women's hourly wages, we cannot observe the dependent variable, i.e., wages, for those women who decided not to go to work, which makes it problematic to estimate, for example, the effect of having children on wages.

## REFERENCES

- Achen, Christopher. 1986. *The Statistical Analysis of Quasi-Experiments*. Berkeley: University of California Press.
- Böhmelt, Tobias, and Tina Freyburg. 2013. The temporal dimension of the credibility of EU conditionality and candidate states' compliance with the *acquis communautaire*, 1998-2009. *European Union Politics* 14(2): 250-272.
- Cragg, John G. 1971. Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica* 39(5): 829-844.
- Diamond, Alexis, and Jasjeet Sekhon. 2013. Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *Review of Economics and Statistics* 95(3): 932-945.
- Gibney, Mark, Linda Cornett, and Reed Wood. 2011. Political terror scale 1976-2009. Available at the political terror scale website: <http://www.politicalterrorsscale.org/>.
- Gilligan, M. J., and E. J. Sergenti. 2008. Do UN interventions cause peace? Using matching to improve causal inference. *Quarterly Journal of Political Science* 3(2): 89-122.
- Greene, William. 2003. *Econometric Analysis*. New Jersey: Prentice Hall.
- Hafner-Burton, Emilie. 2005. Trading human rights: How preferential trade agreements influence government repression. *International Organization* 59(3): 593-629.
- Heckman, James J. 1976. The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement* 5(4): 475-492.
- Heckman, James. 1979. Sample Selection Bias as a Specification Error. *Econometrica* 47: 153-161.
- Hill, Daniel. 2010. Estimating the effects of human rights treaties on state behavior. *Journal of Politics* 72(4): 1161-1174.
- Hillion, Christophe. 2002. EU enlargement: A legal analysis. In: Anthony Arnall and Daniel Wincott (eds). *Legitimacy and accountability in the European Union*. Oxford: Oxford University Press, pp. 401-419.
- Ho, Daniel, Kosuke Imai, Gary King, and Elizabeth Stuart. 2007. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis* 15(3): 199-236.
- Leemann, Lucas. 2014. Strategy and sample selection: A strategic selection estimator. *Political Analysis* 22(3): 374-397.
- Morgan, Stephen L., and Christopher Winship. 2007. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. *Analytical Methods for Social Research*. New York: Cambridge University Press.
- Plümper, Thomas, Christina Schneider, and Vera Troeger. 2006. The politics of EU eastern enlargement: Evidence from a Heckman selection model. *British Journal of Political Science* 36(1): 17-38.
- Rosenbaum, Paul R., and Donald B. Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70: 41-55.
- Sartori, Anne E. 2003. An estimator for some binary-outcome selection models without exclusion restrictions. *Political Analysis* 11(2): 111-138.
- Schimmelfennig, Frank, and Ulrich Sedelmeier. 2004. Governance by conditionality? EU Rule transfer to the candidate countries of Central and Eastern Europe. *Journal of European Public Policy* 11(4): 661-679.
- Sekhon, Jasjeet. 2007. Multivariate and propensity score matching software with automated balance optimization: The matching package for R. *Journal of Statistical Software* 42(1): 1-52.
- Spilker, Gabriele, and Tobias Böhmelt. 2013. The impact of preferential trade agreements on governmental repression revisited. *Review of International Organizations* 8(3): 343-361.

- Sweeney, Kevin J. 2003. The severity of interstate disputes: Are dyadic capability preponderances really more pacific? *Journal of Conflict Resolution* 47(6): 728–750.
- Van de Ven, Wynand, and Bernard Van Praag. 1981. The demand for deductibles in private health insurance: A probit model with sample selection. *Journal of Econometrics* 17(2): 229–252.
- Vance, Colin, and Nolan Ritter. 2014. Is peace a missing value or a zero? On selection models in political science. *Journal of Peace Research* 51(4): 528–540.
- Von Stein, Jana. 2005. Do treaties constrain or screen? Selection bias and treaty compliance. *American Political Science Review* 99(4): 611–622.

# Dyadic Data Analysis

Eric Neumayer and Thomas Plümpert

## INTRODUCTION

Humans are social beings because it is in their interest and their nature to interact with each other. Scholars have studied social relations since at least the work of Georg Simmel and Emile Durkheim, who both argued that social relations should be the very focus of social scientists. Yet, though the social sciences carry the label ‘social’ for good reason, arguably we devote by far the largest share of our attention and resources to the analysis of *individual* preferences, strategies, behaviours and outcomes (Krasikova and LeBreton, 2012). The analysis of the causes and consequences of social interactions, while important, remains a rather neglected field of research.

One important, though by no means the only, way of studying social interactions is by employing dyadic data, defined as data in which the unit of analysis is the relation between two actors, as opposed to monadic data, in which the unit of analysis is an individual actor. Dyadic data require a

link, an interaction, a relation or a contract between two actors. Dyadic data analyses are employed in political science, international relations, international economics, psychology and sociology, among others. Poast (2016: 369) calls it ‘the standard research design in quantitative international relations’. Nevertheless, given that the analysis of social relations, interactions and so on ought to be hardwired into the DNA of the social sciences, why is it that dyadic analyses of social relations are relatively few and far between across the social sciences?<sup>1</sup>

We believe that the widespread neglect of dyadic analyses results from the added difficulty of analysing dyadic data, which shares many of the specification problems and inferential challenges that exist with monadic analyses<sup>2</sup> – with many more on top. Dyadic data can be directed or undirected (see the second section of this chapter), they exacerbate the problem of rare events (third section), allow for regressors from multiple levels of analyses (fourth section) and

introduce an additional form of potentially unobserved heterogeneity (fifth section), and the specification and analysis of spatial interdependence among dyads is much more complex than in monadic data (sixth section). We briefly discuss extensions and alternatives to dyadic designs in the seventh section but conclude in the final section that dyadic analyses are here to stay – and for good reason.

We employ two running examples, namely international trade and international war. Both phenomena cry out for, and have extensively been studied with, dyadic data. They differ, however, in the kinds of specification challenges and inferential problems posed: bilateral trade is continuously measured and not rare, whereas international war is often dichotomously measured (war vs no war or, increasingly, war onset vs no war onset) and is a phenomenon that is rare both in absolute and relative terms. This allows us to discuss problems and solutions separately for continuous and also binary categorical data whenever a problem or solution depends on the measurement scale of the dependent variable.

## DIRECTED VS UNDIRECTED DYADS

Dyads can be undirected or directed (Handcock and Gile, 2010; Neumayer and Plümper, 2010a). While in undirected dyads the two actors  $i$  and  $j$  (we also call them ‘nods’) are functionally identical to each other, such that  $y_{ij}$  is indistinguishable from  $y_{ji}$ , these actors are functionally different in directed dyads, such that  $y_{ij} \neq y_{ji}$ . Directed dyads distinguish between source and target, sender and receiver, proposer and acceptor, attacker and defender, exporter and importer and so on. For simplicity and terminological neutrality, for the most part we will say that directed dyad relations occur between source  $i$  and target  $j$ .

Directed dyads, in turn, fall into two categories: symmetric directed dyads, where each potential source is also a potential target, and

asymmetric directed dyads, where sources and targets are entirely separable so that each nod is exclusively either a source or a target. Examples of asymmetric directed dyads include employment relations between employers and employees, shop–customer relations and asymmetric relations between nations in the form of, for example, international aid from donors to recipients (Elkins et al., 2006; Neumayer and Plümper, 2010a). In asymmetric directed dyadic interactions,  $i$  can do something to  $j$  that  $j$  cannot or does not do to  $i$ .

Whether an analysis employs monadic or dyadic data and, if the latter, employs directed or undirected data often follows more from research tradition, data availability and convenience than from the underlying data generating process. For our two examples, it is possible to analyse why some countries trade more in total than others, analyse war at the monadic level and ask why country  $i$  is involved in a war. It is also possible, however, to analyse these phenomena with undirected dyadic data and ask why there is more total trade between some countries than other country pairs and why country  $i$  is at war with country  $j$  but not others. Finally, employing directed dyadic data allows one – in fact, forces one – to distinguish between exporter and importer and between attacker and defender.

The kind of data employed exerts a strong influence on the research question and on the inferences one can derive from the analysis. Yet, it should be the research question and the intended inferences one wishes to make that determine whether one employs monadic data, undirected dyadic data or directed dyadic data, and not vice versa. Crucially, the analysis of dyadic data supports the inclusion of regressors that vary at the ‘nod’ level (properties of the units that form the dyad), thus allowing inferences that at least resemble inferences from the analysis of monadic data. In contrast, the analysis of monadic data does not allow researchers to make inferences about the dyadic level.

Data generating processes that call for directed dyadic data are fairly common. For

example, in many country dyads, trade is somewhat unbalanced in that the volume of exports from  $i$  to  $j$  differs from the volume of exports from  $j$  to  $i$ . International wars are typically started by one side, though whether this means the side that formally started the conflict can be (entirely) blamed for the war is a different matter, as the debate around the German 'Alleinschuld' (sole guilt) for the First World War testifies.

As an example to illustrate how the choice of unit of analysis impacts upon findings and inferences, consider the so-called 'democratic peace' literature in particular (De Mesquita et al., 1999; Maoz and Russett, 1993). Though war and peace are both social interactions, for a surprisingly long period of time researchers studied conflict parties in isolation. The question why some countries are more likely than others to be involved in a war is important; but is it not even more important to know why country  $i$  is at war with country  $j$ , or why country  $i$  attacked country  $j$  (Gleditsch and Hegre, 1997; Rousseau et al., 1996)? Crucially, results and inferences based on different units of analysis also differ. At the monadic level, the democratic peace research asked whether democracies are more pacific than autocratic regimes and it generated rather mixed results (Benoit, 1996; Weede, 1984): some scholars found that democracies are more pacific; others did not find a statistically significant difference between democracies and autocracies. This variety of results disappears in dyadic data analysis. Democracies may not be less likely to fight wars, but they are significantly less likely to go to war with each other (Maoz and Abdolali, 1989; Maoz and Russett, 1992; Raknerud and Hegre, 1997). As another example, Buhaug et al. (2008) find that in a dyadic 'political centre-periphery' analysis, the discrimination of large ethnic groups increases the likelihood of civil war – a finding absent from the most highly cited monadic study of civil war, by Fearon and Laitin (2003).

## THE PROBLEM OF RARE EVENTS

Moving from monadic to undirected dyadic data and from undirected to directed dyadic data dramatically increases the number of observations but not the number of events or the aggregate of the analysed phenomena. With categorical outcomes at least, the analysis of dyadic data often, though not always, generates two related and at times confused problems: relatively and absolutely rare events. Events are relatively rare when potential outcomes – say, war and peace – are not equally likely: peace years are much more frequent than war years. Events are absolutely rare when the number of one potential outcome is low in absolute terms: for example, there simply are very few war years. Of course, the combination of relatively and absolutely rare events is also possible: international wars are both absolutely and relatively rare in one year and their onset is even rarer than their incidence, since many wars last more than one year. In an ideal world, outcomes would be neither rare in absolute nor relative terms. The worst outcome for quantitative analysts is the combination of an extremely low number of outcomes of a particular type with an extreme imbalance between outcomes, i.e., the combination of relative and absolute rarity.

With respect to our two examples, the problem of rare events is not so much an issue for international trade but it is for international war. If we extend the number of years, war years remain rare relative to peace years, but the absolute number of wars increases. This can help with conventional logit or probit analysis of dichotomous dependent variables, as Paul Allison (2012) explained in a blog entry:

[M]any researchers worry about whether they can legitimately use conventional logistic regression for data in which events are rare. [...] The problem is not specifically the rarity of events, but rather the possibility of a small number of cases on the rarer of the two outcomes. If you have a sample size of 1000 but only 20 events, you have a problem. If



you have a sample size of 10,000 with 200 events, you may be OK. If your sample has 100,000 cases with 2000 events, you're golden. There's nothing wrong with the logistic model in such cases. The problem is that maximum-likelihood estimation of the logistic model is well-known to suffer from small-sample bias.

Yet, there will be events that remain so rare that even after more years are added, conventional maximum-likelihood estimation like logit or probit are unlikely to be useful, since they will sharply underestimate the likelihood of these absolutely rare events to occur. To see how moving to the undirected dyadic or, worse still, to the directed dyadic level exacerbates the problem of rare events, let us take a closer look at international war. For 2003, Uppsala and PRIO's Armed Conflict Data coded two international war onsets with six involved countries. At the same time, roughly 200 independent countries existed. At the monadic level, international war was a rare event in the year 2003 with approximately 3% of units of analysis at war (6/200). Coding undirected dyads of war gives four bilateral war events, namely between Pakistan and India, Iraq and Australia, Iraq and the UK, and Iraq and the United States. Not only has the number of observations with war events actually decreased from six to four but the problem of rare events is massively exacerbated by the fact that 200 independent countries result in 19,900 undirected and 39,800 directed country dyads, thus generating approximately 0.02% of units of analysis at war in undirected dyads (4/19,900) and 0.01% of units at war in directed dyads (4/39,800). Granted, other years may see a slightly larger number of international war onsets, and increasing the number of years included in the sample will reduce the absolute rarity of the war-onset event but not its relative rarity.

In order to reduce relative rarity or even perfectly balance the outcomes, one can match peace-year observations to annual war-year observations based on observables. Alternatively, one can randomly draw a specified number of peace-year observations to at

least equal the number of war-year observations. This is indeed the ingenious method called rare-events logit, which King and Zeng (2001) have suggested for dealing with the problem of relatively rare events. In a nutshell, the procedure includes all rare events in the analysis and repeatedly draws at least an equal number and at most up to five times more observations with the frequent outcome – here, dyads not at war – to the sample and then corrects the estimated coefficients and standard errors to account for the sampling.<sup>3</sup> By drastically reducing the number of observations, rare-events logit has the additional advantage of researchers being able to focus their efforts on collecting more and better data for a much smaller sample. For events that are absolutely rare, Firth (1993) has proposed a technique called penalized maximum likelihood, which reduces small sample bias in logistic models.

## MULTILEVEL ANALYSIS

The analysis of dyadic data increases the 'levels' of analysis and therefore the kinds of regressors that can be included in the analysis. Even in relatively simple model specifications, three different types of factors may influence the relation between actors *i* and *j* that form the dyad *ij*, which is the unit of analysis:

- properties of *i*:  $x_i$
- properties of *j*:  $x_j$
- properties of *ij*:  $x_{ij}$

The standard workhorse assumption of simple additivity between these factors is not necessarily plausible in the analysis of dyadic data. Dyadic data are multilevel data with the two nodes and the dyad constituting separate levels. Potentially complex multiplicative causal relations between explanatory variables have been suggested by theorists.

Consider the example of the gravity model of international trade, which strongly relies on inverse distance and the gross domestic

product (GDP) of *i* and *j* to explain trade relations (Mátyás, 1997). The distance variable is often motivated and theoretically justified by Tobler's (1970) law, which states that 'everything is related to everything else, but near things are more related than distant things'.<sup>4</sup> Tobler's law is consistent with any monotonically increasing function. Alternatively, gravity models can also be motivated by Newton's law of gravity, which states that 'every particle attracts every other particle in the universe with a force which is directly proportional to the product of their masses and inversely proportional to the square of the distance between their centers'.<sup>5</sup> Of course, countries are not bodies, but the logic could apply if we accept the countries' GDP as the equivalent of 'mass'. This is indeed how the gravity model of international trade is typically specified (hence the name), following Tinbergen (1962), namely as:

$$y_{ij} = GDP_i \cdot GDP_j \cdot 1 / distance_{ij} \cdot \varepsilon_{ij}$$

Operationally, the gravity model is either estimated linearly after log-linearizing the above equation, which creates the issue of what to do about observations of zero bilateral trade, or with a Poisson pseudo maximum-likelihood estimator, following Silva and Tenreyro (2006), which has become standard practice now.

Similar multiplicative effects could also exist between the factors accounting for the properties of the dyad *ij* – distance in a gravity model of international trade – and properties of actors *i* and *j*, such that these factors condition each other. For example, the effect of distance on trade volumes depends on the production structure of *i* and *j*. The assumption that distance between two countries reduces trade is of course plausible and can be derived from virtually any model of international trade that allows for the existence of transaction and transportation costs. However, the effect of distance on trade is probably strongest for countries that produce goods that are both heavy and perishable and weakest for

countries that produce modern skill-intensive services that can be exported with the click of a button. Accordingly, the production structure of *i* and *j* enters the model explaining trade between *i* and *j* not just additively but also as a multiplicative term that conditions the effect of distance of trade.

In other contexts, the multiplicative effect may not be the appropriate specification and instead properties of *i* and *j* can have an additive effect so that  $y_{ij} = f(x_i + x_j)$ , a weakest link effect so that  $y_{ij} = f[\min(x_i, x_j)]$ , a strongest link effect so that  $y_{ij} = f[\max(x_i, x_j)]$  or combinations of these options. In international-conflict research, Oneal and Ray (1997) have argued for the existence of a 'weakest link' in the effect of democracy. They suggest that the lowest level of democracy in a dyad of countries is the decisive factor for the probability of international conflict. This hypothesis gives one nod in a dyad a strong influence and the other nod no influence on outcomes. Consistent with this hypothesis, they find that only the lower democracy score in a country dyad has a statistically significant effect, whereas the higher democracy score has an effect that is both small in size and statistically insignificant.

## UNOBSERVED DYADIC HETEROGENEITY

Regression models assume conditional homogeneity. After controlling for all factors included in the model, all remaining variation should be random. Of course, this assumption can hardly ever be satisfied in empirical research, regardless of whether the dependent variable is monadic or dyadic. Empirical models need to simplify the true data generating process, and simplification always leads to a violation of the conditional homogeneity assumption.

Yet, because dyadic data bring multiple levels into the analysis, the relevance of potentially unobserved heterogeneity increases: on

top of potentially unobserved factors at nodes *i* and *j* comes potentially unobserved dyad heterogeneity. For the example of international war, such unobserved dyad heterogeneity may result from numerous factors, including unmeasured animosities between countries in a dyad. As King (2001: 499) explains:

[S]uppose a degree of antipathy exists between pairs of countries, based on cultural, historical, or personal animosities, that has not been measured. (For example, completely accounting for problems between India and Pakistan by the usual list of annual dyadic variables we have measured seems unlikely.) The ‘historical animosity’ variable is (1) unmeasured, probably (2) causally prior to and (3) correlated with democracy, and (4) affects the probability of conflict—precisely the conditions for large omitted variable biases.

The combination of potentially unobserved heterogeneity at the level of either node *i* or node *j* or at the *ij* dyad level has naturally prompted many to call for the inclusion of fixed effects at the unit *i*, unit *j* and/or dyadic *ij* level.<sup>6</sup> For example, according to Wilson et al. (2005: 849), ‘a correct specification of the gravity model is parsimonious in specific economic variables’ and ‘rich in fixed effects’. In fact, because of the multilevel nature of dyadic analysis, the following most stringent fixed-effects specification is possible for a variable of interest that varies at the dyadic level over time, namely unit *i*-specific year fixed effects ( $\eta_{it}$ ), unit *j*-specific year fixed effects ( $\lambda_{jt}$ ), and dyad fixed effects ( $u_{ij}$ ), as in the following specification:

$$y_{ijt} = \alpha + \beta x_{ijt} + \eta_{it} + \lambda_{jt} + u_{ij} + \varepsilon_{ijt}$$

This specification obliterates the need to control for anything that is constant over time within dyads or that varies in any shape or form over time at the level of unit *i* or unit *j*.

However, what works well for a continuously measured variable such as international trade, which has great variation over time and is neither relatively nor absolutely rare (see, for example, Czaika and Neumayer, 2017), can represent the equivalent of ‘throwing

out the baby with the bath water’ (Beck and Katz, 2001) for dyadic data such as international war, which is dichotomous and both relatively and absolutely rare.

And yet, this is exactly what Green et al. (2001: 442) called for in a paper that attracted much attention at the time, based on the contention ‘that analyses of pooled cross-section data that make no allowance for fixed unobserved differences between dyads often produce biased results’. We agree with King’s (2001) assessment that Green et al. (2001) rightly draw attention to an important problem but that the proposed solution does not work for rare dyadic categorical data. Beck and Katz (2001: 490) offer Weibull duration models with gamma heterogeneity and the frailty Cox proportional hazards model that allows for random effects (frailty) as alternatives to the fixed-effects model, exploiting the fact that many dyadic data with binary outcomes resemble event history data. Collecting further data that can control for some dyadic heterogeneity is another option promoted by King (2001) – an option considerably aided if the relevant sample for the analysis of rare events can be drastically pruned by randomly drawing observations of the frequent event to all observations of the rare event (see the earlier section on ‘The Problem of Rare Events’). Last but not least, in Neumayer and Plümper (2017b), we propose a number of robustness tests that can be employed to test the extent to which potentially omitted variables, be they time-invariant or not, change the substantive findings.

A more radical approach to dealing with heterogeneity across dyads is to make dyads more homogeneous via the selection of relevant dyads or, equivalently, by dropping irrelevant dyads. The idea that sampling on the dependent variable – which is widely regarded as a sampling strategy that generates the most biased estimates – can be a solution to the problem of unobserved heterogeneity that ultimately results from the notion that fixed-effects models do not work well with a binary dependent variable and rare events

(Beck and Katz, 2001; King, 2001). The dyad fixed-effects model simply drops all dyads that never fought a war. The selection-of-relevant-dyads approach, on the other hand, distinguishes between those dyads that have never been at war that are relevant for the study of international war and those dyads who have never been at war that are irrelevant or ‘nearly irrelevant’ (Maoz and Russett, 1993: 627). Relevant dyads are those which have been never at war but could, whereas irrelevant dyads have an assumed a priori zero probability of going to war against each other.

Maoz and Russett (1993) have employed a very simple rule-based approach and assumed that country dyads are relevant for the study of international war only where the two countries are geographically contiguous or where at least one of the two countries is a major power. Lemke (1995) finds this rule to be neither necessary nor sufficient and accordingly redefines dyads relevant for the study of international war based on a complex relation between military strength, distance and terrain. In other words, a dyad is irrelevant if the two countries cannot overcome the obstacles of fighting a war due to limited military resources, a large distance between them and terrain that is not conducive to war. For example, two small and poor, landlocked countries on two different continents will not be able to go to war with each other. And indeed, Nepal and Chad never fought a war with each other. However, on the other hand, being landlocked has not always prevented countries from having a navy, which would give them at least some potential for a sea battle (Xiang, 2010).

Proponents of the selection-of-relevant-dyads approach argue that the inclusion of irrelevant dyads biases results and the exclusion of irrelevant dyads therefore improves the reliability of inferences. But does it in reality? Firstly, tens of thousands of ‘irrelevant’ dyads do not provide no information. At the very least, the very low to zero probability of war in these ‘irrelevant’ dyads demonstrates the importance of distance, terrain,

military capability and so on for the study of international war. Secondly, any rule-based selection will create measurement error and selection bias, as Lemke and Reed (2001) recognize. They suggest that even a selection rule as simple as contiguity and major power status produces only negligible bias due to these misspecification errors. Yet, we concur with Clark and Regan (2003) that it is better to explicitly model the opportunity to fight, which gives rise to a separate decision on willingness to fight, conditional on opportunity being present. Ultimately, the opportunity to fight is not fixed, and *probabilistic* modelling of the selection into the relevant sample of country dyads that can then be willing or unwilling to go to war trumps a *deterministic* rule-based selection. Only probabilistic modelling correctly accounts for the fact that there are no strict necessary or sufficient conditions for dyads to become relevant.

## SPATIAL DEPENDENCE OF DYADS

At least implicitly, the vast majority of analyses of dyadic data assume that observations are independent of each other. This is a problematic assumption even in analyses of monadic data, where actors learn from each other, where one actor coerces another or where the action of one actor triggers externalities for other actors, thereby influencing the behaviour of others (Franzese and Hays, 2008; Neumayer and Plümper, 2012). Yet, while independence of units is perhaps unlikely but certainly possible in monadic data, observations in dyadic data – almost by definition – cannot be independent of each other unless each node occurs only in one dyad. Thus, analyses of dyadic data dramatically increase the need to take seriously the dependence structures among observations.

The fact that dyadic data ‘have complex dependence structures’ (King, 2001: 498) has triggered three approaches for dealing with this. Firstly, some have explored how standard

errors can be adjusted to account for the dyadic and necessarily interdependent data structure that generates correlation across dyads in the error term (Aronow et al., 2015; Cameron et al., 2011; Cameron and Miller, 2014; Fafchamps and Gubert, 2007). This approach has yet to penetrate the mainstream and make its way into standard statistical software such as Stata. Secondly, some reject the dyadic approach altogether, arguing it is inconsistent with spatial dependence among dyads and suggesting network analysis instead. We deal with this fundamental objection to the use of dyadic data in the next section. Here we focus on the third approach, which is to model the spatial dependence in the dependent variable, in the explanatory variables or error term or combinations thereof.

At the very least, two dyads have to be partly dependent on each other if one actor is included in more than one dyad. In the language we use here, dyad  $ij$  cannot be entirely independent from dyads  $im$  and  $kj$ . To give an example, trade between the United States and the EU influences trade between the United States and China in multiple ways. In a static perspective, it is obvious that a product a US exporter sells to the EU cannot be sold to China at the same time. In a dynamic perspective, at least in sectors with economies of scale, the exports of a US company to European customers reduce the production cost per item for the US company and thus potentially increase competitiveness of US companies on the Chinese market. Whatever the perspective is, trade between the United States and China is unlikely to be independent from trade between the United States and the EU. In directed dyads, there also could be dependence between dyads  $ij$  and  $ji$ . For example, exports from the United States to China should not be entirely independent of exports from China to the United States. Clearly, as these countries enter into a 'trade war', trade in both directions will be affected.

Directed dyadic data offer a particularly rich set of modelling spatial dependence due to the distinction between sources and

targets (see Neumayer and Plümper, 2010a). As explained above, the most obvious source of spatial dependence in dyads results from the lack of independence between dyads that have one actor or nod, either source  $i$  or target  $j$ , in common. Directed dyad  $ij$  can therefore be influenced by what other sources  $k$  do with respect to the *specific* target  $j$  only, which results in what we have dubbed specific-source contagion:

$$y_{ij} = \rho \sum_{k \neq i} w_{pq} y_{kj} + \varepsilon_{ij}$$

or by what other targets  $m$  do with respect to the *specific* source  $i$  only (specific target contagion):

$$y_{ij} = \rho \sum_{m \neq j} w_{pq} y_{im} + \varepsilon_{ij}$$

Alternatively, directed dyad  $ij$  can be influenced by what other sources  $k$  do with respect to *any* other target  $m$ , not just  $j$ , which results in what we have dubbed aggregate-source contagion:

$$y_{ij} = \rho \sum_{k \neq i} \sum_m w_{pq} y_{km} + \varepsilon_{ij}$$

or by what other targets  $m$  do with respect to *any* source  $k$  and not just  $i$ , which results in aggregate-target contagion:

$$y_{ij} = \rho \sum_k \sum_{m \neq j} w_{pq} y_{km} + \varepsilon_{ij}$$

Finally, dyad  $ij$  can be influenced by what other dyads  $km$  do, which we have dubbed directed dyad contagion:

$$y_{ij} = \rho \sum_{\substack{k,m, \\ ij \neq km}} \omega y_{km} + \varepsilon_{ij}$$

Undirected dyadic data, by contrast, only allow the dyad  $ij$  ( $=ji$ ) to be influenced by other dyads  $km$ , which may either include nod  $i$  or nod  $j$  in them (i.e., including  $im$  and  $kj$ ) or exclude them ( $k,m \neq i,j$ ), resulting in inclusive or exclusive undirected dyad contagion, respectively.

Add to this richness in specification of contagion channels the fact that the link function  $w_{pq}$  that links dyad  $ij$  to other dyads can be modelled as any link between  $i, j, k$  and  $m$  and any combination of these links, it becomes clear that spatial dependence in dyadic data offers incredible modelling flexibility, which can be daunting to those who come with little theory to an empirical research question. Analysing spatial dependence in dyadic data therefore requires more prior theorizing than the equivalent analysis in monadic data.

There is of course no reason why only one of these dependencies should be relevant for any given data generating process. All these dependencies can be simultaneously relevant. However, if researchers want to estimate the spatial effects of different types of dyadic spatial dependence simultaneously, they are likely to run into the problem that the spatial-effect variables are highly correlated with each other. If so, we recommend including only the one or ones that the theory predicts to be the most relevant. In many applications, the spatial effect of dyads that include either  $i$  or  $j$  will be more relevant than the spatial effect of dyads that do not include one of the nodes.

If, as we have argued above, the analysis of dyadic data should typically include an analysis of spatial dependence – or at least make allowance for it – why then is it that such analyses are even rarer at the dyadic than at the monadic level? One reason is that many of the ‘canned’ estimators will not work with dyadic data, not least because it is typically computationally impossible with standard computers to create the 4-adic dataset required for the weighting matrix to link dyads with dyads, for data with a large number of dyads. We have made Stata ado-files available that allow researchers to generate spatial-effect variables for dyadic data, by parsing through each row of a virtual 4-adic dataset (Neumayer and Plümper, 2010b). From the standpoint of an econometrician purist, the disadvantage of our approach is that it does not allow for spatial maximum-likelihood estimation to account for the endogeneity that comes from spatial

feedback loops among dyads. However, it does allow for spatial-2SLS to account for this specific kind of endogeneity, a technique that is only marginally less efficient than spatial maximum likelihood. For example, for a spatial autoregressive model with specific-source contagion, one would use the following variables as instruments for  $\sum_{k \neq i} w_{pq} y_{kj}$ , the spatially lagged variable:  $\sum_{k \neq i} w_{pq} X_{kj}$  as well as, in so far as not already included in the estimation model,  $X_k$  and  $X_j$ . Naturally, the identifying assumptions are that the spatial dependence is not additionally of the spatial- $x$  type (implying spatial dependence in  $X_{kj}$ ), since otherwise  $\sum_{k \neq i} w_{pq} X_{kj}$  cannot serve as an instrument, and that  $X_k$  and  $X_j$  are not themselves endogenous for any reason and only affect the outcome variable through  $\sum_{k \neq i} w_{pq} y_{kj}$ , which is why they were not already included in the estimation model itself. Note, however, that if these identifying assumptions do not hold, then the autoregressive spatial estimation model is misspecified and estimating this model with spatial maximum likelihood will also result in biased estimates. In other words, as always, model misspecification results in biased estimates and spatial maximum likelihood does not provide a miraculous cure against it. The computationally much easier spatial-2SLS therefore dominates spatial maximum likelihood for models containing dyadic spatial dependence unless the number of dyads  $ij$  and the number of time periods  $T$  are small.

## EXTENSIONS AND ALTERNATIVES

As mentioned in the previous section, dyadic data analysis has its discontents. Cranmer and Desmarais (2016: 361) are probably those most explicit in their rejection of dyadic research designs: ‘We believe (...) that dyadic design is almost never appropriate for IR theory and data’. They base this conclusion principally on the claim that ‘a fundamental assumption of dyadic analysis

is that dyads exist in isolation from anything else going on in the world' (2016: 357). We hope to have shown in the previous section that this claim is simply untenable, since it is very much possible to model dyadic interaction without losing the spatial dependencies between dyads out of sight.

That is not to say, however, that dyadic designs are always superior to other modelling options. One challenge to such designs stems from events that are truly multilateral rather than bilateral in nature, i.e., when multiple actors jointly act simultaneously, such as in entering the world wars or when multilateral treaties are created. In principle, this can still be modelled in a dyadic design with spatial interdependence. However, Poast (2016: 371) is right to say that, in principle, a multilateral event represents a distinct phenomenon from interdependence among dyads, and so a *k*-adic design, where *k* stands for the number of actors that act jointly, can prove better in certain circumstances (see Fordham and Poast, 2016; Poast, 2010). Yet, *k*-adic designs – a *k*-dimensional extension to dyadic designs – have their own problems, creating a gigantic number of observations unless the sample can be heavily pruned by employing a strict rule for cutting out 'irrelevant' *k*-ads and/or random sampling of non-event observations to the rare event observations. Moreover, spatial dependence among *k*-ads is even more difficult to model than among dyads.

Cranmer and Desmarais (2011) prefer exponential random graph models and other models from inferential network analysis as alternatives to dyadic designs. Again, we acknowledge that network analysis can offer an interesting modelling choice (see Ward et al., 2011 for a general overview). And yet, we are not convinced of the superiority of these alternative research designs. Even if dyads are not the closest approximation to the data generating process, they can nevertheless represent the most useful research design, which should be the dominant decision-making criterion (Diehl and Wright, 2016). After all, what Cranmer and Desmarais (2016: 356) seem to regard as a

weakness, namely that 'the dyad constitutes the simplest level of measurement at which we can record a relationship', can actually be regarded as its main strength, since social-science analysis depends on simplification in order to make sense of a causally complex world (Clarke and Primo, 2012; Neumayer and Plümper, 2017b). We agree with Diehl and Wright (2016: 367), who come to the conclusion that dyads 'are often *appropriate* and *useful* units of analysis, contingent upon the questions that researchers ask and the outcomes they wish to explain'.

It is also possible to bring the dyadic model with spatial dependence closer to traditional network analysis.<sup>7</sup> This can be done by including two additional factors in the analysis: systemic nod and systemic dyad information. While systemic nod information describes the position of both *i* and *j* in the sample of nodes – their network centrality or the total number of their interactions to mention just two examples – systemic dyad information measures the systemic position of a dyad in a population of dyads. For example, the dyad United States–Russia traditionally obtains a central network position in international relations. The relevance of the relations between the United States and Russia to other dyads is not necessarily accounted for by systemic nod information available for these countries individually. Systemic nod and systemic dyad information can be measured and included in the dyadic analysis as an independent explanatory variable or, alternatively, such information can be used in the weighting matrix for the spatial-effect variables. Thus, one can, for example, test whether spatial dependence among dyads depends on the centrality of dyads – a plausible assumption consistent with many theories of social networks.

## CONCLUSION

Undoubtedly, dyadic data challenge analysts. Specifying models for dyadic data analysis

sufficiently well is vastly more complicated than doing so for models for monadic-data analysis. However, some of these complications – the choice between directed and undirected dyads, the interactions between different levels of analysis, the mutual dependencies between dyads – actually also present new opportunities that often allow researchers to gain novel insights. No doubt, getting the specification of a dyadic model sufficiently right such that inferences are reliable is difficult. However, there is much to learn from analysing dyadic data that cannot be learned from the analysis of monadic data.

The good news is that over the last two decades, the most severe specification problems have been solved or at least reduced. Today, social scientists know – at least in principle – how to deal with the lack of independence of dyads and they know how to address the problems associated with rare events and the multiple heterogeneity issues. The methodological frontier of dyadic data analysis has therefore shifted and moved on. The new debates focus on utilizing spatial analyses of dyadic data for network analysis and whether social scientists should progress to higher dimensions in the form of k-adic data analysis. But dyadic data analysis has been and, we predict, will remain the design of choice for the analysis of relationships between actors in political science, international relations and beyond.

## Notes

- 1 In political science, for each article with 'dyad' or 'dyadic' as a keyword in the Social Sciences Citation Index, roughly 250 articles which do not, at least not explicitly, state to be concerned with dyadic data are published. Aronow et al. (2015: 564) have counted 62 dyadic analyses over a five-year period published in the *American Political Science Review*, *American Journal of Political Science* and *International Organization*. That amounts to an average of merely four per journal per year.
- 2 Common model misspecifications that cause bias include selection, concept invalidity, measurement error, omitted variables, functional form misspecification, causal heterogeneity and context conditionality, structural change, dynamic

misspecification and lack of independence of observations (Neumayer and Plümpert, 2017b).

- 3 The procedure is also known as choice-based or case-control sampling.
- 4 Tobler's (1970) law does not only motivate the gravity model but also spatial models (Franzese and Hays, 2007; Neumayer and Plümpert, 2017a; Plümpert and Neumayer, 2010). Hence, it can be applied to monadic as well as dyadic phenomena.
- 5 [https://en.wikipedia.org/wiki/Newton%27s\\_law\\_of\\_universal\\_gravitation](https://en.wikipedia.org/wiki/Newton%27s_law_of_universal_gravitation).
- 6 Though we do not wish to underestimate the relevance of unobserved heterogeneity, we need to stress that the fixed-effects approach makes the very strong and largely unjustified assumption that all heterogeneity between dyads and thus all model misspecification is strictly time-invariant with constant effects. Once this assumption is removed, differences between the pooled-OLS estimates and fixed-effects estimates could just result from misspecified dynamics, omitted time-varying variables, unexplained trends in the data, misspecified lag structures or similar (Plümpert and Troeger, 2019). The test that Green and co-authors (2001) propose does not work any better than the Hausman test, which rests on the assumption that the fixed-effects model is unbiased: 'The Hausman test is not a reliable tool for identifying bias in typically sized samples; nor does it aid in evaluating the balance of bias and variance implied by the two modelling approaches' (Clark and Linzer, 2015).
- 7 See Hays et al. (2010) for a proposal on how network behaviour can be incorporated into a spatial model at the monadic level, which is straightforward to generalize to the dyadic level.

## REFERENCES

- Allison, P., 2012. Logistic regression for rare events. *Statistical Horizons*, Blog, <https://statisticalhorizons.com/logistic-regression-for-rare-events> (Accessed on 20 November, 2019).
- Aronow, P.M., Samii, C. and Assenova, V.A., 2015. Cluster-robust variance estimation for dyadic data. *Political Analysis*, 23(4), pp. 564–577.
- Beck, N. and Katz, J.N., 2001. Throwing out the baby with the bath water: A comment on Green, Kim, and Yoon. *International Organization*, 55(2), pp. 487–495.
- Benoit, K., 1996. Democracies really are more pacific (in general): Reexamining regime type



- and war involvement. *Journal of Conflict Resolution*, 40(4), pp. 636–657.
- Buhaug, H., Cederman, L.E. and Rød, J.K., 2008. Disaggregating ethno-nationalist civil wars: A dyadic test of exclusion theory. *International Organization*, 62(3), pp. 531–551.
- Cameron, A.C. and Miller, D.L., 2014. *Robust Inference for Dyadic Data*. Unpublished manuscript, University of California-Davis.
- Cameron, A.C., Gelbach, J.B. and Miller, D.L., 2011. Robust inference with multiway clustering. *Journal of Business & Economic Statistics*, 29(2), pp. 238–249.
- Clark, D.H. and Regan, P.M. 2003. Opportunities to fight: A statistical technique for modeling unobservable phenomena. *Journal of Conflict Resolution*, 47(1), pp. 94–115.
- Clarke, K.A. and Primo, D.M., 2012. *A Model Discipline: Political Science and the Logic of Representations*. Oxford: Oxford University Press.
- Clark, T.S. and Linzer, D.A., 2015. Should I use fixed or random effects?. *Political Science Research and Methods*, 3(2), pp. 399–408.
- Cranmer, S.J. and Desmarais, B.A., 2011. Inferential network analysis with exponential random graph models. *Political Analysis*, 19(1), pp. 66–86.
- Cranmer, S.J. and Desmarais, B.A., 2016. A critique of dyadic design. *International Studies Quarterly*, 60(2), pp. 355–362.
- Czaika, M. and Neumayer, E., 2017. Visa restrictions and economic globalisation. *Applied Geography*, 84, pp. 75–82.
- De Mesquita, B.B., Morrow, J.D., Siverson, R.M. and Smith, A., 1999. An institutional explanation of the democratic peace. *American Political Science Review*, 93(4), pp. 791–807.
- Diehl, P.F. and Wright, T.M., 2016. A conditional defense of the dyadic approach. *International Studies Quarterly*, 60(2), pp. 363–368.
- Elkins, Z., Guzman, A.T. and Simmons, B.A. 2006. Competing for capital: The diffusion of bilateral investment treaties, 1960–2000. *International Organization*, 60(4), pp. 811–846.
- Fafchamps, M. and Gubert, F., 2007. The formation of risk sharing networks. *Journal of Development Economics*, 83(2), pp. 326–350.
- Fearon, J.D. and Laitin, D.D., 2003. Ethnicity, insurgency, and civil war. *American political science review*, 97(1), pp. 75–90.
- Firth, D., 1993. Bias reduction of maximum likelihood estimates. *Biometrika*, 80(1), pp. 27–38.
- Fordham, B. and Poast, P., 2016. All alliances are multilateral: Rethinking alliance formation. *Journal of Conflict Resolution*, 60(5), pp. 840–865.
- Franzese, R.J., Jr. and Hays, J.C., 2007. Spatial econometric models of cross-sectional interdependence in political science panel and time-series-cross-section data. *Political Analysis*, 15(2), pp. 140–164.
- Franzese, R.J., Jr. and Hays, J.C., 2008. Interdependence in comparative politics: Substance, theory, empirics, substance. *Comparative Political Studies*, 41(4–5), pp. 742–780.
- Gleditsch, N.P. and Hegre, H., 1997. Peace and democracy: Three levels of analysis. *Journal of Conflict Resolution*, 41(2), pp. 283–310.
- Green, D.P., Kim, S.Y. and Yoon, D.H., 2001. Dirty pool. *International Organization*, 55(2), pp. 441–468.
- Handcock, M.S. and Gile, K.J., 2010. Modeling social networks from sampled data. *The Annals of Applied Statistics*, 4(1), pp. 5–25.
- Hays, J.C., Kachi, A. and Franzese Jr, R.J., 2010. A spatial model incorporating dynamic, endogenous network interdependence: A political science application. *Statistical Methodology*, 7(3), pp. 406–428.
- King, G., 2001. Proper nouns and methodological propriety: Pooling dyads in international relations data. *International Organization*, 55(2), pp. 497–507.
- King, G. and Zeng, L., 2001. Explaining rare events in international relations. *International Organization*, 55(3), pp. 693–715.
- Kinne, B.J., 2012. Multilateral trade and militarized conflict: Centrality, openness, and asymmetry in the global trade network. *The Journal of Politics*, 74(1), pp. 308–322.
- Krasikova, D.V. and LeBreton, J.M., 2012. Just the two of us: Misalignment of theory and methods in examining dyadic phenomena. *Journal of Applied Psychology*, 97(4), pp. 739–757.
- Lemke, D. 1995. The tyranny of distance: Redefining relevant dyads. *International Interactions*, 21(1), pp. 23–38.
- Lemke, D. and Reed, W., 2001. The relevance of politically relevant dyads. *Journal of Conflict Resolution*, 45(1), pp. 126–144.
- Maoz, Z. and Abdolali, N., 1989. Regime types and international conflict, 1816–1976. *Journal of Conflict Resolution*, 33(1), pp. 3–35.

- Maoz, Z. and Russett, B., 1992. Alliance, contiguity, wealth, and political stability: Is the lack of conflict among democracies a statistical artifact?. *International Interactions*, 17(3), pp. 245–267.
- Maoz, Z. and Russett, B., 1993. Normative and structural causes of democratic peace, 1946–1986. *American Political Science Review*, 87(3), pp. 624–638.
- Mátyás, L., 1997. Proper econometric specification of the gravity model. *World Economy*, 20(3), pp. 363–368.
- Neumayer, E. and Plümper, T., 2010a. Spatial effects in dyadic data. *International Organization*, 64(1), pp. 145–166.
- Neumayer, E. and Plümper, T., 2010b. Making spatial analysis operational: Commands for generating spatial-effect variables in monadic and dyadic data. *Stata Journal*, 10(4), pp. 585–605.
- Neumayer, E. and Plümper, T., 2012. Conditional spatial policy dependence: Theory and model specification. *Comparative Political Studies*, 45(7), pp. 819–849.
- Neumayer, E. and Plümper, T., 2017a. *W. Political Science Research and Methods*, 4(1), pp. 175–193
- Neumayer, E. and Plümper, T., 2017b. *Robustness Tests for Quantitative Research*. Cambridge: Cambridge University Press.
- Oneal, J.R. and Ray, J.L., 1997. New Tests of the Democratic Peace: Controlling for Economic Interdependence, 1950–85. *Political Research Quarterly*, 50(4), pp. 751–775.
- Plümper, T. and Neumayer, E., 2010. Model specification in the analysis of spatial dependence. *European Journal of Political Research*, 49(3), pp. 418–442.
- Plümper, T. and Troeger V.E. 2019. Not so harmless after all: The fixed-effects model. *Political Analysis*, 27(1), pp. 21–45.
- Poast, P., 2010. (Mis)Using dyadic data to analyze multilateral events. *Political Analysis*, 18(4), pp. 403–425.
- Poast, P., 2016. Dyads are dead, long live dyads! The limits of dyadic designs in international relations research. *International Studies Quarterly*, 60(2), pp. 369–374.
- Raknerud, A. and Hegre, H., 1997. The hazard of war: Reassessing the evidence for the democratic peace. *Journal of Peace Research*, 34(4), pp. 385–404.
- Rousseau, D.L., Gelpi, C., Reiter, D. and Huth, P.K., 1996. Assessing the dyadic nature of the democratic peace, 1918–88. *American Political Science Review*, 90(3), pp. 512–533.
- Silva, J.S. and Tenreyro, S. 2006. The log of gravity. *The Review of Economics and Statistics*, 88(4), pp. 641–658.
- Tinbergen, J., 1962. *Shaping the World Economy. Suggestions for an International Economic Policy*. New York: Twentieth Century Fund.
- Tobler, W.R., 1970. A computer movie simulating urban growth in the Detroit region. *Economic Geography*, 46(sup1), pp. 234–240.
- Ward, M.D., Stovel, K. and Sacks, A., 2011. Network analysis and political science. *Annual Review of Political Science*, 14, pp. 245–264.
- Weede, E., 1984. Democracy and war involvement. *Journal of Conflict Resolution*, 28(4), pp. 649–664.
- Wilson, J.S., Mann C.L. and Otsuki, T., 2005. Assessing the benefits of trade facilitation: A global perspective. *World Economy*, 28(6), pp. 841–871.
- Xiang, J., 2010. Relevance as a latent variable in dyadic analysis of conflict. *The Journal of Politics*, 72(2), pp. 484–498.



# Model Specification and Spatial Interdependence\*

Scott J. Cook, Jude C. Hays, and Robert Franzese

## INTRODUCTION

Researchers now regularly estimate spatial models in applied political science, both to enhance the validity of their direct (i.e., non-spatial) covariate-effect estimates and to test explicitly spatial theories. While this is a welcome advance over past practices, we worry that much of this first generation of applied spatial research overlooks certain aspects of spatial models. In particular, while different theories imply different spatial-model specifications, statistical tests frequently have power against incorrect alternatives. As a consequence, researchers who fail to discriminate explicitly between the different manifestations of spatial association in their outcomes are likely to erroneously find support for their theoretically preferred spatial process (e.g., contagion or endogenous global spillovers) even where an alternative process instead underlies the association (e.g., diffusion or exogenous local spillovers). To help researchers avoid

these pitfalls, we (1) elaborate the alternative theoretical processes that give rise to a taxonomy of spatial models, (2) indicate why and provide evidence that these alternative processes are frequently mistaken for one another during conventional hypothesis testing, and (3) suggest a set of strategies for effectively discriminating between the seven alternative spatial-lag models (with one, two, or all three of spatially lagged errors, spatially lagged independent variables, and/or spatially lagged dependent variable).

Cross-sectional, or spatial, interdependence is ubiquitous in the social sciences. Theories indicating that the actions of/outcomes in some units are a function of (i.e., depend upon) those of other units – as they are coerced by, compete with, learn from, and emulate one another – span across the sub-fields and substance of political science, for example.<sup>1</sup> The diffusion of political institutions and policy is well established in American and comparative politics, with units learning from and/or emulating the institutions and instruments

of other units. Similarly, political behavior, from voting to violence, is necessarily interdependent as expectations over outcomes are a function of beliefs about the actions of others. The very structure of the global economy indicates the importance of interdependence in the study of comparative and international political economy, evidenced both in deepening economic integration and more prevalent policy coordination or competition. The very field name International Relations, meanwhile, centrally implicates interdependence in that area of study. More generally still, spatial interdependence is present whenever units are affected by the actions, behaviors, and outcomes of other units.

Given the theoretic centrality of spatial interdependence in political science and international relations, early work sought to introduce and extend methods for analyzing this dependence directly (Beck et al., 2006; Franzese and Hays, 2007). Beyond the classic linear model, statistical methods have been developed for spatial analysis of binary outcomes (Franzese et al., 2016; Wilhelm and de Matos, 2013), count data (Hays and Franzese, 2017), durations (Hays and Kachi, 2009; Hays et al., 2015), and endogenous predictors (Betz et al., 2020). Moreover, researchers have built on the dictum that space is ‘more than geography’ and indicated how the specification of the connectivity matrix itself enables researchers to test a range of political theories (Neumayer and Plümper, 2016; Plümper and Neumayer, 2010). As a result, there has been a proliferation of empirical work in political science, which offers theories, estimates models, and conducts tests of spatial interdependence.<sup>2</sup>

While this is a welcome advance over past practices – treating spatial dependence as a nuisance or ignoring it altogether – we worry that much of this first generation of applied spatial research does not fully appreciate or is unfamiliar with certain aspects of spatial models. Importantly, distinct spatial-model specifications arise from different theoretical explanations of spatial clustering

in the outcomes: *i*) endogenous interaction effects (e.g., spillovers in the outcomes), *ii*) exogenous interaction effects (e.g., spillovers in the predictors), and/or *iii*) interactions or clustering in the residuals (Elhorst, 2010).<sup>3</sup> Problematically, these theoretically distinct statistical models are quite similar and so produce similar patterns in empirical data, which complicates specification testing (Anselin, 2001; Gibbons and Overman, 2012). Specifically, diagnostic tests have power against incorrect alternatives (testing rejects A in favor of B, when, in fact, C is present and causes the rejection, not B), making it difficult to statistically distinguish between these various models. To the extent that researchers attach theoretic importance to these different model specifications, which they should, and subsequently draw substantively meaningful inferences off these diagnostic tests, it is important to understand how and the extent to which these tests can distinguish between these alternatives. Thus, while we can now estimate a variety of spatial models in many different contexts, these ambiguities, left unaddressed, limit what we can learn from analyses utilizing spatial methods.

To begin to redress these limitations here, we first detail and describe the possible sources of spatial clustering and the econometric models that are implied when any combination of these sources is present. While a general model that allows for all three sources of spatial clustering is discussed, we show that this model is weakly identified based on structural assumptions and therefore can provide only a precarious guide to our specification search. This precludes a Hendry-like general-to-specific specification search, as has been advocated in time-series modeling (in political science by De Boef and Keele, 2008). Instead, researchers generally must constrain one of the possible sources of spatial clustering in order to discriminate effectively between the remaining alternatives. While research design or theory should be the preferred bases on which to justify this constraint, we offer guidance for

researchers in situations where these solutions are not available.

Our intention is not to discourage the use of spatial methods, as we feel spatial analysis is necessary and appropriate whenever one has cross-sectional or time-series-cross-sectional observational data.<sup>4</sup> Instead, we simply advocate that researchers exercise greater caution when estimating these models, especially when attempting to articulate and test specific theories of spatial interdependence. Taking ‘space’ seriously does not simply mean estimating a spatial model but rather estimating the *appropriate* spatial model. In the following section, we outline the variety of alternative spatial models, show how easy it is to mistake one of these models for another when drawing inferences, and suggest tests to aid researchers in identifying and specifying appropriate models for estimation. Subsequently, we evaluate the small-sample performance of these tests under a variety of simulated conditions.

## SPECIFYING SPATIAL MODELS

In prior work, we have highlighted the substantive/theoretical ubiquity of interdependence across political science. While the emergence of applied spatial research in political science suggests broad agreement on the importance of spatial theories, some research may have too quickly turned to articulating and testing specific mechanisms (e.g., emulation vs learning) and sources (e.g., distance vs trade) for spatial dependence across a range of issue areas without first devoting sufficient attention to the various broader ways in which spatial dependence can manifest in observational data. Before discriminating between competing theories of the bases of diffusion, researchers must first evidence that there is some form of diffusion. Researchers need to be aware of the various possible sources of spatial correlation in their outcomes and adopt models

that appropriately nest and test between these competing alternatives. Therefore, we open by discussing the potential sources of spatially correlated outcomes, before outlining the spatial-econometric models implied by each.<sup>5</sup>

According to Anselin (2010), spatial heterogeneity is the uneven distribution of a trait, event, or relationship across a region. Therefore, it is present whenever we observe spatial clustering in the outcomes across some set of sample units. By which we mean that when there is non-zero covariance among these units’ outcomes:

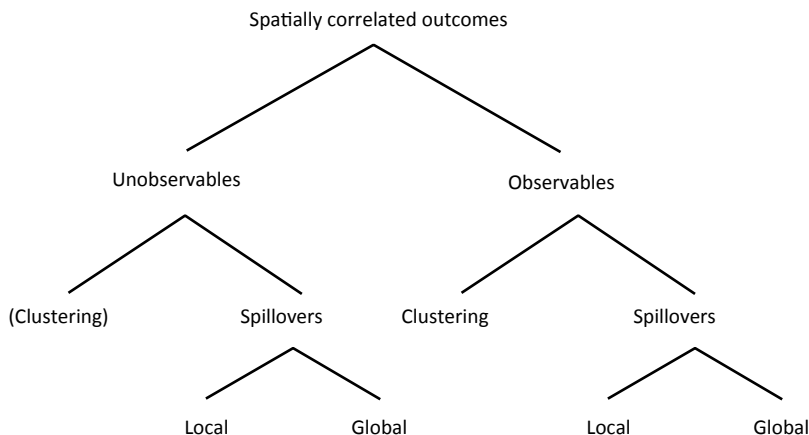
$$\begin{aligned} \text{cov}(y_i, y_j) &= E(y_i y_j) \\ &- E(y_i) \times E(y_j) \neq 0 \text{ for } i \neq j \end{aligned} \quad (1)$$

i.e., whenever variation in the outcome is not randomly distributed across units. This only becomes problematic for non-spatial analyses, however, when the (spatial) distribution of these outcomes is not entirely explained by the (spatial) distribution of predictors. In these instances, additional unmodeled factors give rise to the spatial correlation we observe in our outcomes, the failure to account for which potentially threatens the accuracy of our estimates and the validity of our inferences.

To elaborate the various manifestations of spatial association more fully, consider Figure 39.1. As we see, correlation in the outcomes arises from spatial (inter)dependence in the observable and/or unobservable inputs.<sup>6</sup> Broadly, there are two mechanisms that produce spatially correlated outcomes: *i*) spatial clustering and/or *ii*) spatial spillovers or interactions. As with the outcomes, spatial clustering in the observables (unobservables) occurs when the level, presence, or change of an observed (unobserved) determinant in one unit is correlated with but not a function of (not caused by) the value of that factor in other (spatially proximate) units:

$$y_i = f(x_i, \varepsilon_i) \text{ and } \text{cov}(x_i, x_j) \neq 0 \text{ for } i \neq j \quad (2a)$$

$$y_i = f(x_i, \varepsilon_i) \text{ and } \text{cov}(\varepsilon_i, \varepsilon_j) \neq 0 \text{ for } i \neq j \quad (2b)$$



**Figure 39.1 Manifestations of spatial association**

where  $y$  is the outcome,  $x$  is a predictor, and  $\varepsilon$  is the unobserved error, with subscripts  $i$  and  $j$  identifying cross-sectional units. Here, the predictors and/or errors are spatially clustered, which, in turn, produces spatial clustering in the outcomes.<sup>7</sup> This does not require or suggest interaction between the units, simply that proximate actors possess similar characteristics (e.g., natural endowments that span across units) that, when manipulated, cause these unit outcomes to vary concurrently. That is, a common factor in the observables or unobservables results in correlated group effects.<sup>8</sup> For example, policy or technological innovations that change in the costs of inputs or demand (holding supply fixed) impact the revenues of all producers of a good, even where there is no direct interaction between them.

Alternatively, spatially correlated outcomes can arise due to spatial spillovers, when the outcomes of one unit are a function of (are caused by) the outcomes, actions, and behaviors of *other* units:

$$y_i = f(x_i, x_j, \varepsilon_i) \quad (3a)$$

$$y_i = f(x_i, \varepsilon_i, \varepsilon_j) \quad (3b)$$

$$y_i = f(x_i, y_j, \varepsilon_i) = f(x_i, (x_j, \varepsilon_j), \varepsilon_i) \quad (3c)$$

These we label *interdependence*, which seems to be the spatial process most commonly assumed by contemporary applied researchers. In this case, there are spillovers and/or externalities that arise from the observables (Equation 3a), unobservables (Equation 3b), or outcomes (Equation 3c) of other units. Note that here we need not assume that the observables or unobservables are governed by a spatial process (are spatially correlated) – although they certainly may be – merely that there is cross-unit dependence, where the outcome in  $i$  is a function of the observables and/or unobservables in unit  $j$ . Theories of diffusion or contagion, or of strategic decision-making, for example, would generally imply such interdependent processes.

While many of our theories suppose interdependence in the outcomes, this necessarily implies that the relation of  $y_i$  and  $y_j$  operates through the combined spatial effects of the observables ( $x_j$ ) and unobservables ( $\varepsilon_j$ ) (Equation 3c). Anselin (2003) discusses that for specification, then, a more fundamental consideration is whether these externalities are global or local (the third dimension of Figure 39.1), i.e., whether actors only affect their immediate neighbors, peers, etc., as assumed by a local process, or, as in Tobler’s oft used expression ‘everything is related to everything’, suggesting a global process in

which actors affect proximate actors, who in turn affect their proximate actors, and so on. Perhaps more clearly, the distinction is between whether spillovers in the observables ( $\mathbf{X}$ ) and unobservables ( $\varepsilon$ ) in my neighbors affect me *directly*, or they affect me *indirectly* through my neighbors' outcomes ( $y_i$ ).<sup>9</sup>

Our theoretical propositions about which combination of these spatial effects produces spatial clustering in the outcomes imply different econometric specifications. Specifically, we have discussed three relevant dimensions which should inform spatial specification: *i*) whether spatial heterogeneity in the outcome is caused by observable or unobservable factors (or both), *ii*) whether these spatial effects arise from clustering or spillovers (or both), and *iii*) if spillovers, whether these spillovers are local or global.<sup>10</sup> Table 39.1 lists the spatial models most commonly discussed in the literature.<sup>11</sup>

Beginning with the most restrictive of these models, the non-spatial linear-regression model assumes that any spatial correlation in the outcomes is entirely a function of spatial correlation in the predictors:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (4)$$

That is, to account for the spatial correlation in outcomes, we need simply to include appropriate predictors ( $\mathbf{X}$ ), as regularly done in non-spatial analysis. We emphasize this simple point as it seems to be misunderstood in some applied literature.<sup>12</sup> Moreover, it underscores the importance of model specification more

generally when undertaking spatial analysis, as misspecified models – those omitting relevant spatially clustered predictors – will exhibit spatial dependence in the residuals (and, in turn, give power to spatially lagged (in)dependent variables). As such, a better specified model is one obvious solution when confronting spatially clustered residuals.<sup>13</sup>

In estimating these models, researchers assume a spherical error variance–covariance matrix (and, by extension, that  $\rho = \lambda = 0$ ;  $\boldsymbol{\theta} = \mathbf{0}$ ), i.e., that the residuals are not spatially correlated. This can be easily tested through a variety of post-estimation diagnostic tests, including the familiar Moran's I and Lagrange Multiplier tests (Franzese and Hays, 2008). Should these tests reject the null, indicating spatial correlation in the residuals, further remedies are needed to avoid inefficiency and possible bias in our parameter estimates. Most applied spatial work in political science engages in this type of exploratory spatial analysis to justify the use of further spatial methods. However, these tests merely suggest *a* spatial process and generally are not very helpful for making specification choices from among the broad class of possible spatial models.

Of these models, the most widely discussed have been the spatial error model (**SEM**), the spatial lag model (**SAR**), and, more recently, the spatially lagged  $\mathbf{X}$  model (**SLX**). Each assumes that any spatial correlation in the outcomes arises from a single source, exogenous observables, unobservables, or

**Table 39.1 Common spatial econometric models**

Name	Structural model	Restrictions
General nesting model	$\mathbf{y} = \rho\mathbf{W}\mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\mathbf{X}\boldsymbol{\theta} + \mathbf{u}, \mathbf{u} = \lambda\mathbf{W}\mathbf{u} + \boldsymbol{\varepsilon}$	None
Spatial Durbin error model	$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\mathbf{X}\boldsymbol{\theta} + \mathbf{u}, \mathbf{u} = \lambda\mathbf{W}\mathbf{u} + \boldsymbol{\varepsilon}$	$\rho = 0$
Spatial autocorrelation model	$\mathbf{y} = \rho\mathbf{W}\mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \mathbf{u} = \lambda\mathbf{W}\mathbf{u} + \boldsymbol{\varepsilon}$	$\boldsymbol{\theta} = \mathbf{0}$
Spatial Durbin model	$\mathbf{y} = \rho\mathbf{W}\mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$	$\lambda = 0$
Spatial autoregressive	$\mathbf{y} = \rho\mathbf{W}\mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$	$\lambda = 0; \boldsymbol{\theta} = \mathbf{0}$
Spatially lagged Xs	$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$	$\rho = \lambda = 0$
Spatial error model	$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \mathbf{u} = \lambda\mathbf{W}\mathbf{u} + \boldsymbol{\varepsilon}$	$\rho = 0; \boldsymbol{\theta} = \mathbf{0}$
(Spatial) Linear model	$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$	$\rho = \lambda = 0; \boldsymbol{\theta} = \mathbf{0}$

outcomes, restricting the other possibilities to zero. SEMs imply that the pattern of spatial dependence is attributable to unmeasured covariates that are orthogonal to the included regressors, resulting in a non-spherical error variance-covariance matrix.<sup>14</sup> Under these conditions, parameter estimates are unbiased but inefficient (and standard errors are incorrectly estimated). Efficient parameter estimates and correct standard errors can be obtained by accounting for the spatial structure of the residuals, as done in the SEM:

$$y = X\beta + u, \text{ where } u = \lambda W u + \varepsilon \quad (5)$$

where  $W$  is an  $N \times N$  connectivity matrix with elements  $w_{ij}$  indicating the (pre-specified, exogenous) relative connectivity (i.e., relationship) from unit  $j$  to unit  $i$  and  $\lambda$  indicating the strength of the spatial interdependence along this pre-specified pattern of connections.<sup>15</sup> In the terms of Table 39.1, this model assumes global spillovers in the unobservables, i.e., that the residuals are governed by a spatial autoregressive process.<sup>16</sup> This will also be the preferred specification when we believe there is clustering in the unobservables. Unlike with observable predictors, we have no means of introducing this heterogeneity into the systematic component of the model directly and must assume that these unobserved components are orthogonal to the observed ones (and so no bias issue), but accounting for the structure of the residuals should still provide some insurance against inefficiency resulting from this type of clustering and produce more accurate standard error estimates.<sup>17</sup>

If, instead, researchers believe that there are spillovers in the observables, one of the other single-source spatial models should be estimated to *i*) avoid bias in the estimated non-spatial effect-parameters and *ii*) obtain estimates of these spatial (spillover) effects. Where theory and substance suggest these spillovers/externalities are local, the SLX model should be preferred:

$$y = X\beta + WX\theta + \varepsilon \quad (6)$$

Alternatively, where theory indicates these spillovers/externalities are global and in the outcome, the widely used SAR model is called for:<sup>18</sup>

$$y = \rho Wy + X\beta + \varepsilon \quad (7)$$

This will likely be familiar to most readers, as it has quickly become the workhorse model of applied spatial work in political science (and elsewhere). While both SLX and SAR models allow for spillovers in observables, they differ over whether they model these as local or global processes, as discussed earlier, and whether there are spatial effects in the unobservables. More theoretically, they also differ over whether we believe there is cause to understand the spillovers of the observables as direct,  $x_j \Rightarrow y_i$ , as is more likely with social aggregates, like GDP for example, or indirect,  $(x_j, \varepsilon_j) \Rightarrow y_j \Rightarrow y_i$ , as is more likely with strategic interdependence among decision makers like as in public policies, for example.

We have noted that the similarity of these models creates challenges for diagnostic tests. While this may not be obvious from the structural forms given in Equations 5–7, we can re-express them to highlight the similarities. Taking the reduced form of  $u$  and substituting and rearranging terms, the SEM model becomes

$$y = \lambda Wy + X\beta - \lambda WX\beta + \varepsilon \quad (8)$$

The similarities between the SEM model and the SLX model (given in Equation 6) and the SAR model (given in Equation 7) are now readily apparent, as it is composed of a spatial lag of the outcomes ( $\lambda Wy$ ) and spatial lags of the predictors ( $\lambda WX\beta$ ). Similarly, taking the reduced form of the SAR model in Equation 7 and its expansion produces

$$y = (I - \rho W)^{-1}(X\beta + \varepsilon) \quad (9a)$$

$$y = X\beta + \rho WX\beta + \rho^2 W^2 X\beta \dots + \varepsilon + \rho W\varepsilon + \rho^2 W^2 \varepsilon \dots \quad (9b)$$



Again, the similarities between the **SAR** and **SLX** models are now apparent, with the only differences being the higher-order polynomials of the spatial lag of **X** and the spatial error process. As a consequence, unmodeled spatial spillovers/externalities in the observable predictors, in the unobservables, or in the outcomes will result in a rejection of the zero null for the spatial-effect parameter in *any* of these single-source models.

To ward against this possibility, spatial econometricians have increasingly recommended the two-source models:

$$\text{SDM: } \mathbf{y} = \rho \mathbf{W}\mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon} \tag{10}$$

$$\text{SAC: } \mathbf{y} = \rho \mathbf{W}\mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \tag{11}$$

where  $\mathbf{u} = \lambda \mathbf{W}\mathbf{u} + \boldsymbol{\varepsilon}$

$$\text{SDEM: } \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\mathbf{X}\boldsymbol{\theta} + \mathbf{u}, \tag{12}$$

where  $\mathbf{u} = \lambda \mathbf{W}\mathbf{u} + \boldsymbol{\varepsilon}$

and a more general model still: the so-called General Nesting Spatial Model (**GNS**),

$$\text{GNS: } \mathbf{y} = \rho \mathbf{W}\mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\mathbf{X}\boldsymbol{\theta} + \mathbf{u}, \tag{13}$$

where  $\mathbf{u} = \lambda \mathbf{W}\mathbf{u} + \boldsymbol{\varepsilon}$

which imposes no constraints on the three spatial parameters ( $\rho$ ,  $\lambda$ ,  $\theta$ ).<sup>19</sup> Given that this model subsumes all the alternatives presented thus far, one might think to engage in a Hendry-like general-to-specific specification search (Hendry, 1995), thereby avoiding the pitfalls encountered when adopting a specific-to-general approach. While this strategy has much to recommend it and is commonplace in the time-series literature, there are two problems that prevent adopting the general-to-specific approach in the spatial context.

First, the **GNS** model is weakly identified. As discussed in Gibbons and Overman (2012), the **GNS** is the analog to Manski's

(1993) well known linear-in-means neighborhood-effects model:

$$y = \underbrace{\rho_1 E[y | a]}_{\text{Endog. Effects}} + x\boldsymbol{\beta} + \underbrace{E[x' | a]\boldsymbol{\gamma}}_{\text{Exog. Effects}} + v, \tag{14a}$$

where  $u = \underbrace{\rho_2 E[u | a]}_{\text{Corr. Errors}} + \boldsymbol{\varepsilon}$

$$y = x\boldsymbol{\beta} + E[x' | a] \frac{(\rho_1 \boldsymbol{\beta} + \boldsymbol{\gamma})}{1 - \rho_1} + \frac{\rho_1}{1 - \rho_1} E[v | a] + u \tag{14b}$$

This parallel should raise some red flags given the well known identification problems of the Manski model. As indicated in Equation 14b, it is impossible to separately identify the endogenous and exogenous spatial effects in this model.<sup>20</sup> With spatial econometric methods, however, one does not simply estimate one sum 'neighborhood' effect: each unit in a sample is known to be connected to others through **W**, and this matrix almost always provides more information than neighborhood membership. For example, within a given 'neighborhood', there are first-, second-, and higher-order neighbors.<sup>21</sup> As a result, spatial-econometric models are *usually* able to use the pre-specification of **W** to achieve identification in most cases.<sup>22</sup>

Focusing on an example with a single predictor, some algebraic manipulation of the **GNS** model in Equation 13 allows it to be re-written as

$$\mathbf{y} = (\rho + \lambda)\mathbf{W}\mathbf{y} - \rho\lambda\mathbf{W}^2\mathbf{y} + \mathbf{x}\boldsymbol{\beta} + (\theta - \lambda\boldsymbol{\beta})\mathbf{W}\mathbf{x} - \lambda\theta\mathbf{W}^2\mathbf{x} + \boldsymbol{\varepsilon} \tag{15a}$$

$$\mathbf{y} = q_1\mathbf{W}\mathbf{y} + q_2\mathbf{W}^2\mathbf{y} + \mathbf{x}q_3 + q_4\mathbf{W}\mathbf{x} + q_5\mathbf{W}^2\mathbf{x} + \boldsymbol{\varepsilon} \tag{15b}$$

where the spatial parameters are weakly identified by the second-order terms in the polynomial. The reduced form of the **GNS** provides five parameters from which we can recover the four structural parameters.

Substituting  $q_1$  into  $q_2$  and  $q_4$  into  $q_5$  gives a set of quadratic relationships for  $\lambda$ :

$$\lambda^2 = q_2 + \lambda q_1$$

$$\lambda^2 = \frac{-(q_5 + \lambda q_4)}{\beta}$$

These equations provide a unique solution for  $\lambda$  and, in turn, the other parameters:

$$\lambda = \frac{-(\beta q_2 + q_5)}{\beta q_1 + q_4}$$

$$\rho = q_1 + \frac{(\beta q_2 + q_5)}{\beta q_1 + q_4}$$

$$\theta = q_4 - \frac{\beta(\beta q_2 + q_5)}{\beta q_1 + q_4}.$$

The problem is that these parameters are identified solely by the structural assumptions of the functional form implied by autoregression and the pre-specified  $\mathbf{W}$ , and the performance of the **GNS** model deteriorates rapidly as these assumptions become more appreciably incorrect, as we show later in our Monte Carlo experiments.

How, then, should researchers who are interested in undertaking spatial analysis proceed? Broadly, there are two strategies one can pursue. The first is to constrain one of the spatial parameters to zero, thereby allowing firmer identification of the remaining free parameters and more robust estimation of the relevant two-source model.<sup>23</sup> The second is to add additional structure to the model in the form of unique weights matrices for the observables and unobservables. While possible, this second approach seems unappealing to us as a general strategy, given that we can think of no reason why we would generically expect to have strong prior information to indicate that unobserved effects are spatially governed in a manner distinct from observed predictors.<sup>24</sup> Accordingly, we focus on evaluating the efficacy of the first strategy, constraining one or more parameters, as a more generally applicable approach.

Implicitly, this is the approach currently advocated by most spatial econometricians, who have increasingly recommended one or another of the two-source models. However, to date, researchers have received conflicting advice over which model should be preferred as a general model, with some strongly advocating the **SDM** and others the **SAC**. Elhorst (2010: 10) offers a fun account that highlights this discord: ‘In his keynote speech at the first World Conference of the Spatial Econometrics Association in 2007, Harry Kelejian advocated [**SAC** models], while James LeSage, in his presidential address at the 54th North American Meeting of the Regional Science Association International in 2007, advocated [**SDM**] models’. Moreover, most of the work systematically exploring the small-sample performance of these models generally has done so with data-generating processes that satisfy the constraints assumed by the statistical model.

Instead of simply advocating one model over another, as is commonly done, we believe researchers should adopt a more systematic approach to motivating these constraints. First, one could use research design, such as natural experiments, to eliminate one (or more) of the three possible sources. This is the strategy suggested by Gibbons and Overman (2012), both to evade the issues that arise from the unidentified **GNS** and avoid models only identified off-structure (e.g., spatial econometric models, generally).<sup>25</sup> Focusing exclusively on those contexts where natural experiments are available, however, bounds the range of issues that can be studied. As such, we consider approaches where such strategies are not possible.<sup>26</sup>

A natural alternative in such instances is to use theory to guide these constraints. Where theory can eliminate one of the possible sources, we should be more confident in our selection of the appropriate two-source model. Even where we do not have strong theory to confidently eliminate one

of these sources, we suggest a third alternative: use the aim of the research to guide the model selection. That is, where researchers are principally interested in obtaining unbiased estimates of the non-spatial *parameters*,<sup>27</sup> the spatial Durbin model should be preferred. This should provide the most insurance against possible omitted variable bias by explicitly introducing both forms of observable spillovers into the systematic component of the model. However, where researchers are explicitly interested in evaluating spatial theories, we believe one of the other two-source models (**SAC** or **SDEM**) are best. Each frees one parameter to capture spillovers in observables (either  $\rho$  or  $\theta$ ) while accounting for spatial effects in the unobservables ( $\lambda$ ). To us, distinguishing between spatial spillovers in observables and spatial effects in unobservables is the most significant consideration. Importantly, this will help prevent researchers from drawing erroneous conclusions about diffusion and/or spillovers where none exist, i.e., where spatial clustering in the outcomes is determined in whole or part by spatial effects in unobservables. Where such spillovers still find support, we have only lost the ability to statistically and empirically distinguish whether they were truly global or local – a cost that, by comparison, seems less severe.

Using either theory or research focus to guide specification, however, also naturally risks a much more problematic cost: estimating the incorrect model (and so, generally, calculating incorrect effect estimates). This can occur in four ways with the estimation of two-source spatial models: (1) the truth is all three spatial effects; (2) the truth is two sources but our statistical model imposes the wrong constraint, yielding the wrong two-source model; (3) the truth is a single-source model; (4) the truth is a non-spatial model. In either of the first two, we risk bias in the estimates of the included spatial and non-spatial parameters, as is always the case with spatially misspecified models.

Thirdly, if the truth is a single source among our included two, the estimation should reveal that. If our two-source model does not include the true single source, the combination of estimated coefficients on the included should produce that omitted third, but we would incorrectly find support for *both* included spatial parameters being non-zero, even though the truth is that only the omitted third is non-zero (the fourth, non-spatial case should be unproblematic, as the estimation would return to zero for the spatial parameters.)

The possibility that an omitted single-source process would be reproduced through the combination of two-source parameter estimates has been well established for the **SEM** model, which can be re-expressed as a spatial Durbin model (noted above and re-expressed here):

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \text{ where } \mathbf{u} = \lambda\mathbf{W}\mathbf{u} + \boldsymbol{\varepsilon} \quad (17a)$$

$$\mathbf{y} = \lambda\mathbf{W}\mathbf{y} + \mathbf{X}\boldsymbol{\beta} - \lambda\mathbf{W}\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (17b)$$

In this case, we can test the common-factor restriction of Burridge (1981),  $\boldsymbol{\theta} = -\lambda\boldsymbol{\beta}$ , after estimating an SDM to evaluate whether it can be constrained to the SEM. Similarly, we can see that the SAR model can be re-expressed as a higher-order variation of the SDEM:

$$\mathbf{y} = \rho\mathbf{W}\mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (18a)$$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \rho\mathbf{W}\mathbf{X}\boldsymbol{\beta} + \rho^2\mathbf{W}^2\mathbf{X}\boldsymbol{\beta} + \dots + (\mathbf{I} - \rho\mathbf{W})^{-1}\boldsymbol{\varepsilon} \quad (18b)$$

Thus, the only difference between the SDEM and the SAR model is the higher-order polynomials of  $\mathbf{W}\mathbf{X}$  in the latter.<sup>28</sup> Finally, while expressing the relationship between the SLX and the SAC model is not as straightforward, the basic intuition for why a true effect of  $\theta$  in the SLX model would cause significant findings for both  $\rho$  and  $\lambda$  in the SAC model

parallels the above discussions in that the estimates of each is a function of  $\mathbf{WX}$ :

$$\mathbf{y} = \rho \mathbf{W}\mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \mathbf{u},$$

where  $\mathbf{u} = \lambda \mathbf{W}\mathbf{u} + \boldsymbol{\varepsilon}$  (19a)

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \rho \mathbf{W}\mathbf{X}\boldsymbol{\beta} + \rho^2 \mathbf{W}^2 \mathbf{X}\boldsymbol{\beta} + \dots + (\mathbf{I} - \rho \mathbf{W})^{-1} (\mathbf{I} - \lambda \mathbf{W})^{-1} \boldsymbol{\varepsilon}$$
 (19b)

Both this and the SAR–SDEM relation do not allow for a simple common-factor restriction test (as in the SEM–SDM case). Therefore, rather than testing constraints on parameters, one could use tests that compare the performance of non-nested models. For example, the ‘closeness’ test in Vuong (1989) can evaluate whether the two models differ significantly in their ability to explain the data. In this context, a failure to reject the null hypothesis would indicate support for the more parsimonious single-source model. We do not explore this approach at length here, but it may warrant further consideration in subsequent work.

In the next section, we explore the consequences of imposing the wrong constraints when estimating spatial models.

**MONTE CARLO ANALYSIS**

In our simulations, we explore the possibility of detecting interdependence in outcomes and spillovers from covariates in cross-sections of data when there is spatial clustering in both observables and unobservables using the relevant models from Table 39.1. We define clustering as a common spatial or group fixed effect. Substantively, clustering differs from both interdependence and spillovers in that changes in covariates and disturbances inside one unit do not cause outcomes to change in other units. The simulation DGP is

$$\mathbf{y} = \rho \mathbf{W}\mathbf{y} + \boldsymbol{\beta}\mathbf{x} + \theta \mathbf{W}\mathbf{x} + \mathbf{u}$$

where  $\mathbf{y}$  is an  $N \times 1$  vector of outcomes,  $\mathbf{x}$  is an  $N \times 1$  covariate vector,  $\mathbf{u}$  is an  $N \times 1$  vector of disturbances,  $\mathbf{W}$  is an  $N \times N$  spatial weights matrix,  $\rho$  is the spatial interdependence parameter,  $\boldsymbol{\beta}$  is the ‘direct-effect’ parameter, and  $\theta$  is the spatial spillover parameter.

The individual elements of the vectors  $\mathbf{x}$  and  $\mathbf{u}$  are generated as

$$x_{ig} = \eta_g^x + \varepsilon_{ig}^x \text{ and } u_{ig} = \eta_g^u + \varepsilon_{ig}^u$$

where  $x_{ig}$  and  $u_{ig}$  refer to the covariate and disturbance for unit  $i$  in spatial group  $g$ ,  $\eta_g^x$  and  $\eta_g^u$  are the common spatial effects, distributed as standard normal variates (clustering), and  $\varepsilon_{ig}^x$  and  $\varepsilon_{ig}^u$  are the unit-specific components of the covariate and disturbance, which are also distributed as standard normal variates.

The spatial weights matrix identifies intra-group connectivity and takes the form

$$\mathbf{W} = \begin{bmatrix} \mathbf{W}_1 & 0 & \dots & 0 \\ 0 & \mathbf{W}_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \mathbf{W}_G \end{bmatrix}$$

Thus, the complete weights matrix has a block diagonal structure for  $\mathbf{G}$  groups, when the units or individuals in the sample are stacked by groups. We set the number of groups ( $\mathbf{G}$ ) to 15, the number of members in each group ( $n_g$ ) to 20, and the degree of intragroup connectivity at 40%. We assume the connectivity weights are uniform and sum to one. That is, the weights are  $1/n_c$ , where  $n_c$  is the number of intragroup connections. This weights matrix is motivated by the fact that we usually do not know the relevant spatial groups. Should North Africa be grouped with Sub-Saharan Africa? Does Pennsylvania belong in the Northeast or Midwest? We do however observe intragroup relationships such as contiguity.

We evaluate the small-sample performance of the **SAR**, **SAC**, **SLX**, **SDEM**, **SDM**, and **GNS** models under four experimental conditions: (1) no spillovers and no interdependence ( $\theta = 0, \rho = 0$ ), (2) spillovers and no interdependence ( $\theta = 0.2, \rho = 0$ ), (3) no spillovers and interdependence ( $\theta = 0, \rho = 0.2$ ), and (4) both spillovers and interdependence ( $\theta = 0.2, \rho = 0.2$ ). We set  $\beta = 2$  in all of our experiments. Furthermore, clustering in the covariate and in the disturbances, as generated above, are present in all experiments.

Table 39.2 provides the ML estimates for the direct covariate effect ( $\hat{\beta}$ ). It is notable that all of the models perform reasonably well across the experiments, with

the exception of **SAR**. The direct effect is underestimated on average with this model. Clustering in the disturbances strengthens their correlation with the spatial lag, above and beyond the correlation that exists when the structural disturbances are independent and identically distributed (i.i.d.) This generates an inflating simultaneity bias in  $\hat{\rho}$ , which induces an attenuating bias in  $\hat{\beta}$ . Moreover, estimation using the SAR model performs relatively poor in root-mean-squared-error terms (largely a function of the bias), and the standard error estimates are overconfident (we should note, however, that SAR's underestimation of  $\beta$  tends to be in some partial way compensated by its overestimation of  $\rho$  in

**Table 39.2 ML estimates of covariate-coefficient estimate ( $\hat{\beta}, \beta = 2, N = 300, 1,000$  trials)**

	(1)	(2)	(3)	(4)
	$\theta = 0, \rho = 0$	$\theta = 0.2, \rho = 0$	$\theta = 0, \rho = 0.2$	$\theta = 0.2, \rho = 0.2$
<b>SAR</b>				
Bias	-0.18	-0.17	-0.20	-0.18
RMSE	0.21	0.19	0.22	0.20
Overconfidence	1.47	1.42	1.47	1.41
<b>SAC</b>				
Bias	-0.01	-0.01	-0.01	-0.01
RMSE	0.07	0.07	0.07	0.07
Overconfidence	1.05	1.05	1.05	1.06
<b>SLX</b>				
Bias	0.00	0.00	0.01	0.01
RMSE	0.07	0.07	0.08	0.08
Overconfidence	0.95	0.95	0.92	0.92
<b>SDEM</b>				
Bias	0.00	0.00	0.00	0.00
RMSE	0.06	0.06	0.06	0.06
Overconfidence	1.05	1.05	1.03	1.03
<b>SDM</b>				
Bias	0.00	-0.01	-0.03	-0.04
RMSE	0.06	0.06	0.07	0.08
Overconfidence	1.06	1.06	1.05	1.04
<b>GNS</b>				
Bias	0.00	0.00	-0.01	-0.01
RMSE	0.06	0.06	0.06	0.06
Overconfidence	1.03	1.04	1.06	1.07

terms of yielding an estimated total effect  $(\mathbf{I} - \hat{\rho}\mathbf{W})^{-1}\hat{\beta}$  closer to the true value, two.

The results for the spatial interdependence-parameter estimates ( $\hat{\rho}$ ) are presented in Table 39.3. Here, we see the inflation bias (in SAR and SDM, especially) driven by the unmodeled spatial clustering in the disturbances. The standard error estimates are highly overconfident as well. In SAR, across all four experiments, the standard deviation in the sampling distribution for  $\hat{\rho}$  is more than double the size of the average estimated standard error. The combination of an inflation bias and overconfident standard errors means the rejection rate is extremely high when the null hypothesis is true. In other words, estimation with SAR produces a high rate of false positive rejections when there

is unmodeled clustering (this is Galton's problem).

Estimation with SAC does better than with SAR or SDM in terms of bias, root-mean-squared-error performance, and standard error accuracy. The improvement stems from the fact that SAC accounts for the clustering in the disturbances (unmodeled/unobserved factors) by allowing them to follow a spatial AR process. This is not a perfect representation of the true DGP, but the AR specification is easy to implement when the spatial groups are not known, and there are substantial gains from doing so. The SAC provides protection against false positive rejections. The cost is a loss of power, which is large in column (3). However, the rate at which the SAC model correctly rejects the null hypothesis is

**Table 39.3 ML estimates of interdependence ( $\hat{\rho}, \beta = 2, N = 300, 1,000$  trials)**

	(1)	(2)	(3)	(4)
	$\theta = 0, \rho = 0$	$\theta = 0.2, \rho = 0$	$\theta = 0, \rho = 0.2$	$\theta = 0.2, \rho = 0.2$
<b>SAR</b>				
Bias	0.29	0.32	0.23	0.45
RMSE	0.30	0.33	0.24	0.46
Overconfidence	2.24	2.17	2.25	2.18
False positives (0.10 level)	97.4%	99.2%		
Power (0.10 level)			99.9%	99.9%
<b>SAC</b>				
Bias	-0.08	0.01	-0.09	0.20
RMSE	0.12	0.10	0.14	0.22
Overconfidence	1.16	1.21	1.19	1.24
False positives (0.10 level)	28.8%	18.2%		
Power (0.10 level)			36.9%	72.6%
<b>SDM</b>				
Bias	0.70	0.70	0.56	0.76
RMSE	0.70	0.70	0.56	0.76
Overconfidence	1.38	1.38	1.33	1.36
False positives (0.10 level)	100%	100%		
Power (0.10 level)			100%	100%
<b>GNS</b>				
Bias	0.11	-0.03	-0.30	-0.22
RMSE	0.80	0.79	0.75	0.66
Overconfidence	7.06	6.20	4.75	3.84
False positives (0.10 level)	97.9%	97.5%		
Power (0.10 level)			92.2%	89.0%

sensitive to experimental conditions. If we increase the strength of interdependence, for example, the power will improve. Both the **SDM** and **GNS** models perform poorly, producing biased estimates and overconfident standard errors.

Table 39.4 provides the ML estimates for the spillover parameter ( $\hat{\theta}$ ). Whenever there is no interdependence ( $\rho = 0$ ), estimates from the **SLX** model do well in terms of bias but not in terms of efficiency. The variance in the sampling distribution is relatively large. Also, the standard errors are highly overconfident. Across the experiments, the standard deviations for the empirical sampling distributions are about 2.5 times large than the average estimated standard error. Because of the overconfident standard errors, the **SLX**

model produces a high rate of false positive rejections, even when there is no interdependence. When there is interdependence, omitted variable bias causes the performance of **SLX** to deteriorate further. Similar to the **SAC** improvement over **SAR**, estimation with **SDEM** does better than with **SLX** in terms of bias, root-mean-squared-error performance, and standard error accuracy. **SDEM** provides some protection against false positive rejections; the cost for this protection is a loss of power. Again, both the **SDM** and **GNS** models perform poorly, producing biased estimates and overconfident standard errors.

To sum, clustering in unobservables – for example, unobserved or unmodeled clustered factors – complicates our ability to detect interdependence in outcomes and spillovers

**Table 39.4 ML estimates of spillover effect ( $\hat{\theta}$ ,  $\beta = 2$ ,  $N = 300, 1,000$  trials)**

	(1)	(2)	(3)	(4)
	$\theta = 0, \rho = 0$	$\theta = 0.2, \rho = 0$	$\theta = 0, \rho = 0.2$	$\theta = 0.2, \rho = 0.2$
<b>SLX</b>				
Bias	0.00	0.00	0.46	0.49
RMSE	0.32	0.32	0.60	0.63
Overconfidence	2.41	2.41	2.62	2.62
False positives (0.10 level)	47.8%		74.8%	
Power (0.10 level)		56.6%		89.7%
<b>SDEM</b>				
Bias	0.01	0.01	0.41	0.42
RMSE	0.21	0.21	0.46	0.47
Overconfidence	1.12	1.12	1.08	1.08
False positives (0.10 level)	13.6%		70.0%	
Power (0.10 level)		31.6%		92.1%
<b>SDM</b>				
Bias	-1.60	-1.70	-1.55	-1.64
RMSE	1.61	1.71	1.56	1.65
Overconfidence	1.29	1.31	1.28	1.32
False positives (0.10 level)	99.9%		99.9%	
Power (0.10 level)		99.9%		99.9%
<b>GNS</b>				
Bias	-0.23	0.05	0.58	0.81
RMSE	1.84	1.89	1.80	1.77
Overconfidence	6.16	5.71	4.59	3.89
False positives (0.10 level)	97.6%		97.7%	
Power (0.10 level)		98.9%		98.7%

from observable covariates in cross-sections of data. When one suspects both interdependence and exogenous spillovers, it would seem natural to estimate either the **SDM** or **GNS** models, but this is not advisable under these conditions. The **SDM** allows for both interdependence and spillovers, but it ignores the clustering in disturbances. This omission generates a bias in the estimates for the interdependence parameter ( $\rho$ ) and the spillover parameter ( $\theta$ ). Why not also allow for spatial correlation in the disturbances? This is what the **GNS** does. Unfortunately, this model is only identified from strong structural assumptions (functional form and **W**), and, in this case, the structural assumption about the disturbances is incorrect, so failure of these other strong assumptions has severe ramifications. Therefore, **GNS** tends not to perform any better than **SDM**. Both models frequently produce statistically significant estimates of interdependence and spillover parameters with the wrong sign!

When one suspects clustering on unobservables, it does not seem advisable to estimate either the **SDM** or **GNS** models. Instead, estimating either **SAC** or **SDEM** would seem to be a more prudent strategy (and preferable to **SAR** and **SLX** as well). While design should be leveraged to select between these models where possible, often this will not be an option and researchers will instead have to eliminate either interdependence or spillovers (plus spatial error dependence) on theoretical grounds.<sup>29</sup> This makes it difficult to offer a general prescription; however the nature of one's data will often be instructive. When the outcomes of interest are social aggregates – such as unemployment rates, crime rates, or the aggregate demand for cigarettes (these are common outcomes in the spatial-econometrics literature) – outcome contagion makes little sense. The unemployment rate in one locality does not literally cause the unemployment rate in another; rather, economic conditions cluster spatially and economic conditions in  $j$  cause unemployment in  $i$

(exogenous spillovers). On the other hand, when the outcomes are choices made by strategically interdependent actors – as is common in political science – interdependence is far more plausible: tax rates in  $i$  do likely respond to tax rates in  $j$ .

Ultimately, however, researchers are simply deciding whether theory indicates that changes to my neighbors' covariates affect me directly (as in **SDEM**) or indirectly through the changes they elicit in my neighbors' outcome (as in **SAC**). To us, this consideration seems less consequential (though not inconsequential because **Wy** implies multipliers whereas **WX** does not) than determining whether the spatial clustering we observe in the outcomes arises from spillovers of either exogenous (**WX**) or endogenous (**Wy**) type or merely through the presence of (spatially) common unobservables, which both **SDEM** and **SAC** better enable us to do.

## CONCLUSION

In general, there are two primary conclusions with which we hope to leave readers and practitioners. The first conclusion is the importance of undertaking appropriate diagnostics to explicitly test the restrictions implied by one's model, considering the different sources of spatial association, exogenous and endogenous, observed and unobserved. While this will seem obvious to readers more familiar with model specification in other literatures (e.g., time series), these issues have not been as well articulated in the spatial literature that guides political scientists to date. This is especially important in spatial-analytic contexts, where researchers are more likely to attach theoretic importance to these findings and, as such, should exercise greater care when specifying their models. The second is that no single model can or should serve universally as baseline specification that guards against misspecification. While some in the spatial-econometric literature have advocated



strongly for the **SAR**, **SLX**, or spatial Durbin models, we present a variety of theoretically plausible and empirically likely conditions where each of these models will cause researchers to draw faulty inferences. We have argued and presented support for the case that the heretofore relatively neglected **SAC** or **SDE** models will have broader utility as prudent defaults; but even those, we would acknowledge, can perform poorly under some plausible conditions.

For interested readers, we expand on the discussion presented here in Cook et al. (forthcoming, b) in several ways. First, we consider alternative specifications of the weights matrix, beyond the block-group structure examined here. Second, we increase the true effect size of the spatial lag of the predictor. Here, we have used a common coefficient size for each of the processes, however this implies a larger total effect size for the spatial lag of the outcome than the spatial lag of the predictor. Third, we consider not just coefficient estimates and hypothesis tests but substantive effects as well – i.e., the derivatives  $dy/dx$  (see Franzese, Chapter 31, this *Handbook*) – comparing the efficacy of the various models in capturing pre-spatial, post-spatial, and total effects. Fourth, we consider extensions to TSCS data, where temporal dynamics may also impair spatial-model specification. Finally, we illustrate our approach to model selection with an empirical example (democracy and income) to aid applied researchers.<sup>30</sup>

## ACKNOWLEDGEMENTS

We thank the participants at the 2015 Texas A&M Conference on Innovations in Comparative Political Methodology, the 2015 Summer Methods Meeting of the Society for Political Methodology, the 2016 Meeting of the American Political Science Association, and the 2018 Workshop Modeling Spatial and Network Interdependence in International Relations for their useful feedback on this project. Any remaining errors are ours alone.

## Notes

- 1 See Franzese and Hays (2008) for a fuller account of the substantive range of ‘spatial’ theories advanced in political science. In addition, Cook et al. (2019) discuss the application of spatial-econometric models to research in public administration and public management.
- 2 This trend is likely to continue growing as these methods become more familiar to researchers and packages facilitating their easy estimation become available in widely used statistical software languages, such as the ‘spdep’ package in *R* and the ‘sp’ suite in *Stata 15*.
- 3 Briefly noting the models that these would imply: spatial clustering can manifest due to unobserved factors common to proximate units, suggesting the spatial error model (**SEM**), or through exogenous perturbations to the predictors in my neighbor(s), which can influence me directly, motivating a spatially lagged X (**SLX**) model, or indirectly, by affecting my neighbors’ outcome and thereby my own outcome, as in a spatial autoregressive (**SAR**) model. Or it might be any combination thereof, suggesting one of several models that are more general.
- 4 We suspect this is often true in experimental data on human subjects as well.
- 5 For clarity, we confine our attention in this paper to the cross-sectional analysis of continuous data. While many of the themes and topics generalize to a broader set of circumstances, we save peculiarities confronted when dealing with qualitative outcomes and/or panel/time-series-cross-sectional for address in other work.
- 6 Although most of the literature uses the terms *observables* and *unobservables*, the actual issue is whether these factors are *observed* and included in the model’s systematic component or *unobserved* and left in the unmodeled residual component. We will continue to follow convention throughout, but the reader is encouraged to understand (*un*) *observed*s for all references to (*un*) *observables*.
- 7 Generally, this is discussed as the predictor and/or residual being governed by a spatial autoregressive process. However, it may also be that the predictor is a function of spatially correlated (but not autoregressive) factors. The consequences with respect to parameter estimates in the model of *y* are identical.
- 8 More formally, Andrews (2005) states common-factor residuals and predictors satisfy the following:

$$u_i = \mathbf{C}'_g \mathbf{u}_i^*$$

$$x_i = \mathbf{C}'_g \mathbf{x}_i^*$$

where  $\mathbf{C}_g$  is a random common (e.g., group) factor with random factor loadings  $\mathbf{u}_i^*$  &  $\mathbf{x}_i^*$ .

Therefore, if units  $i$  and  $j$  are each members of group  $g$ , they are jointly impacted by the respective loading.

- 9 The distinction also closely parallels that between moving-average (MA) and autoregressive (AR) processes in time-series contexts. Roughly, spatial-lag  $\mathbf{X}$  (and spatial-MA error) processes are local, MA-like, and spatial-lag  $y$  (and spatial-AR error) processes are global, AR-like (in the errors only, not the outcomes, as in the SAR-error case).
- 10 This is analogous to Anselin's (2003) two-dimensional taxonomy for externalities.
- 11 Note that this is a partial list. All of the models presented here assume parameter constancy, first-order spatial dependence (when present), and global (and not local) spatial autocorrelation in the unobservables (when present). As noted, these are the most common alternatives in the literature and importantly include those advocated by LeSage and Pace (2009) and Elhorst (2010).
- 12 For example, Buhaug and Gleditsch (2008) argue that conflicts cluster in space because the characteristics that produce conflict also cluster in space. If correct, this would be captured simply via the inclusion of the relevant country-characteristics. Instead, they estimate a model with spatially lagged independent variables (e.g., democracy in contiguous countries), these  $\mathbf{WX}$ s actually relate to a different argument as we discuss later.
- 13 As always, the distribution of our residuals – spatial or otherwise – is entirely dependent on the specification of the systematic component of our model.
- 14 In the remaining models, we will continue to assume that the residuals are orthogonal after the appropriate spatial specification is set. The possible endogeneity of the predictors present further complications as discussed in Betz et al. (2020).
- 15 Under spatial dependence in orthogonal residuals, standard errors can also be consistently estimated, leaving the parameter-inefficiency unaddressed, by using appropriately designed robust standard errors (Driscoll and Kraay, 1998).
- 16 The local (i.e., moving average) analog to this model would be given as

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon + \gamma\mathbf{W}\varepsilon$$

where the residual is decomposed into a spatial and non-spatial component. However, unlike the more common **SEM**, there is not autoregression in the residuals and therefore there is no inverse required in the reduced form, as noted by Anselin (2003). This model is not widely used in practice, likely because researchers have little information

to justify this constraint, instead preferring the perhaps greater generality of the **SEM** model.

- 17 Note that this is not true of panel or time-series-cross-sectional data, where we can use spatial fixed effects to account for time-invariant heterogeneity in the unobservables directly. An example of this can be found in Cook et al. (2019).
- 18 In actuality, the **SAR** model suggests global spillovers in both the observables and unobservables as we can see from the reduced form given below.
- 19 We note again that each of these models assumes a global autocorrelation in  $\mathbf{y}$  and/or  $\varepsilon$  and that only first-order processes are considered.
- 20 Instead, all that is identified is the total spillover effect; this is *Manski's reflection problem*.
- 21 This should suggest the importance of  $\mathbf{W}$  given that the degree to which the weights matrix accurately reflects the true spatial relationships among the units is paramount. Both our ability to detect whether spatial dependence is present and to identify which source of spatial effects are present depend upon the accuracy of  $\mathbf{W}$ .
- 22 In this instance, the spatial analog is

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{WX}(\beta\rho_1 + \gamma) + \rho_1\mathbf{WX}(\beta\rho_1 + \gamma) + \rho_1^2\mathbf{W}^2\mathbf{X}(\beta\rho_1 + \gamma) + \dots + \varepsilon$$

- 23 While we do not fully elaborate it here, the intuition – beyond simply being identified – as to why two parameter specification checks work well follows directly from Anselin et al.'s (1996) robust Lagrange Multiplier tests (here given for spatial error):

$$LM_{\lambda}^* = \frac{(\hat{\varepsilon}'\mathbf{W}\hat{\varepsilon} / \hat{\sigma}_{\varepsilon}^2 - \psi\hat{\varepsilon}'\mathbf{W}\mathbf{y} / \hat{\sigma}_{\varepsilon}^2)^2}{T[1 - \psi]}$$

which treats  $\rho$  – the spatial heterogeneity attributable to the spatial lag of the outcomes – as a nuisance parameter, adjusting for its effect on the likelihood. In effect, removing the portion of  $\text{cov}(\hat{\varepsilon}, \mathbf{W}\hat{\varepsilon})$  that can be attributable to  $\text{cov}(\hat{\varepsilon}, \mathbf{W}\mathbf{y})$ . Equivalently, we could construct additional pre-specification tests (or simply estimate models) that hold fixed the effect of one alternative while evaluating the second.

- 24 Even when these exist, the likely high degree of correlation between the weights matrices would likely leave a still weakly identified model.
- 25 See also Egami (2018) for a strategy attempting nonparametric causal-inference tests of spatial spillovers in observational data.
- 26 Egami's (2018) approach requires no temporally simultaneous interdependence and is designed to test for, but not estimate (see Franzese, Chapter 31, this *Handbook*), spatial effects. We are inter-

- ested in spatial-effect estimation in other contexts.
- 27 Unbiased estimation of isolated parameters is sufficient for testing purposes, for effect or response estimation; however, generally one needs more (see Franzese, Chapter 31, this *Handbook*).
  - 28 While this does not as easily permit a Burridge-type restriction, we could specify a higher-order SDEM model and then perform an F-test of zero coefficients on these higher-order polynomials. Rejection would indicate that the standard SDEM model is insufficient. To be clear, we would not be able to reject the possibility that the truth is some higher-order SDEM from this analysis. This problem is analogous to that discussed by Beck (1991) in the time-serial context, where the AR(1) model can be closely approximated by a higher-order MA model. While we have no information to discriminate between those two, researchers in these situations should typically prefer the more parsimonious SAR model.
  - 29 As an alternative, one could estimate both **SAC** and **SDEM**. If one rejects  $\lambda = 0$  in both models and  $\rho = 0$  and  $\theta = 0$  in the **SAC** and **SDEM** models, respectively, it is likely that all three sources of clustering in the outcome are present. Power concerns make it more difficult to interpret the other combinations of possible results.
  - 30 Interested readers can also find an empirical application in Cook et al. (2015), an earlier version of this paper.

## REFERENCES

- Andrews, D. W. (2005), 'Cross-section regression with common shocks', *Econometrica* 73(5), 1551–1585.
- Anselin, L. (2001), Spatial econometrics, in B. Baltagi (ed.) *A Companion to Theoretical Econometrics*. Oxford: Blackwell, pp. 310–330.
- Anselin, L. (2003), 'Spatial externalities, spatial multipliers, and spatial econometrics', *International regional science review* 26(2), 153–166.
- Anselin, L. (2010), 'Thirty years of spatial econometrics', *Papers in regional science* 89(1), 3–25.
- Anselin, L., Bera, A. K., Florax, R. and Yoon, M. J. (1996), 'Simple diagnostic tests for spatial dependence', *Regional science and urban economics* 26(1), 77–104.
- Beck, N. (1991), 'Comparing dynamic specifications: The case of presidential approval', *Political analysis* 3(1), 51–87.
- Beck, N., Gleditsch, K. S. and Beardsley, K. (2006), 'Space is more than geography: Using spatial econometrics in the study of political economy', *International studies quarterly* 50(1), 27–44.
- Betz, T., Cook, S. J. and Hollenbach, F. M. (2020), 'Spatial interdependence and instrumental variable models', *Political science research and methods* 1–16, doi:10.1017/psrm.2018.61.
- Buhaug, H. and Gleditsch, K. S. (2008), 'Contagion or confusion? Why conflicts cluster in space1', *International studies quarterly* 52(2), 215–233.
- Burridge, P. (1981), 'Testing for a common factor in a spatial autoregression model', *Environment and planning A* 13(7), 795–800.
- Cook, S., An, S.-H. and Favero, N. (2019), 'Beyond policy diffusion: Spatial econometric models of public administration'. *Journal of public administration research and theory*, 29(4): 591–608.
- Cook, S. J., Hays, J. C. and Franzese, R. J. (2015), 'Model Specification and Spatial Interdependence'. [http://www.sas.rochester.edu/psc/polmeth/papers/Cook\\_Hays\\_Franzese.pdf](http://www.sas.rochester.edu/psc/polmeth/papers/Cook_Hays_Franzese.pdf) (Accessed on 19 November, 2019).
- Cook, S. J., Hays, J. C. and Franzese, R. J. (forthcoming, b), *Empirical Analysis of Spatial Interdependence*. Cambridge: Cambridge University Press.
- De Boef, S. and Keele, L. (2008), 'Taking time seriously', *American Journal of Political Science* 52(1), 184–200.
- Driscoll, J. C. and Kraay, A. C. (1998), 'Consistent covariance matrix estimation with spatially dependent panel data', *Review of economics and statistics* 80(4), 549–560.
- Egami, N. (2018), 'Identification of causal diffusion effects using stationary causal directed acyclic graphs', *arXiv preprint*. arXiv:1810.07858.
- Elhorst, J. P. (2010), 'Applied spatial econometrics: raising the bar', *Spatial economic analysis* 5(1), 9–28.
- Franzese, R., Hays, J. and Cook, S. (2016), 'Spatial- and spatiotemporal-autoregressive probit models of interdependent binary outcomes', *Political science research and methods* 4(1), 151–173.
- Franzese, R. J. and Hays, J. C. (2007), 'Spatial econometric models of cross-sectional

- interdependence in political science panel and time-series-cross-section data', *Political analysis* 15(2), 140–164.
- Franzese, R. J. and Hays, J. C. (2008), Empirical methods of spatial interdependence, in Janet M. Box-Steffensmeier, Henry E. Brady, and David Collier (eds.) *Oxford Handbook of Political Methodology*. Oxford: Oxford University Press, pp. 570–604.
- Gibbons, S. and Overman, H. G. (2012), 'Mostly pointless spatial econometrics?', *Journal of regional science* 52(2), 172–191.
- Hays, J. C. and Franzese, R. J. (2017), A comparison of the small-sample properties of several estimators for spatial-lag count-models, in R. J. Franzese (ed.) *Advances in Political Methodology*. Cheltenham: Elgar Research Collections, Edward Elgar, pp. 180–207.
- Hays, Jude C. and Kachi, Aya. (2015) Interdependent duration models in political science. In Franzese, R. J. (ed.) *Quantitative Research in Political Science (Vol. 5)*. Thousand Oaks (CA): Sage, pp. 33–62.
- Hays, J. C., Schilling, E. U. and Boehmke, F. J. (2015), 'Accounting for right censoring in interdependent duration analysis', *Political analysis* 23(3), 400–414.
- Hendry, D. F. (1995), *Dynamic Econometrics*. Oxford: Oxford University Press.
- LeSage, J. and Pace, R. K. (2009), *Introduction to Spatial Econometrics*. Boca Raton (FL): CRC press.
- Manski, C. F. (1993), 'Identification of endogenous social effects: The reflection problem', *The review of economic studies* 60(3), 531–542.
- Neumayer, E. and Plümper, T. (2016), 'W', *Political science research and methods* 4(1), 175–193.
- Plümper, T. and Neumayer, E. (2010), 'Model specification in the analysis of spatial dependence', *European journal of political research* 49(3), 418–442.
- Vuong, Q. H. (1989), 'Likelihood ratio tests for model selection and non-nested hypotheses', *Econometrica*, 57(2): 307–333.
- Wilhelm, S. and de Matos, M. G. (2013), 'Estimating spatial probit models in r.', *The R journal* 5(1): 130–143.

# Instrumental Variables: From Structural Equation Models to Design-Based Causal Inference

Christopher L. Carter and Thad Dunning

## INTRODUCTION

Instrumental-variables (IV) analysis bridges structural equation modeling and design-based methods for causal inference. Early studies employed instrumental variables to overcome the endogeneity of price and quantity in a structural system of supply and demand curves; by finding a third variable that was correlated with the supply but not the demand of a good (or vice versa), scholars sought to map how supply and demand respond to changes in prices. More recently, researchers have used insights from IV analysis to estimate ‘complier average causal effects’ (CACEs) in randomized experiments—that is, effects for units that comply with their assignment to receive a given treatment. Thus, the use of IV spans observational and experimental research.

In both kinds of applications, IV analysis appeals to an assumption of random or as-if random assignment of units to causal conditions. In structural equation modeling,

randomization implies statistical independence of the causal variable(s) and the error term in a regression model – that is, exogeneity. In the observational world, such assignment occurs naturally or is otherwise out of the control of the researcher – often thereby raising concerns about whether treatment assignment is really as-good-as random. When plausible, however, this assumption allows researchers to obviate concerns about confounding variables that complicate drawing causal inferences from observational data. An IV approach thus promises to marry the realism and macro focus of observational research to the rigor of experimental methods.

Yet, in observational and experimental work alike, (as-if) random assignment alone does not guarantee valid causal inference under the IV framework. Other assumptions are also needed, and these often cannot be fully tested from the data. Furthermore, the assumptions invoked by structural equation models that are fit to observational data (e.g., supply and demand curves) carry different

weight and meaning, as compared to those required to estimate CACEs in experiments.

We draw attention to these points of overlap and divergence between different usages of instrumental variables in this chapter. We begin with a discussion of early IV work in the structural equation modeling (SEM) framework, highlighting the key assumptions and potential places where they may break down. We then discuss applications of instrumental variables to design-based (often experimental) research under a potential outcomes model. We detail similarities and differences in the assumptions that these two types of applications entail. A key distinction involves the stipulation of a linear response schedule with constant effects across units in the SEM framework. Relaxing this assumption under the potential outcomes framework allows for clear definition of heterogeneous unit-level causal effects, which proves particularly important in experiments with non-compliance. Moreover, the potential outcomes framework disaggregates and clarifies other key assumptions often left implicit in the SEM framework. Yet, both approaches face important challenges in generalizing effects beyond the variation induced by a particular instrument. In a final section, we illustrate these points by comparing two different IV strategies – one observational and the other experimental – for investigating the effect of price changes on demand for coffee.

## IV IN STRUCTURAL EQUATION MODELS

The use of instrumental variables originated in simultaneous equation models, in which researchers sought to estimate supply and demand curves from equilibrium values of price and quantity (Angrist and Krueger, 2001; Stock and Trebbi, 2003: 179). Because supply and demand curves map how quantity supplied and demanded responds to changes in prices, they can be considered ‘response

schedules’ or ‘structural equations’, where the regression of quantity,  $Q$ , on price,  $P$ , carries a causal interpretation (Freedman, 2009; Imbens, 2014, 9).<sup>1</sup> A researcher may stipulate, for instance, that demand is determined according to

$$Q_t = \beta_0 + \beta_1 P_t + \beta_2 X_t + \gamma_t, \quad (1)$$

where  $Q_t$  is the quantity of a product demanded at time  $t$ ,  $P_t$  is its price,  $X_t$  is a matrix of exogenous variables affecting demand, and  $\gamma_t$  is a random error (disturbance) term.

A challenge for estimating equation (1), however, is that the quantity of the good supplied is also a function of  $P_t$ . Suppose the supply curve is given by

$$Q_t = \beta_3 + \beta_4 P_t + \beta_5 Z_t + \gamma_t, \quad (2)$$

where  $Z_t$  is a matrix of variables affecting supply.<sup>2</sup> Were the supply curve to remain fixed while the demand curve shifted, data on equilibrium levels of price and output could allow a researcher to trace out the demand equation. Yet, both curves may shift as a function of shared market conditions. In an early analysis of the impact of tariffs in markets for butter and flaxseed, the mathematician and economist Philip G. Wright noted this problem: ‘If both supply and demand conditions change, price-output data yield no information as to either curve. Unfortunately ... [this case] is the more common’ (Wright, 1928: 296). Indeed, if  $X_t = Z_t$  in equations (1) and (2) – that is, the same variables affect the quantity of the good demanded and supplied – then data on quantities and prices cannot uniquely identify the supply and demand curves.

Wright (1928) proposed an initial solution to this problem by using variables that affected supply without independently shaping demand (and vice versa).<sup>3</sup> When such variables can be found, the columns of the matrix  $X_t$  in equation (1) are not identical to the columns of  $Z_t$  in equation (2). Using what came to be called ‘instrumental variables’ – that is, variables in  $X_t$  that are excluded from  $Z_t$ , and vice versa – Wright determined the elasticity of the

supply (and demand) functions of flaxseed. One instrument Wright used to estimate supply elasticity was the price of a flaxseed substitute, cottonseed. This example already suggests difficulties in finding viable instrumental variables, however: shocks to substitutes might affect not only the demand for, but also the supply of flaxseed, perhaps because producers anticipate shifts in the demand curve.

Estimating equations (1) and (2) raises related difficulties. Manipulation of the price of the good affects quantity in both equations: supply and demand are jointly determined within a system of structural equations. Moreover, unmeasured variables that affect the quantity of demand may also affect supply, resulting in endogeneity – that is, correlation between disturbances and an explanatory variable (Freedman, 2006: 699; Freedman, 2009; Imbens, 2014: 9). In that case, the Ordinary Least Squares (OLS) estimate of  $\beta_1$  in equation (1) is biased by  $(P'P)^{-1}P'E(\gamma|P)$ , when  $E(\gamma|P) \neq 0$  (Freedman, 2009: 181). Yet, as long as cottonseed is correlated with flaxseed price but uncorrelated with the disturbance term from the demand equation, IV analysis can provide a consistent estimator of demand elasticity (Angrist and Krueger, 2001: 70).

Wright's work went largely unnoticed and played little role in the development of the IV method in econometrics.<sup>4</sup> In fact, there was no further work on instrumental variables until the 1940s, when Reiersøl's (1945) dissertation demonstrated that model parameters can be identified using the additional 'instrumental set of variables' (Angrist and Krueger, 2001; Morgan, 1990; Aldrich, 1993). Building on the further work of Geary (1949) and Durbin (1954), Sargan (1958) demonstrated the consistency of the IV estimator. Wald (1940) had previously shown the consistency of an equivalent 'grouping' estimator.<sup>5</sup>

While some of this early research sought to address measurement error in independent variables, the IV framework has gained its most prominent use in addressing the problem of omitted variable bias. A researcher interested in a causal effect of an explanatory

variable  $X_i$  on an outcome variable  $Y_i$  may stipulate the response schedule,

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i. \quad (3)$$

If unmeasured variables not included in equation (3) are correlated with the explanatory variable, such that  $\epsilon_i$  and  $X_i$  are statistically dependent, OLS will yield biased and inconsistent estimates of  $\beta_1$ . However, a third variable,  $Z_i$ , that is correlated with  $X_i$  but not  $\epsilon_i$ , offers a way to identify  $\beta_1$ . Specifically, consistent estimation of  $\beta_1$  can be obtained from the 'first-stage' regression of  $X_i$  on  $Z_i$  and then a second-stage regression of  $Y_i$  on the fitted values of  $X_i$ , or  $\hat{X}_i$ , from the first stage.<sup>6</sup> In matrix notation, this 'two-stage least squares' (2SLS) estimation of  $\beta_1$  can be written as

$$\hat{\beta}_{1,2SLS} = (\hat{X}'\hat{X})^{-1}\hat{X}'Y, \quad (4)$$

where  $\hat{X} = Z(Z'Z)^{-1}Z'X$ .

Other derivations of the multivariate IV least squares (IVLS) estimator can be rearranged to show their equivalence with equation (4) (Freedman, 2009: 178–9). In the bivariate model in equation (3), equation (4) is equivalent to dividing the regression coefficient of the 'reduced-form' regression of  $Y_i$  on a variable  $Z_i$  by the regression coefficient obtained in the first-stage regression of  $X_i$  on  $Z_i$ .<sup>7</sup>

$$\hat{\beta}_{1,2SLS} = \frac{\widehat{\text{Cov}}(Y_i, Z_i)}{\widehat{\text{Cov}}(X_i, Z_i)} = \frac{\widehat{\text{Cov}}(Y_i, Z_i)}{\widehat{\text{Cov}}(X_i, Z_i)} \cdot \frac{\widehat{\text{Cov}}(X_i, Z_i)}{\widehat{\text{Cov}}(X_i, Z_i)}. \quad (5)$$

We return to equation (5) in the next section on IV analysis in design-based inference, where the reduced-form regression of  $Y_i$  on  $Z_i$  is referred to as 'intent-to-treat' (ITT) analysis.

IV analysis requires several crucial assumptions for consistent estimation of  $\beta_1$  by the method in equations (4) or (5). Some of these assumptions are mechanical, meaning that the calculation of the 2SLS estimator requires

them to be true. To solve equation (4), the number of units must be at least as large as the number of independent variables (i.e.,  $n > q \geq p$  where  $n$  is the number of observations,  $q$  is the number of columns in  $Z$  and  $p$  is the number of columns in  $X$ ); and  $Z'X$  and  $Z'Z$  must have full rank of  $p$  and  $q$  respectively.<sup>8</sup>

Additionally, we require in practice a sufficiently strong covariance between  $X_i$  and  $Z_i$ : the so-called ‘weak instrument’ problem exacerbates finite-sample bias in the IV estimator.<sup>9</sup> Indeed, in finite samples with instruments that are only weakly related to the endogenous regressors, the asymptotic unbiasedness of the 2SLS estimator in a hypothetical, infinitely large sample – i.e., its consistency – may be of limited practical utility (Staiger and Stock, 1997). We can diagnose weak instruments by examining the relationship between  $X_i$  and  $Z_i$  in the data; as a rule of thumb, F-statistics of less than 10 indicate a weak instrument (Staiger and Stock, 1997).

Other key assumptions of structural equation models are more difficult or impossible to test. Each assumption merits careful consideration in applications of the method. First, and perhaps most fundamentally, valid IV analysis of structural equation models requires that the data were generated according to the posited response schedule – that is, a regression model such as equation (3):

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i.$$

Because the model stipulates the effect of hypothetical interventions to alter values of  $X_i$ ,  $\beta_1$  is said to carry a causal interpretation: it is the causal effect of  $X_i$  on  $Y_i$ . However, in observational studies – by definition – no researcher intervened in the system to manipulate the value of  $X_i$  (Freedman, 2009). Whether the model captures what would happen if, say, a researcher varied  $X_i$  experimentally is usually a matter of conjecture. We return to this idea of invariance to manipulation in our discussion of design-based inference.

The stipulation of this model embeds several auxiliary postulates, with specific implications

for IV estimation. First – as often noted in methodological discussions of IV analysis, and as sometimes discussed in applications as well – the assumption that the response schedule is correctly specified implies an ‘exclusion restriction’. That is, the instrument is *excluded* from equation (3). Thus,  $Z_i$  does not have a ‘direct’ effect on  $Y_i$ : it does not itself belong in the response schedule and, if it is related to  $Y_i$ , it is only through its effect on  $X_i$ . We refer here to the exclusion restriction, although some scholars use this term to refer to the combination of this assumption and the independence of  $Z_i$  and  $\epsilon_i$ ; we treat the latter as a distinct assumption (Angrist and Pischke, 2008: 117).<sup>10</sup> Additional collection of qualitative and quantitative data can help to rule out plausible alternative channels through which  $Z_i$  might have a direct effect on  $Y_i$ . Yet, for reasons we discuss further below, convincingly demonstrating that the instrument only affects the outcome through the endogenous regressor of interest raises considerable difficulties.

In addition, the structural model critically implies a set of linearity and constancy assumptions. Equation (3) stipulates that the response schedule is linear in the parameter  $\beta_1$ : thus, the effect is proportional to the value of  $X_i$ . In addition, for each unit  $i$ , the response  $Y_i$  only depends on the value of the regressor  $X_i$ : the exposure to this treatment of other units  $j \neq i$  is irrelevant. This is an analogue to the ‘non-interference’ assumption, a component of the ‘stable-unit treatment value assumption’ (SUTVA), in the context of design-based inference under the potential outcomes model. Thus, for each unit  $i$ , the treatment effect is constant, in the sense that it does not depend on the treatment assignment of other units, which might be compromised by communication or learning from other subjects in a study. The response schedule also presumes a treatment effect that is constant *across* all units  $i$ :  $\beta_1$  is the same for every unit in the study. We further discuss these assumptions of linear and constant effects across units (and contrast it to the assumption of idiosyncratic unit effects in the potential outcomes framework) in the next section.<sup>11</sup>



Finally, models such as equation (3) assume a different kind of constancy assumption: effects are constant (or homogenous) across components of  $X$ . This assumption has received somewhat less attention yet is critical for understanding the leverage that IV analysis may – or may not – provide. Imagine a researcher who is interested in the effect of income ( $X_i$ ) on attitudes toward taxation ( $Y_i$ ). Among participants in a lottery, lottery winnings ( $Z_i$ ) can be used as an instrument for income. Income ( $X_i$ ) is the sum of winnings from the lottery and income from other sources ('earned income'): call these  $X_{1i}$  and  $X_{2i}$ . The model in equation (3) assumes that the effects of these two components of  $X_i$  are the same. If this is not the case, then, in calculating a 2SLS estimate of  $\beta_1$ , we are getting the effect of a particular type of income shock – specifically, windfall gains  $X_{1i}$  (Dunning, 2008). We are not getting the effect of an increase to earned income  $X_{2i}$ . Perhaps, then, the model we should be considering is, in fact,  $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i$ . If  $\beta_1 \neq \beta_2$  in this equation, then assuming a constant effect of  $\beta_1$  in equation (3) is misleading. However, we cannot estimate the model with  $X_{1i}$  and  $X_{2i}$  without another instrument for  $X_{2i}$ . Were we to find such an instrument, concerns about the assumption that  $\beta_2$  is constant across components of  $X_{2i}$  may arise. Dunning (2008) gives additional examples where possible heterogeneous partial effects in IV analyses raise concerns about model specification. One way to reduce concerns of heterogeneous partial effects may be to define concepts more precisely a priori and limit causal claims to those aspects of a general concept that are actually measured through the IV analysis.<sup>12</sup> The point is that the stipulation of the response schedule is a key consideration for IV analysis.

If the many modeling assumptions hold – but  $X_i$  is endogenous, or statistically dependent on  $\epsilon_i$  – and there exists an instrumental variable such that

$$Z_i \perp \epsilon_i, \quad (6)$$

where  $\perp$  is read as 'is independent of', then the IV estimator in equation (5) consistently estimates  $\beta_1$  from equation (3). This too is a matter of model specification: like the exclusion restriction, statistical independence of the instrument and the disturbance term implies that  $Z_i$  does not belong in the response schedule. If it did, then the response schedule in equation (3) would be incorrectly specified. Given the model, however, random assignment of values of the instrument may imply  $Z_i \perp \epsilon_i$ . The assumption could also hold in a natural experiment where treatment is merely 'as-if' randomly assigned. Yet, the burden is then on the researcher to demonstrate why  $Z_i$  might be plausibly uncorrelated in expectation with pre-treatment causes of  $X_i$  and  $Y_i$ . Sovey and Green (2011) and Dunning (2012), among others, discuss tests that can be used to assess the validity of this assumption.

Relative to the design-based approach discussed next, the SEM framework adds additional complications for assessing the assumption in equation (6). Researchers often use multivariate IVLS regression – thus, the matrix form of the estimator in equation (4). They tend to focus on a single endogenous regressor and on whether a single instrumental variable is as good as randomly assigned; they also include putatively 'exogenous' columns of the matrix  $X$  in the matrix of independent variables,  $Z$ . Little attention is typically paid to assessing the assumption that those other columns of  $Z$  are exogenous – that is, as good as randomly assigned – as required for valid estimation by the 2SLS estimator for multiple regression. We return later to discussing these assumptions, after introducing the use of IV analysis in design-based analysis under the potential outcomes framework.

## IV ANALYSIS IN DESIGN-BASED CAUSAL INFERENCE

The rise of experimental social science has provided a new use for instrumental variables

as a tool for estimating complier average causal effects. When conducting an experiment, researchers randomly assign units to treatment or control conditions. Interest is often in estimating the average causal effect (ACE) for the group of units included in the experimental study (referred to variously as the ‘study group’ or ‘experimental population’). Scholars typically stipulate the Neyman potential outcomes model, also called the Neyman–Rubin–Holland model (Splawa-Neyman et al., 1990; Rubin, 1974; Holland, 1986). According to this model, each unit has a potential outcome under treatment,  $Y_i(1)$  – that is, the outcome that would materialize if it were assigned to the treatment group – and another potential outcome  $Y_i(0)$  that would materialize if it were assigned to control. The two potential outcomes cannot be simultaneously observed for the same unit, because a unit assigned to the treatment group cannot be assigned to control; this is the ‘fundamental problem of causal inference’ (Holland, 1986). Nor can a researcher observe the average of the potential outcomes under treatment for the experimental population without losing access to the average of the potential outcomes under control. In an experiment, however, units are assigned at random to treatment and control groups. It is as if the treatment group is a random sample from the experimental population; and the control group is another random sample from the same population. The mean of the treatment sample can therefore be used to estimate the average potential outcome under treatment for all units in the study group, and the mean of the control sample similarly estimates the average potential outcome under control. The difference of the means is an unbiased estimator for the average causal effect.

This mode of inference is sometimes called ‘design-based’, because the only stochastic element in the model is the random assignment to treatment and control groups – which is controlled by the researcher as a matter of research design (Cox, 2009).<sup>13</sup> Scholars have also used the term more broadly to denote

strategies for controlling for confounding variables that depend centrally on research design – rather than on regression adjustment, as in standard SEM frameworks (Freedman, 2009; Dunning, 2012). Design-based approaches are thus sometimes contrasted with ‘model-based’ research, even though models for causal and statistical inference play a central role in both. The key difference, as we discuss in the next section, concerns the nature of the assumptions that must be made.

In design-based inference in experiments, the CACE – and IV analysis – enters the picture when some units, despite having been assigned to the treatment condition, do not actually receive the treatment. Differential take-up of treatment generates a problem of non-compliance with treatment assignment. Imagine a case where a government offers a temporary employment program to unemployed citizens; many citizens apply, far more than the program can fund. The government decides to use a lottery to decide which applicants may participate. However, not all of those selected ultimately participate. Some have already located other employment, others may have already migrated elsewhere in search of employment, and still others may simply lose interest in participating. Similarly, some of those who were not offered enrollment may ultimately participate, say, if there is non-take-up by those originally selected to participate.

With non-compliance, the difference-of-means estimator is ITT analysis: it measures the effect of assignment to the program.<sup>14</sup> The effect of treatment assignment on outcomes, such as future employment or political support for the incumbent, may be of substantial policy as well as scholarly interest. Estimating it could tell us, for example, the likely marginal returns of offering the program to additional participants. Still, the estimator does not readily measure the effect of treatment receipt – that is, actual participation in the employment program. The assigned-to-treatment and assigned-to-control groups include non-compliers; this may ‘dilute’

the effect of treatment assignment. How to estimate the effect of program participation is not immediately obvious, however. We cannot naively compare those who received treatment to those who did not: those are self-selected groups, and participants may differ from non-participants in ways other than exposure to treatment. Put differently, these self-selected groups contain distinct mixes of compliers and non-compliers, and asymmetry may confound valid inference about the effect of treatment receipt.<sup>15</sup>

Here, IV analysis can assist in the estimation of an average causal effect among compliers – the CACE. To do so, we extend the potential outcomes model to allow for non-compliance.<sup>16</sup> Thus, we imagine that there are three types of subjects in the study pool: compliers, always-takers, and never-takers. Under the model, these types are fixed at the level of the subject; type is not affected by the assignment to levels of treatment. Compliers are those units who would receive the treatment if assigned to the treatment group – but otherwise receive the control. Always-takers receive the treatment, and never-takers receive the control, regardless of their assignment. A fourth type, defiers – who receive the treatment if assigned to the control group but receive the control if assigned to treatment – are ruled out; this assumption is required for identification of the CACE (Freedman, 2006).<sup>17</sup> The trick is then to separate the responses of compliers, always-takers, and never-takers – in order to isolate the effect of treatment assignment among compliers. At the unit level, we often cannot directly observe who is a complier and who is not, as these definitions involve counterfactuals – that is, potential outcomes (Imbens, 2014). For example, among those assigned to the control group who actually receive the control protocol, we do not observe whether they would have taken the treatment had they been assigned to the treatment group.

However, we can estimate the group-level distribution of compliance types—and the average responses by type. Imagine first that there are no always-takers: this is a situation

of ‘single crossover’ or ‘one-way non-compliance’ (Gerber and Green, 2012). In this case, we can tell which type is which among units assigned to the treatment group: the never-takers cross over to receive the control protocol while the compliers receive treatment.<sup>18</sup> Thus, we observe the average responses of the group of compliers in the assigned-to-treatment group. In the assigned-to-control group, however, the compliers and never-takers look the same: they both follow the control-group protocol.

Nonetheless, due to random assignment, we can estimate the proportion of each type in the study group. Indeed, the proportion of each type in the assigned-to-treatment group is an unbiased estimator for the corresponding proportions in the experimental population, since the treatment group is a random sample from the whole set of units in the experiment. In particular, the fraction of compliers in the treatment group – which we can observe in the case of single crossover from treatment to control – estimates the fraction of compliers in the experimental population. Moreover, the responses of never-takers in the treatment and control groups should be the same, in expectation: by assumption, treatment assignment has no effect on the response of never-takers, since they receive the control condition whether they are assigned to the treatment or the control group. Since we observe the overall response in the assigned-to-control group, and we impute the response of never-takers from the assigned-to-treatment group, we can therefore estimate the responses of the compliers in the control group. The assumption that treatment assignment does not affect the response of never-takers is akin to the exclusion restriction in the SEM framework, as we discuss further in the next section, though the potential outcomes framework helpfully clarifies the important distinction between the exclusion restriction and as-if random assignment. Together, random assignment and the exclusion restriction therefore allows us to estimate the responses of the never-takers in the assigned-to-control group – and thus the

compliers. An estimate of the CACE is just the average difference between the assigned-to-treatment and the assigned-to-control groups – that is, what ITT analysis gives us – divided by the estimated proportion of compliers in the study group.

The single-crossover model can be extended to the case of two-sided non-compliance, or ‘double crossover’ (Freedman, 2006; Dunning, 2012; Gerber and Green, 2012). In this case, we estimate the proportion of compliers by subtracting the proportion of the assigned-to-control group that actually receives the treatment from the proportion of the assigned-to-treatment group that receives treatment. Thus, when treatment assignment is a binary variable (e.g.,  $Z_i = 1$  when assigned to treatment,  $Z_i = 0$  when assigned to control), we can use the ‘Wald estimator’,

$$\hat{\beta}_{1,Wald} = \frac{\bar{Y}_1 - \bar{Y}_0}{\bar{X}_1 - \bar{X}_0}, \quad (7)$$

where  $\bar{Y}_1$  is the sample average in the assigned-to-treatment group,  $\bar{Y}_0$  is the sample average in the assigned-to-control group,  $\bar{X}_1$  is the proportion who receive treatment in the assigned-to-treatment group and  $\bar{X}_0$  is the proportion who receive treatment in the assigned-to-control group. The difference of means in the numerator of equation (7) is thus ITT analysis: it estimates the average causal effect of treatment assignment.<sup>20</sup> Note that  $\bar{X}_1$  includes both compliers and always-takers, while  $\bar{X}_0$  includes only always-takers. Effectively, the denominator subtracts off the proportion of always-takers in the control group from the joint proportion of compliers and always-takers in the treatment group. Because we expect the estimated proportion of compliers to be the same across the groups assigned to treatment and control (due to random assignment), the denominator of the Wald estimator estimates the proportion of compliers in the full study group.<sup>21</sup>

Why is equation (7) an IV estimator? Numerically, it is equivalent to a 2SLS procedure in which we regress  $Y$  on the fitted

values of  $X$ , which were obtained from a first-stage regression of  $X$  on  $Z$ . Indeed, with one treatment and one control group, and as we show in the appendix,

$$\hat{\beta}_{1,Wald} = \hat{\beta}_{1,2SLS}, \quad (8)$$

where  $\hat{\beta}_{1,2SLS}$  is given by equation (5). Conceptually, treatment assignment serves as an instrumental variable for treatment receipt in a similar sense to that developed in the previous section: it is correlated with an endogenous variable (treatment receipt), is randomly assigned, and by assumption does not influence outcomes, other than through its effect on treatment receipt. Indeed, the proof that  $\hat{\beta}_{1,Wald}$  is a consistent estimator for the CACE depends on both the randomization of the instrument and the stipulation that treatment assignment does not affect the responses of always-takers and never-takers – a kind of exclusion restriction.<sup>22</sup>

The beauty of the Wald estimator lies in its simplicity. If we have two potential treatment assignments, we can calculate an estimate for the complier average causal effect knowing only the first-stage difference in means (the numerator) and the estimated proportion of compliers in the study group (the denominator). This simplicity also rests on a model that seems reliably to capture core elements of the data-generating process: in any experiment, for example, the physical properties of the random assignment of units to treatment and control groups seem to justify the metaphor of drawing potential outcomes at random from an urn.

Nonetheless, as in all causal and statistical inference – and certainly as also in the SEM framework – design-based analysis under the potential outcomes model involves maintained hypotheses. A key assumption is the response schedule itself. The Neyman model assumes each unit has potential outcomes – in its simplest formulation, a potential outcome under control and a potential outcome under treatment. While potential outcomes are free to vary across units, they are considered fixed, deterministic properties of each

unit, and the treatment assignment of one unit does not affect the response of another. When extended to account for non-compliance, moreover, the model assumes that units are always-takers, never-takers, or compliers – but not defiers. In addition, assignment to treatment only affects outcomes for compliers; the response of always-takers and never-takers is invariant to treatment assignment. As in the SEM framework, such modeling assumptions merit careful consideration in applications of the design-based approach.

**SEM VS. DESIGN-BASED IV: A COMPARISON OF ASSUMPTIONS**

How, then, does the use of instrumental variables in the SEM framework compare to design-based approaches? The discussion so far highlights points of convergence – but also important areas of divergence. We detail similarities and contrasts in the core assumptions of the models in Table 40.1. Each row or set of rows in the table includes an assumption in the SEM framework (first column) and a corresponding or contrasting analogous assumption in the design-based approach (second column). Fundamental distinctions in the approaches

involve the assumed response schedule; the population for which key estimands are defined; stipulations on the stochastic process; and the manner of formulating validity conditions on instruments. In boldface, we indicate those assumptions that can be assessed, if at least partially, with data; we later explain the coding decisions.

First, for SEMs, the response schedule is a linear equation such as equation (3). Thus, the effect of  $X_i$  on  $Y_i$  is given by the constant of proportionality  $\beta$ . Linear structural equation models involve assumptions akin to potential outcomes because the response schedule traces out counterfactual responses at different values of  $X_i$  (Freedman, 2009). Yet, the levels of  $X_i$  (or  $Z_i$ ) are not typically directly manipulated, and linearity implies that the response surface varies smoothly as a function of  $X_i$ . By contrast, models in the Neyman tradition stipulate unit-level potential responses to two or several categorical treatment conditions. In experiments, assignment to these conditions is directly manipulated by a researcher.

Second, the response schedule under SEM also implies an assumption of a constant effect across units. By contrast, the design-based approach explicitly allows for treatment effects to vary across compliance types. This assumed unit-level heterogeneity of effects

**Table 40.1 SEM vs. design-based IV: a comparison of the assumptions**

<i>Structural equation modeling</i>		<i>Design-based IV</i>	
Linear response schedule	Linearity in parameters	Neyman potential outcomes	Unit-level potential responses to categorical treatment conditions
	Constant effect across units		Varying effect across units
	Constant effects across treatment components		Constant effect across treatment components
	Infinite (or undefined) population		<b>Finite experimental population</b>
	Random disturbance term (i.i.d)		Random sampling of potential outcomes (not i.i.d)
	$Y_i$ depends on $X_i$ , not on $X_{j \neq i}$		<b>Non-interference/SUTVA</b>
	$Z_i$ does not belong in response schedule (i.e., $Z_i \perp \square_i$ and exclusion restriction)		<b>(As-if) random treatment assignment</b>
	<b>Rank assumptions</b>		Exclusion restriction
	<b>Strength of instrument</b>		No defiers (monotonicity)
			<b>Strength of instrument</b>

Note: Bolded assumptions indicate those that can be potentially or partially tested from the data.

is useful because it readily illuminates key assumptions – for example, the idea of monotonicity, discussed momentarily – which are otherwise buried in the stipulation of common effects across units in the SEM framework (Imbens, 2014: 346). It also allows easy characterization of varying effects for specific sub-groups, including ‘local average treatment effects’ (LATE), such as the CACE.

This complier average causal effect is less readily characterized in an SEM model in which effects are presumed constant across units (Sovey and Green, 2011).

One apparent point of convergence among the SEM and design-based approaches is that both appear to stipulate constant effects across components of a treatment. Yet, the issues this raises, as we discuss later, appear less troublesome for the design-based approach than in structural equation modeling: that effects are constant across components of treatment in an experiment seems weaker and more plausible, compared to, say, the assumption that different types of income have the same effect on attitudes (Dunning, 2008).

Next, the two approaches further imply different assumptions about the population for which key estimands are defined. In the design-based approach, the target of inference is clear: it is the average of potential outcomes under treatment and control (and their difference) for the set of units in the study group – also known as the ‘experimental population’. Since this study group is typically (though not always) a convenience sample, there need be no broader population to which formal statistical inferences are drawn: the ACE is defined for the experimental population (and effects for sub-groups, such as the CACE, are similarly defined in reference to compliers in the study group).<sup>23</sup> Thus, in statistical treatments, the design-based approach is sometimes known as ‘finite population’ analysis. We code the presence of a finite experimental population as testable in Table 40.1, but, indeed, this is directly observable.

This clarity on the target population is not always present in the SEM approach. To be

sure, an equation such as (3) will be fit to data for a particular group of units; but the equation aspires to a level of generality that does not appear restricted to a particular set of data. This impression is heightened by assumptions on stochastic process. In the SEM framework, ‘Nature’ draws random disturbance terms,  $\epsilon_i$  in equation (3); in a classical regression model, these are independent and identically distributed (i.i.d). However, how ‘Nature’ draws error terms at random and with replacement, and from what broader population, is not clearly articulated. In contrast, the design-based approach assumes that potential outcomes are fixed in the particular study group at hand. Randomness enters only in the metaphor of sampling potential outcomes from an urn – that is, in sampling from this experimental population. Thus, random assignment to treatment or control groups determines which potential outcomes are observed. Moreover, these draws from the urn are not generally i.i.d: they are made without replacement, and the treatment and control samples are statistically dependent.

Both approaches require that the outcome for a given individual depends only on whether that individual received treatment – and not on the assignment of other individuals. Thus, under an SEM, such as equation (3),  $Y_i$  depends only on  $X_i$  and not on any other unit’s value of the endogenous regressor,  $X_j$ . The design-based framework makes an analogous stipulation: a unit’s potential outcomes are fixed and do not depend on the treatment receipt of any other unit. This is ‘non-interference’, or a component of what Rubin (1978) calls the Stable Unit Treatment Value Assumption (SUTVA). In the design-based framework, for example, a common concern is that units that were assigned to receive the treatment may talk with or otherwise affect units that were assigned to control. The stipulation of non-interference can be seen as an identifying restriction: if potential outcomes depend only on a given unit’s treatment receipt, but also on the treatment receipt of other units, the number of

parameters (potential outcomes) in the model multiplies quickly, and this increases the difficulty of identifying key causal parameters of interest. However, unlike in the SEM tradition, manipulation of an experimental design can provide the means to test the existence of such spillovers between treatment and control groups. For example, a researcher might assign clusters of households to the treatment and control group but then further assign individuals at random to treatment and control within the treatment households. Comparison of the responses of control individuals in the treatment and control households allows assessment of the presence of spillovers; see, for example, Nickerson (2008). For this reason, in Table 40.1, we code the non-interference assumption as potentially testable in the design-based approach.

Next, the two approaches differ in their approach to key validity assumptions on the instrument. The SEM framework assumes that the instrument  $Z_i$  does not belong in the response schedule given by equation (3). This, in turn, implies both  $Z_i$  independent of  $\epsilon_i$  – as secured by randomization of the instrument – and what we call the exclusion restriction. Yet, SEM does not clearly separate these two assumptions (Imbens, 2014: 346), while the design-based IV approach treats each assumption as distinct. An instrument needs to be, on the one hand, (as-if) randomly assigned, which allows for a causal interpretation of the first-stage regression of  $Y_i$  on  $Z_i$  (i.e., the ITT) (Angrist and Pischke, 2008: 152–153). The second assumption requires that the instrument only affects the outcome through the endogenous regressor. This exclusion restriction implies that potential outcomes for a given level of  $X_i$  do not change based on the value of  $Z_i$ .<sup>24</sup> Angrist and Pischke (2008: 153) use the example of the Vietnam draft lottery, a ‘natural’ experiment, to illustrate why these two validity assumptions should be treated as distinct. To serve in Vietnam, young men were randomly assigned a number based on their birthday; lower numbers were selected first to serve. The random assignment of draft order

fulfills the first validity assumption (i.e., statistical independence of  $Z_i$  and  $\epsilon_i$ ). Yet, being assigned a low draft number might affect the outcome (i.e., future earnings) not only through the endogenous regressor of interest (i.e., higher probability of military service), but also through other channels (e.g., enrolling in a university in hopes of getting a deferment). Compared to the stipulation that  $Z_i \perp \epsilon_i$  in SEMs, the assumption of as-if random assignment in the design-based approach can be directly, if only partially, tested. In addition to a priori knowledge or theory about the randomization process, this assumption can be assessed using balance and placebo tests, which answer the question of whether the data are consistent with randomization to treatment conditions. By contrast, in neither approach (SEM or design-based) can the exclusion restriction be directly assessed.

Note, then, that none of the assumptions of the SEM framework discussed so far can be readily tested from data. Others, however, must be true in order to calculate the 2SLS estimator. We refer to these as ‘rank assumptions’ in Table 40.1. For example, the number of units,  $n$ , must exceed the number of instruments,  $q$ , which must also be at least as large as the number of endogenous covariates,  $p$ ; also, the matrices  $Z'X$  and  $Z'Z$  must be full rank,  $p$  and  $q$ , respectively. In the first section, we referred to these assumptions as ‘mechanical’: given particular matrices  $X$  and  $Z$ , they can be readily tested. We therefore put this item in boldface in Table 40.1. More deeply, however, the rank of the matrices also reflects substantive modeling decisions – such as the exclusion of covariates that might otherwise be included in  $X$  or  $Z$  but cannot because if they were, the number of independent variables would outstrip the number of observations. Thus, identification is accomplished through model specification. As Freedman (2009: 144) puts it, ‘Many statisticians frown on under-identified models: if a parameter is not identifiable, two or more values are indistinguishable, no matter how much data you have. On the

other hand, most applied problems *are* under identified. Identification is achieved only by imposing somewhat arbitrary assumptions’.

In the design-based approach, the no-defiers (monotonicity) assumption can similarly be seen as an identification restriction. Defiers are those who receive the opposite treatment from the one they were assigned: that is, they receive the control if assigned to treatment, and the treatment if assigned to control. If there exist defiers, the relationship between treatment assignment and treatment receipt is non-monotonic. The existence of both defiers and compliers also means there are more structural parameters than we can estimate from the data. While we can estimate the proportions of compliers – using data on the proportions of never-takers (i.e., non-compliers in the treatment group) and always-takers (i.e., non-compliers in the control group) – we cannot estimate the proportion of defiers. Thus, if there are indeed defiers, the IV model will be under-identified (Freedman, 2006: 706). In that case, the Wald estimator in equation (7) does not consistently estimate the complier average causal effect. The no-defier condition is not directly testable, so we do not bold it in Table 40.1. Nonetheless, the assumption is often viewed as one of the more plausible in design-based IV applications (Freedman, 2006: 700). Certainly, when defiers constitute a very small proportion of the study group, identification and estimation issues from violations of the monotonicity assumption should be limited (Angrist et al., 1996, 451), and certain designs allow for us to dismiss the monotonicity assumption entirely. In cases of one-sided non-compliance, where researchers (or governments, nature, etc.) prevent the control group from having access to the treatment, there are, by construction, neither always-takers nor, more importantly, defiers.

Finally, both the SEM and design-based approaches require a sufficiently strong relationship between the instrument(s) and endogenous regressor(s). Because weak instruments explain little of the systematic

variation in  $X$ , the predicted values of  $X$ , that is  $\widehat{X}$ , approach  $X$ . The 2SLS estimator in equation (4) is thus biased in the same direction as the OLS estimator (Bound et al., 1995). In both IV approaches, this assumption can be tested directly from the data by examining the strength of the relationship between  $X$  and  $Z$  – see the discussion in our first section.

Overall, the discussion in this section suggests several conclusions. First, design-based approaches to IV tend to be more modest in terms of the underlying assumptions. The potential outcomes framework relaxes certain assumptions stipulated in the linear response schedule under SEM (e.g., linearity in parameters, constant effects across units). Moreover, the target of inference – the average causal effect for a particular study group or the average effect for a sub-group of compliers – is readily characterized and estimated; the model does not presume to extrapolate those effects to units outside the experimental population. Next, while many IV assumptions under SEM remain implicit in the assumption of the response schedule, IV analysis under the potential outcomes framework does a clearer job of disaggregating the key assumptions. Finally – as indicated by the greater number of bolded items in the second column of Table 40.1 – the assumptions of design-based analyses tend to be more directly testable, for example, by assessing balance on pre-treatment covariates across treatment and control groups or through modification of design. The next section illustrates these points through an empirical example.

## **AN ILLUSTRATION: THE DEMAND FOR COFFEE**

How do changes in prices affect demand for coffee? The question recalls those motivating Wright’s original work on IV, in which a key issue is the identification of the demand curve for agricultural goods. Yet, one could also approach this question experimentally, by



randomly assigning prices to coffee products and assessing how the demand changes in response. Here, we therefore describe two different approaches to answering this question: one in the SEM tradition, another in the design-based framework. The example further illustrates tradeoffs and limitations, as well as areas of convergence between the approaches.

Thus, as in Wright (1928), one option for studying this relationship would be to find an instrument that affects demand but not supply. A researcher might seek to use, say, rainfall as an instrument for coffee prices. Researchers have used rainfall (or deviations from average rainfall) as an instrument in a variety of settings, for example, estimating the effects of economic growth on dependent variables including civil war in Africa (Miguel et al., 2004) and land invasions in Brazil (Hidalgo et al., 2010). Scholars have also increasingly used rainfall to estimate the effects of turnout on support for certain political parties in the United States (Hansford and Gomez, 2010; Horiuchi and Kang, 2018; Fujiwara et al., 2016), Germany (Arnold and Freier, 2016), and Spain (Artés, 2014).

Imagine, then, a researcher wants to use changes in rainfall patterns in Uganda to instrument for changes to world coffee prices. Ultimately, she wants to test whether higher prices reduce the demand for coffee. Because Ugandan rainfall should only affect coffee demand through its effect on coffee supply and thus prices, the researcher thinks it might be a valid instrument. If the researcher were to use rainfall as an instrument to estimate a model in the form of equation (3), what assumptions must be met for a causal interpretation?

The key stipulations are found in the first column of Table 40.1. Each might raise concerns in this example. We mention only several. The demand schedule might be non-linear – that is, a demand ‘curve’ – rather than proportional to coffee prices. The elasticity of demand might vary across units, as a function of the availability of substitutes (say, tea), violating the assumption of constant

effects across units, and the assumption of constant partial effects – according to which the treatment effect does not vary across the components of the endogenous regressor – might especially suggest issues. In this case, prices can change in response to a variety of events, for example, changes induced by variation in weather patterns may have very different effects than price changes induced by a merger between two large coffee producers.<sup>25</sup> Thus, in the rainfall example, we may not be identifying the effect of price changes generally but, rather, price changes induced by a particular impetus – rainfall. Additionally, the assumption that demand in unit  $i$  does not depend on the exposure to coffee prices in unit  $j$  may also be suspect: in an interdependent world economy, spillover is perhaps much more common than non-interference.

Next, concerns might focus on the validity assumptions on the instrument – specifically, that rainfall does not belong in the response schedule. These validity assumptions entail that rainfall is independent of the disturbance term in the response schedule linking price changes to demand and that rainfall only affects coffee demand through price changes. Researchers may be able to argue that changes to rainfall in Uganda are as good as randomly assigned, adding credibility to the assumption of  $Z_i \perp \epsilon_i$ . However, the exclusion restriction assumption is considerably more difficult to test. Rainfall may change demand for coffee through channels other than supply and, thus, price. Angrist and Krueger (2001: 79) discuss the possibility that ‘sophisticated commercial buyers at the New York Coffee, Sugar and Coca Exchange, where coffee futures are traded, use weather data to adjust holdings in anticipation of price increases that may not materialize in fact’.<sup>26</sup>

This last point raises a final, empirical question regarding the strength of the instrument: is a change in rainfall patterns in Uganda enough to change world coffee prices? We could test this assumption using data from rainfall in Uganda and world coffee prices. An F-statistic greater than 10

in the regression of world coffee prices on Ugandan rainfall may suggest the latter is a sufficiently strong instrument. Yet, this rule-of-thumb has weaknesses. The ideal range of rainfall for coffee production is 45–70 inches per year; less than 30 inches is considered too dry, while more than 100 inches is considered too wet for coffee to successfully grow (Shaw, 1955: 278). Thus, we might expect the relationship between rainfall in Uganda and coffee prices to be U-shaped. The first-stage regression in equation (4) is the linear projection of  $X$  onto  $Z$ , however. While non-monotonicity in the relationship between rainfall and prices may not affect our interpretation of  $\beta$  (Imbens, 2014, 346) – given the modeling assumption of a constant effect – estimating a linear first stage for data that is non-linear may lead us to dismiss an instrument as weak even if it, in fact, has a strong relationship to the endogenous regressor.

The example of rainfall-induced variation in coffee prices thus illustrates several challenges of IV analysis under the SEM framework. Particularly troublesome is the assumption of correct specification of the linear response schedule, which gives rise to a number of other assumptions that must be carefully addressed but often cannot be directly tested – and which are not very clearly illuminated by the model. The design-based approach may provide a way of addressing some of these concerns a priori. Through robust experimental designs, researchers can attempt to reduce many of the issues that arise in the SEM framework. This approach has its own limitations and unverifiable assumptions. A feature and perhaps virtue of this approach, however, is that limited generalizability is baked into the estimand rather than obfuscated by an apparently general structural equation.

Consider, then, the direct experimental manipulation of coffee prices. Drawing partially on the innovative study of Hainmueller et al. (2014), we imagine a case in which researchers would like to work with supermarkets to manipulate coffee prices and

ultimately derive the price elasticity of demand for coffee.<sup>27</sup> The hypothetical researchers work with 100 supermarkets. Managers from 50 of the supermarkets are told to raise their prices while managers from the other 50 are told to hold their prices constant. The researchers track coffee sales in the 100 grocery stores for four weeks and then compare how sales changed across the treatment and control groups. How might the researchers analyze their experiment?

One option is ITT analysis: coffee sales in the 50 supermarkets that were told to raise prices may be compared with the 50 that were told to keep their prices constant. This approach takes advantage of the element of the design over which the researcher had most control – the initial randomization of units to treatment and control conditions. Moreover, it can be analyzed with a simple and transparent difference-in-means estimator that relies on relatively weak assumptions. In cases where researchers can engage in extensive monitoring to reduce non-compliance, as Hainmueller et al. (2014) do, this strategy provides a robust manner of estimating the treatment effect of interest. However, when such monitoring is not present and/or non-compliance is high, researchers may wish to estimate a complier average causal effect.

What might non-compliance look like in our hypothetical price-manipulation experiment? Never-takers are defined as those stores that never raise prices; always-takers are those that always raise prices; and defiers are those that raise prices when told not to raise and that do not raise when told to raise. The CACE would thus constitute the treatment effect for stores that would raise their coffee prices when instructed to and would not raise their prices otherwise. Under the potential outcomes framework, we could estimate a CACE using the ratio in equation (7). What assumptions are required?

A first assumption in valid estimation of the CACE is that the instrument is (as good as) randomly assigned. In the hypothetical case here, researchers controlled

random assignment, which they can subsequently check using balance tests.<sup>28</sup> A second assumption is that treatment assignment affects the outcome only through treatment receipt. In this case, does telling a supermarket manager to raise coffee prices affect coffee sales other than through the actual increase in coffee prices? Perhaps telling a store to raise coffee prices will result in the store not only increasing coffee prices, but also lowering tea prices. Or perhaps the price increase will lead managers to change the placement of merchandise, such that lower priced coffee is moved to occupy a more visible place in the store. The researchers may send monitors to check whether coffee prices are, in fact, being changed, which serves as both a test of the instruments' strength and also a measure of compliance. However, it would be difficult, if not impossible, for the researchers to account for all the changes made in response to the coffee price change announcement that might also affect coffee sales. Finally, we assume absence of defiers, or monotonicity. Is it plausible that there exist stores in the sample that raise prices when told not to and do not raise prices when told to? It seems unlikely in this case, although we might imagine some managers might look to defy an outside researcher who tells them how to control prices in their own store.

Both implicit and explicit in the discussion above are a number of tradeoffs regarding IV under the SEM and design-based approaches. In the SEM framework, many of the assumptions are embedded in the response schedule itself. Researchers should carefully justify ex-post the correct specification of the response schedule, including the assumptions that follow from it—but they may fail to do so. The design-based approach under potential outcomes allows for a weakening of the constant unit-level effects assumption but clarifies a new assumption – monotonicity.<sup>29</sup> This approach also draws on the potential outcomes framework, which relaxes the assumption of the linear response schedule and disaggregates

the assumptions otherwise implied by the model into distinct parts. Concerns about these assumptions are addressed through careful research design; natural and field experiments provide a way to find or develop robust instruments that are plausibly exogenous and that only affect the outcome through the endogenous regressor. The design-based approach thus has a number of desirable properties which include both the clear statement of and the possibility to test key assumptions.

A key limitation to both approaches, however, involves generalizability of interpretation, which the coffee demand example illustrates well. Often, IV analysis involves an intervention that addresses only one component of the treatment of interest. For example, it is hard to know exactly what the price manipulation experiment tells us about the relationship between coffee price and demand. Artificial manipulation of a coffee price may truly isolate the general effect of interest, but more likely, it tells us about only one component of a 'price' treatment – the actual changing of prices by a store. This change occurs at the end of the supply chain and tells us little about the result of price changes due to farming, tariffs, increased fuel prices, etc. The same can be said of using rainfall as an instrument for coffee price changes. However, in the SEM model, this claim is particularly muddled by the specification of the linear model, where the researcher is claiming to identify through  $X_i$  in equation (3) the effect of 'prices'. In reality, and as discussed above, price changes induced by rainfall may have a very different effect than price changes induced by other 'interventions'. From this perspective, an advantage of the design-based approach is then not just the clarity and relative testability of the key assumptions – but that the framework makes clear its limitations. Here, modesty is a virtue: the SEM approach is subject to the same kinds of weaknesses but, because of the lack of specificity embedded in the model specifications, it tends to overstate its ability to deliver on its ambitions.<sup>30</sup>

## CONCLUSION

Since Wright's initial work on supply and demand, social scientists have used instrumental variables to study the effects of independent variables that would otherwise be difficult (if not impossible) for the researcher to randomly assign. Some instruments, like rainfall, rely on plausibly exogenous natural variation that affects an endogenous independent variable of interest. Ramsay (2011) studies as-if random variation in natural disasters to understand how a country's level of democracy responds to a change in oil prices. Other instruments rely on lotteries, where the instrument is – due to actual randomization – independent of pre-treatment causes of  $X$  and  $Y$ . Researchers have used the Vietnam draft lottery as an instrument for military service (Angrist, 1990; Erikson and Stoker, 2011) and lottery winnings as an instrument for income (Doherty et al., 2006). While these cases assure that the instrument,  $Z$ , is – in expectation – uncorrelated with the disturbance term from the regression, there remain important concerns about both violations of the exclusion restriction and the correct specification of the response schedule.

In observational work, challenges often arise for validating the identifying assumptions, which researchers have shown varying degrees of willingness to acknowledge and address (Sovey and Green, 2011: 194; Staiger and Stock, 1997: 597, fn. 2). While certain assumptions can be directly tested from the data (relevance, rank, identifiability of parameters) and others may be plausible by design (independence of instrument and pre-treatment causes of  $X$ ,  $Y$ ), the remaining assumptions of structural equation modeling generally raise concerns that often cannot be fully allayed. The exclusion restriction and specification of the response schedule remain particularly troublesome.

IV analysis in the structural equation modeling framework offers a potential solution to

a key problem: identifying the causal effect of  $X$  on  $Y$  given the stipulation of a particular structural equation in which the model is presumed but  $X$  is assumed to be endogenous. However, the SEM framework may also prove restrictive; it imposes an assumption of a linear response schedule that does not allow for estimation of heterogeneous treatment effects. It further builds monotonicity into the estimand rather than addressing it as an assumption (Imbens, 2014: 346).

The potential outcomes framework overcomes some of these limitations by making more explicit the key underlying assumptions, which are often more plausible, less restrictive, and easier to test from the data. Linearity and constant effects assumptions are relaxed under this approach, and other assumptions, like the exclusion restriction and random assignment of units to values of the instrument, are directly stated, such that they might be separately addressed and evaluated. The monotonicity assumption, which is added under heterogeneous treatment effects, is generally viewed to be plausible, although its credibility should be judged based on the specific intervention.

Ultimately, despite the promise of IV, there remain key limitations. Both the SEM and design-based approaches suffer a common challenge of generating results that generalize beyond the intervention that gave rise to the instrument. Often, instruments affect the independent variable of interest through a specific (and perhaps narrow) channel. Generalizing from that particular component of the treatment to formulating broader claims about the treatment as a whole should be done only after consideration of other factors that may have induced change in the independent variable. Robust IV analysis thus requires that researchers consider both the story that can be told from the data given the assumptions and the story that *cannot* be told, given the inherent difficulties of generalization under the IV framework.

## Notes

- 1 We refer in this paper to 'structural equation models' in this sense. One stream of research uses the term more specifically to refer to systems of equations linking unobservable 'latent' constructs: see, for example, Bollen (1989).
- 2 We use the typical language of supply and demand 'curves' here, even though the response schedules in equations (1) and (2) are linear in  $P_i$ .
- 3 While Philip G. Wright's name is on the piece, the key finding is in Appendix B, which some believe was written by his son, Sewall. The elder Wright also discusses his son's closely related methods of causal path analysis. However, Stock and Trebbi (2003), using stylometric analysis, conclude that Philip was the most likely author.
- 4 Wright's innovation was only recognized in the 1970s when Goldberger (1972) highlighted Wright's contribution to structural equation methods (Aldrich, 1993: 270, fn. 34).
- 5 We show the equivalence of the IV estimator and Wald's grouping estimator below.
- 6 The fitted values  $\hat{X}_i$  are sometimes called 'predicted' values of  $X_i$ , though 'post-dicted' is usually more accurate. Importantly, we cannot simply use the values of  $\hat{X}$  to calculate the variance-covariance matrix of  $\hat{\beta}_1$ , as this produces inconsistent estimation of  $\sigma^2$  (Greene, 2003: 79).
- 7 A derivation can be found in the Appendix – see (A.1).
- 8 A model in which  $q = p$  is 'just-identified' while the case with  $q > p$  is 'over-identified'.
- 9 There is, of course, also a mechanical reason for this, related to the previous paragraph and the rank condition; if  $\text{Cov}(X_i, Z_i) = 0$ , then the estimator of  $\beta_1$  found in equation (5) is undefined.
- 10 Given a model like equation (3), the unconfoundedness of the instrument and the exclusion restriction are implied by  $\square_i \perp\!\!\!\perp Z_i$  (Imbens 2014).
- 11 Heckman and Robb (1986), Imbens and Angrist (1994), Angrist et al. (1996), Rosenzweig and Wolpin (2000), Freedman (2006) and Heckman et al. (2006) all draw attention to this IV assumption.
- 12 To be sure, improving conceptual precision by moving down Sartori's (1970) 'ladder of abstraction' may lessen the perceived impact of the research: a paper on the effects of windfall earnings on political attitudes may generate less interest than one that purports to estimate the effect of income more generally.
- 13 This usage of 'design-based' in statistics differs from a related but distinct use of the term in educational research.
- 14 As we discuss below, the ITT analysis is equal to the reduced-form estimate discussed above.
- 15 Relatedly, while manipulation checks can provide a useful measure of whether subjects understood or experienced the treatment in the way the researcher expected, treatment effects should not be calculated as conditional on having passed a manipulation check, because the check is necessarily post-treatment (Aronow et al., 2015; Montgomery et al., 2018).
- 16 See, *inter alia*, Angrist et al. (1996); Freedman (2006); Gerber and Green (2012); Dunning (2012); Imbens (2014).
- 17 Angrist et al. (1996) call the no-defiers assumption 'monotonicity': being assigned to treatment should never make it *less* likely that a unit actually receives treatment (see also Imbens 2014, 17).
- 18 This also assumes we can observe who receives the treatment, for example, who follows the protocol in a drug trial (which is distinct from the even harder problem of observing counterfactual compliance types).
- 20 Numerically, the value is equivalent to the reduced-form regression of  $Y$  on  $Z$ .
- 21 This is equivalent algebraically to the first-stage regression of  $X$  on  $Z$ .
- 22 Note that equation (7) suffers from ratio-estimator bias: the denominator is a random variable. However, by Slutsky's theorem, the estimator is consistent (asymptotically unbiased) – see Freedman (2006) or Dunning (2012).
- 23 Some work does nonetheless distinguish, though not always with clarity, between a sample average treatment effect (SATE) and a population average treatment effect (PATE), where the study group is itself viewed as a sample from a broader population.
- 24 Both of these assumptions, along with monotonicity and a strong instrument, are necessary for valid estimation of the CACE (Angrist and Pischke, 2008: 154)
- 25 Similar critiques have arisen around the use of rainfall as an instrument for economic growth. Dunning (2008) suggests that rainfall may induce a very specific type of economic growth that is quite different from growth induced by, for example, technological change in agriculture or an increase in foreign aid.
- 26 Research using rainfall as an instrument for economic growth and turnout has often been critiqued with respect to the exclusion restriction. In the case of Miguel et al. (2004), rainfall may lead to flooding on roads and bridges, making it difficult to transport soldiers and thus decreasing the likelihood of conflict (Sovey and Green, 2011; Dunning, 2012). Sarsons (2015) shows that the relationship between rainfall and conflict in India is strongest in areas downstream of dams, where agricultural income is less susceptible to rainfall shocks due to access to irrigation. As for research on turnout and party support, Horiuchi and Kang (2018) demonstrate that

weather directly changes voter support for parties, with rainfall making voters more likely to support Republicans. In fact, most of the benefit obtained by Republicans in rainy elections can be attributed to voters changing their preferences, rather than differential levels of turnout.

- 27 Using a randomized control trial with 26 grocery stores in New England, Hainmueller et al. (2014) manipulate both the price and labeling of coffee to understand whether consumers are willing to pay a higher price for fair trade coffee.
- 28 Unlike the analysis, which was performed using store-weeks as units, the balance tests were performed at the store level, giving a sample size of only 26, which may limit their power.
- 29 However, as we noted, the assumption of monotonicity exists implicitly in the SEM framework. Allowing for defiers would imply more structural parameters than can be estimated from the data.
- 30 For a related point in the context of fixed effects regressions, see Aronow and Samii (2016).

## APPENDIX

We derive the algebraic equivalence of the two-stage least-squares (equivalently, the IV) estimator in equation (5) and the Wald estimator of the Complier Average Causal Effect for a finite population in equation (7) in the bivariate case with one treatment and one control group. Here,  $Y_i$  is the outcome variable;  $X_i = 1$  if unit  $i$  receives treatment and otherwise  $X_i = 0$ ; and  $Z_i = 1$  if unit  $i$  is assigned to treatment and otherwise  $Z_i = 0$ . The number of units is given by  $N$ , and the number of units assigned to treatment is  $m < N$ . Without loss of generality, index the units assigned to treatment by  $i = 1, \dots, m$  and the units assigned to control by  $i = m + 1, \dots, N$ . Thus, we have

$$\begin{aligned}
 \hat{\beta}_{1,2SLS} &= \frac{\widehat{Cov}(Y_i, Z_i)}{\widehat{Cov}(X_i, Z_i)} \\
 &= \frac{\sum_{i=1}^N (Y_i - \bar{Y})(Z_i - \bar{Z})}{\sum_{i=1}^N (X_i - \bar{X})(Z_i - \bar{Z})} = \frac{\sum_{i=1}^N (Y_i Z_i) - \bar{Y} \sum_{i=1}^N Z_i - \bar{Z} \sum_{i=1}^N Y_i + N \bar{Y} \bar{Z}}{\sum_{i=1}^N (X_i Z_i) - \bar{X} \sum_{i=1}^N Z_i + \bar{Z} \sum_{i=1}^N X_i + N \bar{X} \bar{Z}} \\
 &= \frac{\sum_{i=1}^N (Y_i Z_i) - m \bar{Y}}{\sum_{i=1}^N (X_i Z_i) - m \bar{X}} \left( \text{because } \sum_{i=1}^N Z_i = m, \bar{Z} = \frac{m}{N} \right) \\
 &= \frac{m(\bar{Y}_1 - \bar{Y})}{m(\bar{X}_1 - \bar{X})} \left( \text{because } \sum_{i=1}^m Y_i Z_i = m \bar{Y}_1 \right) \\
 &= \frac{m \left( \bar{Y}_1 - \frac{m}{N} \bar{Y}_1 - \frac{N-m}{N} \bar{Y}_0 \right)}{m \left( \bar{X}_1 - \frac{m}{N} \bar{X}_1 - \frac{N-m}{N} \bar{X}_0 \right)} \left( \bar{Y} = \frac{\sum_{i=1}^m Y_i + \sum_{i=m+1}^N Y_i}{N} = \frac{m}{N} \bar{Y}_1 + \frac{N-m}{N} \bar{Y}_0 \right) \\
 &= \frac{\frac{m(N-m)}{N} (\bar{Y}_1 - \bar{Y}_0)}{\frac{m(N-m)}{N} (\bar{X}_1 - \bar{X}_0)} \\
 &= \frac{\bar{Y}_1 - \bar{Y}_0}{\bar{X}_1 - \bar{X}_0}
 \end{aligned}
 \tag{A.1}$$

The first step uses the definition of the sample covariance; we divide through by  $n / n$ . Next, we multiply out terms, then use the definition of  $Z = 1$  as the units assigned to treatment (and thus, the sum,  $\sum Z_i$ , is  $m$ , while the mean,  $\bar{Z}$ , is  $m / N$ ) and cancel terms. In the following step, we use the fact that the product,  $Y_i Z_i$ , will be zero when  $Z_i = 0$  and  $Y_i$  when  $Z_i = 1$ . The sum of this product will thus equal the mean outcome for the treated units,  $\bar{Y}_1$ , times the number of treated units,  $m$ . The next step uses the fact that the mean,  $\bar{Y}$ , is simply a weighted average of the mean outcome under treatment,  $\bar{Y}_1$ , and the mean outcome under control,  $\bar{Y}_0$ . The final steps factor out common terms and reduce the equation to the Wald estimator.

## REFERENCES

- Aldrich, J. (1993). Reiersøl, Geary and the Idea of Instrumental Variables. *Economic & Social Review*, 24(3): 247–273.
- Angrist, J. D. (1990). Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records. *The American Economic Review*, 80(3): 313–336.
- Angrist, J. D., Imbens, G. W. and Rubin, D. B. (1996). Identification of Causal Effects Using Instrumental Variables. *Journal of the American Statistical Association*, 91(434): 444–455.
- Angrist, J. D. and Krueger, A. B. (2001). Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments. *Journal of Economic Perspectives*, 15(4): 69–85.
- Angrist, J. D. and Pischke, J.-S. (2008). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton: Princeton University Press.
- Arnold, F. and Freier, R. (2016). Only Conservatives are Voting in the Rain: Evidence from German Local and State Elections. *Electoral Studies*, 41: 216–221.
- Aronow, P. M., Baron, J. and Pinson, L. (2015). *A Note on Dropping Experimental Subjects Who Fail a Manipulation Check*. SSRN Scholarly Paper ID 2683588, Social Science Research Network, Rochester, NY.
- Aronow, P. and Samii, C. (2016). Does Regression Produce Representative Estimates of Causal Effects? *American Journal of Political Science*, 60(1): 250–267.
- Artés, J. (2014). The Rain in Spain: Turnout and Partisan Voting in Spanish Elections. *European Journal of Political Economy*, 34: 126–141.
- Bollen, K. A. (1989). *Structural Equation Models with Latent Variables*. New York: John Wiley & Sons.
- Bound, J., Jaeger, D. A. and Baker, R. M. (1995). Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogenous Explanatory Variable is Weak. *Journal of the American Statistical Association*, 90(430): 443–450.
- Cox, D. (2009). Randomization in the Design of Experiments. *International Statistical Review / Revue Internationale de Statistique*, 77(3): 415–429.
- Doherty, D., Gerber, A. S. and Green, D. P. (2006). Personal Income and Attitudes toward Redistribution: A Study of Lottery Winners. *Political Psychology*, 27(3): 441–458.
- Dunning, T. (2008). Model Specification in Instrumental-Variables Regression. *Political Analysis*, 16(3): 290–302.
- Dunning, T. (2012). *Natural Experiments in the Social Sciences: A Design-Based Approach*. New York: Cambridge University Press.
- Durbin, J. (1954). Errors in Variables. *Revue de l'Institut International de Statistique / Review of the International Statistical Institute*, 22(1/3): 23–32.
- Erikson, R. S. and Stoker, L. (2011). Caught in the Draft: The Effects of Vietnam Draft Lottery Status on Political Attitudes. *The American Political Science Review*, 105(2): 221–237.
- Freedman, D. A. (2006). Statistical Models for Causation: What Inferential Leverage Do They Provide? *Evaluation Review*, 30(6): 691–713.
- Freedman, D. A. (2009). *Statistical Models: Theory and Practice*. New York: Cambridge University Press.
- Fujiwara, T., Meng, K. and Vogl, T. (2016). Habit Formation in Voting: Evidence from Rainy Elections. *American Economic Journal: Applied Economics*, 8(4): 160–188.
- Geary, R. C. (1949). Determination of Linear Relations between Systematic Parts of Variables with Errors of Observation the Variances of Which Are Unknown. *Econometrica*, 17(1): 30–58.

- Gerber, A. S. and Green, D. P. (2012). *Field Experiments: Design, Analysis, and Interpretation*. New York: WW Norton.
- Goldberger, A. S. (1972). Structural Equation Methods in the Social Sciences. *Econometrica*, 40(6): 979–1001.
- Greene, W. H. (2003). *Econometric Analysis*. Upper Saddle River: Pearson Education.
- Hainmueller, J., Hiscox, M. J. and Sequeira, S. (2014). Consumer Demand for Fair Trade: Evidence from a Multistore Field Experiment. *The Review of Economics and Statistics*, 97(2): 242–256.
- Hansford, T. G. and Gomez, B. T. (2010). Estimating the Electoral Effects of Voter Turnout. *The American Political Science Review*, 104(2): 268–288.
- Heckman, J. J. and Robb, R. (1986). Alternative Methods for Solving the Problem of Selection Bias in Evaluating the Impact of Treatments on Outcomes. In Wainer, H., editor, *Drawing Inferences from Self-Selected Samples*, pages 63–107. New York: Springer New York.
- Heckman, J. J., Urzua, S. and Vytlacil, E. (2006). Understanding Instrumental Variables in Models with Essential Heterogeneity. *The Review of Economics and Statistics*, 88(3): 389–432.
- Hidalgo, F. D., Naidu, S., Nichter, S. and Richardson, N. (2010). Economic Determinants of Land Invasions. *The Review of Economics and Statistics*, 92(3): 505–523.
- Holland, P. W. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association*, 81(396): 945–960.
- Horiuchi, Y. and Kang, W. C. (2018). Why Should the Republicans Pray for Rain? Electoral Consequences of Rainfall Revisited. *American Politics Research*, 46(5): 869–889.
- Imbens, G. W. (2014). Instrumental Variables: An Econometricians Perspective. *Statistical Science*, 29(3): 323–358.
- Imbens, G. W. and Angrist, J. D. (1994). Identification and Estimation of Local Average Treatment Effects. *Econometrica*, 62(2): 467–475.
- Miguel, E., Satyanath, S. and Sergenti, E. (2004). Economic Shocks and Civil Conflict: An Instrumental Variables Approach. *Journal of Political Economy*, 112(4): 725–753.
- Montgomery, J. M., Nyhan, B. and Torres, M. (2018). How Conditioning on Posttreatment Variables Can Ruin Your Experiment and What to Do about It. *American Journal of Political Science*, 62(3): 760–775.
- Morgan, M. S. (1990). *The History of Econometric Ideas*. New York: Cambridge University Press.
- Nickerson, D. (2008). Is Voting Contagious? Evidence from Two Field Experiments. *American Political Science Review*, 102: 49–57.
- Ramsay, K. W. (2011). Revisiting the Resource Curse: Natural Disasters, the Price of Oil, and Democracy. *International Organization*, 65(3): 507–529.
- Reiersøl, O. (1945). *Confluence Analysis by Means of Instrumental Sets of Variables*. Uppsala: Almqvist & Wiksells.
- Rosenzweig, M. R. and Wolpin, K. I. (2000). Natural 'Natural Experiments' in Economics. *Journal of Economic Literature*, 38(4): 827–874.
- Rubin, D. B. (1974). Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology*, 66(5): 688–701.
- Rubin, D. B. (1978). Bayesian Inference for Causal Effects: The Role of Randomization. *The Annals of Statistics*, 6(1): 34–58.
- Sargan, J. D. (1958). The Estimation of Economic Relationships using Instrumental Variables. *Econometrica*, 26(3): 393–415.
- Sarsons, H. (2015). Rainfall and Conflict: A Cautionary Tale. *Journal of Development Economics*, 115: 62–72.
- Sartori, G. (1970). Concept Misformation in Comparative Politics. *The American Political Science Review*, 64(4): 1033–1053.
- Shaw, E. B. (1955). *World Economic Geography: With an Emphasis on Principles*. New York: Wiley.
- Sovey, A. J. and Green, D. P. (2011). Instrumental Variables Estimation in Political Science: A Readers Guide. *American Journal of Political Science*, 55(1): 188–200.
- Splawa-Neyman, J., Dabrowska, D. M. and Speed, T. P. (1990). On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9. *Statistical Science*, 5(4): 465–472.
- Staiger, D. and Stock, J. H. (1997). Instrumental Variables Regression with Weak Instruments. *Econometrica*, 65(3): 557–586.



Stock, J. and Trebbi, F. (2003). Who Invented Instrumental Variable Regression? *Journal of Economic Perspectives*, 17: 177–197.

Wald, A. (1940). The Fitting of Straight Lines if Both Variables are Subject to Error. *The*

*Annals of Mathematical Statistics*, 11(3): 284–300.

Wright, P. G. (1928). *The Tariff on Animal and Vegetable Oils*. New York: Macmillan Company.



# Causality and Design-Based Inference

Jake Bowers and Thomas Leavitt

## DESIGN-BASED CAUSAL INFERENCE

No one knows the true causal effect of an intervention. In an experiment, a researcher can assign some units to treatment and others to control; yet, one cannot see how treatment units would have acted were they assigned to control nor how the control units would have acted were they assigned to treatment.<sup>1</sup> In the face of this fundamental ignorance, statisticians have developed two prominent approaches to inferring unobservable causal effects using data that can be observed. An analyst can either (1) generate a guess about (usually average) treatment effects or (2) posit a hypothesis about the effects of a treatment (such as the hypothesis that a treatment had no effects) and then assess the consistency of observable data with that null hypothesis, relative to a class of alternative hypotheses (such as the hypothesis that a treatment had a positive effect).

In what follows, we will define criteria by which a procedure qualifies as ‘good’ in the

context of both estimation and testing and subsequently explain the role that research design plays in whether estimators and tests satisfy these criteria. We consider estimators and tests about causal effects first in the context of a randomized study design under full control of the researcher and second in cases in which the researcher does not fully control the study design. We show the ways in which either complete knowledge or assumptions (and the ways in which they could be violated) about the study design constitute what Fisher (1935, 14) referred to as a ‘reasoned basis’ for inference.

## CAUSALITY AND RESEARCH DESIGN

### *Defining Causal Effects*

Consider a study in which there are  $1, \dots, N$  units and the index  $i \in \{1, \dots, N\}$  runs over these units. Each individual,  $i$ , can be in

either the treatment condition,  $z_i = 1$ , or the control condition,  $z_i = 0$ . Under the Stable Unit Treatment Value Assumption (SUTVA) (Cox, 1958; Rubin, 1980, 1986), each individual has a treatment potential outcome,  $y_{t,i}$  (unit  $i$ 's outcome if given the intervention), and a control potential outcome,  $y_{c,i}$  (unit  $i$ 's outcome if not given the intervention).<sup>2</sup> An individual causal effect,  $\tau_i$ , for each of the  $i \in \{1, \dots, N\}$  units is a function of each unit's two potential outcomes,  $\tau_i \equiv f(y_{c,i}, y_{t,i})$ , such as  $\tau_i \equiv \frac{y_{t,i}}{y_{c,i}}$  or  $\tau_i \equiv y_{t,i} - y_{c,i}$ . For this chapter, we focus specifically on the additive, individual causal effect defined as  $\tau_i \equiv y_{t,i} - y_{c,i}$ . Researchers, however, can never observe both potential outcomes for each unit; instead, one can observe only  $y_i$ , which can be equal to either  $y_{c,i}$  or  $y_{t,i}$ , depending on whether unit  $i$  is assigned to treatment ( $z_i = 1$ ) or control ( $z_i = 0$ ). We therefore represent observed outcomes by the function  $y_i = z_i y_{t,i} + (1 - z_i) y_{c,i}$ . Researchers may want to make an inference about the  $1, \dots, N$  individual causal effects, which we collect into the vector  $\boldsymbol{\tau}' = [\tau_1 \ \tau_2 \ \dots \ \tau_N]$ , or about a function of these individual causal effects, such as the average causal effect,

$$\bar{\tau} = \left( \frac{1}{N} \right) \sum_{i=1}^N \tau_i.$$

We say that researchers want to 'make an inference' because neither  $\boldsymbol{\tau}$  nor  $\bar{\tau}$  can be directly observed. Researchers, of course, don't simply want to 'make an inference': they want to make inferences that can reliably track the true causal quantity of interest. This chapter shows how inferential procedures based on the research design can have such a reliable relationship with true causal quantities and explains what it means for a procedure to be 'based on research design'.

### **Defining a Research Design**

Although a research design can certainly be more than this, for the purposes of this chapter

a research design refers to the process by which units come to be in one study condition instead of another, i.e., each  $z_i$  comes to equal 1 or 0. More formally, we denote the collection of the values of  $z_i$  for all  $i \in \{1, \dots, N\}$  units by the vector  $\mathbf{z}' = [z_1 \dots z_N]$  and define a research design as (1) a set of possible ways (events) in which the whole vector  $\mathbf{z}$  could occur and (2) a probability distribution on this set of possible events. In a controlled study design (i.e., an experiment), we think of a researcher as 'assigning' conditions to all units in the study. When a researcher does not control how a unit  $i$  takes on a value of  $z_i$ , we think of that unit as 'selecting' into its own condition. As we lay the groundwork of concepts and notation, we write 'assignment' and assume control by the researcher, but we will apply the general framework later to uncontrolled research designs in which units 'select' into study conditions.

The set of possible ways in which  $\mathbf{z}$  can occur depends on the process by which units are assigned to study conditions. If individuals can be in either the treatment or control condition irrespective of any other individual in the population, then we call this process individual assignment. As a simple example, consider the 'coin flip' assignment process: in this case, the proportion of  $N$  individuals in either the treatment or control condition can vary across different assignments. We refer to this process as *simple* individual assignment. A researcher can implement simple individual assignment via an actual physical, stochastic process, such as  $N$  flips of a (potentially biased) coin – although in practice researchers will typically use random number generators (RNGs).

Under completely unconstrained simple individual assignment, the number of units in the treatment condition can range from 0 to  $N$  and the number of units in the control condition can likewise range from  $N - 0$  to  $N - N$ . More formally, we write the set,  $\Omega$ , of possible ways that a researcher can

assign all individuals to study conditions as follows:

$$\Omega = \{0,1\}^N = \left\{ \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \\ 0 \end{bmatrix}, \dots, \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 1 \end{bmatrix} \right\}. \tag{1}$$

We can write the number of possible assignments in the set  $\Omega$  by  $|\Omega|$  (the ‘cardinality of Omega’), under simple assignment as follows:

$$|\Omega| = \binom{N}{0} + \binom{N}{1} + \dots + \binom{N}{N-1} + \binom{N}{N} = \sum_{n_t=0}^N \binom{N}{n_t},$$

where  $n_t = \sum_{i=1}^N z_i$  is the number of units in the treatment condition, which can range from 0 to  $N$ , and  $\binom{N}{n_t} = \frac{N!}{n_c!n_t!}$  is the number of ways to choose  $n_t$  units from a total of  $N$  units. Conversely,  $n_c = \sum_{i=1}^N (1-z_i)$  is the number of units in the control condition, which can range from  $N - 0$  to  $N - N$ . In practice, researchers who control the assignment process will typically forbid assignments in which all units are in either condition (the first and last assignments in Equation (1)), in which case  $|\Omega| = \sum_{n_t=1}^{N-1} \binom{N}{n_t}$ .

The ‘coin flipping’ design helps us introduce the formal elements of a research design. In practice, individual coin flips can lead to lopsided designs in which many units are in one condition or another. An alternative design that enables the researcher to control the numbers of units in each condition is *complete* assignment.

Complete individual assignment differs from simple, individual assignment only in that the value of  $n_t$  is fixed across all possible assignments. We have described simple individual assignment via the example of coin flips. Complete individual assignment can

be thought of as draws from an urn. Imagine, for example, that an urn contains  $N$  balls, of which  $n_t$  are blue balls and  $N - n_t = n_c$  are red balls. The researcher could draw the first ball from the urn and assign the first unit in the study to the treatment condition if the ball is blue and to the control condition if the ball is red. The second draw could follow the same rule for the assignment of the second unit, and so on and so forth until no more balls remain in the urn. This form of assignment ensures that exactly  $n_t$  units are in the treatment condition and  $N - n_t = n_c$  units are in the control condition. More formally, complete individual assignment excludes any assignment,  $\mathbf{z}$ , with more or less treatment units,  $n_t$ , than that which is predetermined by the researcher. Therefore, under complete assignment, the number of possible assignments is simply  $|\Omega| = \binom{N}{n_t}$ .

Simple and complete assignment can also happen at the *cluster* (as opposed to the *individual*) level. In this setup, we not only have a set of  $1, \dots, N$  individuals, but also a set of  $1, \dots, K$  clusters, where each cluster,  $k \in \{1, \dots, K\}$ , contains  $N_k \geq 1$  individual units and  $N = \sum_{k=1}^K N_k$ . In cluster assignment

designs, all of the units in the  $k$ th cluster are assigned to either the treatment condition,  $z_{i,k} = 1$ , or the control condition,  $z_{i,k} = 0$ . In simple cluster assignment, the number of possible assignments is given by  $|\Omega| = \sum_{k_t=0}^K \binom{K}{k_t}$ ,

where  $k_t$  denotes the number of treatment clusters, although (just as in simple individual assignment) researchers will typically ensure that  $k_t \notin \{0, K\}$ . Under complete cluster assignment, the number of treatment clusters is fixed, such that the number of assignments is  $|\Omega| = \binom{K}{k_t}$ .

Lastly, *blocked assignment* is when individuals or clusters are assigned (either simply or completely) to study conditions within

blocks, which we index from  $b \in \{1, \dots, B\}$ . Blocks are typically constructed on the basis of individuals' or clusters' values of baseline covariates. Baseline covariates are measured prior to assignment and hence their values are fixed regardless of the condition to which a unit or cluster is assigned. Under simple individual block assignment, the number of possible assignments is  $|\Omega| = \prod_{b=1}^B \left( \sum_{n_{t,b}=1}^{N_b-1} \binom{N_b}{n_{t,b}} \right)$ , where

$N_b$  is the number of units in block  $b$ ,  $n_{t,b}$  is the number of units in the treatment condition in block  $b$  and  $n_{t,b} \notin \{0, N_b\}$  for all  $b$ . Under complete individual block assignment, the number of possible assignments is  $|\Omega| = \prod_{b=1}^B \binom{N_b}{n_{t,b}}$ . One can analogously deduce the number of possible assignments under either simple or complete cluster block assignment. As we will explain in subsequent sections, block assignment carries important implications for properties of both estimators and hypothesis tests.

Given a set of possible assignments,  $\Omega$ , arising from an assignment mechanism, the remaining component of a research design is a probability distribution on this set of assignments. In a *uniform randomized experiment*, the probability of each assignment is simply  $\frac{1}{|\Omega|}$  for all  $\omega \in \Omega$ , whereby each assignment has an identical probability of realization. Yet the probability distribution on  $\Omega$  need not be uniform, even in a randomized experiment. Design-based inference means only that the stochastic properties of estimators and tests be based on this probability distribution on  $\Omega$ , regardless of whether that distribution is uniform or not. As we now move to discussions of both estimation and testing, notice throughout that whenever we refer to random quantities, the randomness of those quantities stems solely from the probability distribution on  $\Omega$ .

**An Illustrative Example**

In the sections to follow, we demonstrate our arguments via a simple hypothetical example

that consists of  $N = 6$  units and an individual assignment process (complete individual assignment) in which three units are assigned to treatment ( $n_t = 3$ ) and to control ( $n_c = 3$ ). Let's further imagine that (unbeknownst to the researcher) the six units' potential outcomes and individual causal effects are as follows in Table 41.1.

Based on the complete individual assignment process in which there are  $N = 6$  units and of which  $n_t = 3$  are assigned to treatment, the set of  $|\Omega| = \binom{6}{3} = 20$  possible assignments is given by Equation (2):

$$\Omega = \left\{ \begin{bmatrix} 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \dots, \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \end{bmatrix} \right\}. \quad (2)$$

The assignment that one draws from the set,  $\Omega$ , determines which potential outcomes one observes. One can observe treatment potential outcomes only for units assigned to treatment, and control potential outcomes only for units assigned to control (recall the function  $y_i = z_i y_{t,i} + (1 - z_i) y_{c,i}$ , which determines the potential outcome that one observes for each individual  $i$ ). As Table 41.2 shows, for each of the  $\binom{6}{3} = 20$  possible assignments, there are  $\binom{6}{3} = 20$  corresponding possible

**Table 41.1 True values of  $y_c$ ,  $y_t$ , and  $\tau$ , where  $\tau_i = y_{t,i} - y_{c,i}$  for the study population**

$y_c$	$y_t$	$\tau$
20	22	2
8	12	4
11	11	0
10	15	5
14	18	4
1	4	3

**Table 41.2 All possible realizations of experimental data from a completely randomized study with six units and three treatment units**

$z_1$	$y_c$	$y_t$	$y_1$	$z_2$	$y_c$	$y_t$	$y_2$	$z_{19}$	$y_c$	$y_t$	$y_{19}$	$z_{20}$	$y_c$	$y_t$	$y_{20}$	
1	?	22	22	1	?	22	22	0	20	?	20	0	20	?	20	
1	?	12	12	1	?	12	12	0	8	?	8	0	8	?	8	
1	?	11	11	0	11	?	11	...	1	?	11	11	0	11	?	11
0	10	?	10	1	?	15	15	0	10	?	10	1	?	15	15	
0	14	?	14	0	14	?	14	1	?	18	18	1	?	18	18	
0	1	?	1	0	1	?	1	1	?	4	4	1	?	4	4	

realizations of observed data, where ‘?’ throughout this chapter denotes an unobserved and hence unknown potential outcome.

Only one such possible realization of data in Table 41.2 can be observed; but, knowing that there are 20 possible realizations allows the researcher to use procedures – e.g., estimators or hypothesis tests – to make inferences about unobservable causal quantities based on the single observed realization. We want procedures for drawing causal inferences to have properties that are ‘good’ (a notion that we will define more precisely in later sections). These properties describe or measure a procedure’s performance in two contexts: (1) studies with a fixed, finite size (number of units) and (2) a hypothetical scenario in which the size of a given study increases towards  $\infty$  while all other relevant factors remain constant. We refer to the latter context as one of *asymptotic growth*, which we conceptualize below.

If we have an experimental pool of six units, which is not a sample using a known procedure from a well-defined population, what does ‘asymptotic growth’ mean? We follow Brewer (1979) and Middleton and Aronow (2015) in using the idea of ‘copies’ as a way to talk about how estimators and tests behave as study sizes increase. In short, this conception of asymptotic growth states that (1) the original population of  $N$  units is copied  $h-1$  times such that there are  $h$  copies in total, (2) within each of the  $h$  copies, exactly  $n_t$  units are assigned to the treatment condition and the remaining  $n_c = N - n_t$  units

are assigned to the control condition and (3) the  $h$  copies are then collected into a single population with  $hN$  total units,  $hn_t$  treatment units and  $hn_c$  control units.

In the context of our working example, this conception of growth stipulates that the study population of  $N = 6$  units is embedded in a sequence of populations of increasing sizes in which the initial population is simply copied  $h - 1$  times.

Notice that over this sequence of increasing finite populations shown in Table 41.3, all relevant factors other than  $N$  remain constant: the proportions of treatment and control units remain fixed and the mean of control and treatment potential outcomes remain fixed, as do their variances and their covariance. Notice, however, that the number of possible assignments increases over this sequence of increasing

finite populations from  $\binom{6}{3} = 20$  to  $\binom{12}{6} = 924$

and from  $\binom{18}{9} = 48620$  to  $\binom{24}{12} = 2704156$

and so forth.

We will show that, in either the finite context – given in Table 41.1 – or in the asymptotic context – given in Table 41.3 – whether a procedure is ‘good’ depends on whether it maintains fidelity to the research design – i.e., the probability distribution on the set  $\Omega$ . In other words, we will show that in a randomized experiment, a ‘good’ procedure is one that heeds the dictum of Senn (2004, 3729) who, in the voice of R. A. Fisher, states that ‘[a]s ye randomise so shall ye analyse’.

**Table 41.3** Finite populations under asymptotic growth in which  $h \in \{1, 2, 3, 4, \dots\}$

			$y_c$	$y_t$	$\tau$				$y_c$	$y_t$	$\tau$				$y_c$	$y_t$	$\tau$
			20	22	2				20	22	2				20	22	2
			8	12	4				8	12	4				8	12	4
			11	11	0				11	11	0				11	11	0
			10	15	5				10	15	5				10	15	5
			14	18	4				14	18	4				14	18	4
			1	4	3				1	4	3				1	4	3
			10	15	5				10	15	5				10	15	5
			14	18	4				14	18	4				14	18	4
			1	4	3				1	4	3				1	4	3
			10	15	5				10	15	5				10	15	5
			14	18	4				14	18	4				14	18	4
			1	4	3				1	4	3				1	4	3
			10	15	5				10	15	5				10	15	5
			14	18	4				14	18	4				14	18	4
			1	4	3				1	4	3				1	4	3
			10	15	5				10	15	5				10	15	5
			14	18	4				14	18	4				14	18	4
			1	4	3				1	4	3				1	4	3
			10	15	5				10	15	5				10	15	5
			14	18	4				14	18	4				14	18	4
			1	4	3				1	4	3				1	4	3

**ESTIMATION**

As we have mentioned above, no one can observe both potential outcomes for any given unit in a given study population. One can, however, generate a guess about some function of the study population’s individual causal effects (e.g., the mean causal effect) using observed outcomes. We call this unobservable causal quantity the *estimand*. The *estimator*, by contrast, refers to the procedure that generates a guess about the

estimand. An *estimate* is the actual output of the estimator once it is applied to a given data set.

One estimand is the mean causal effect,  $\bar{\tau} = \left(\frac{1}{N}\right) \sum_{i=1}^N \tau_i$ , which, to return to the example from Table 41.1, is  $\bar{\tau} = \frac{2+4+0+5+4+3}{6} = 3$ .

A procedure for generating a guess about  $\bar{\tau}$  is the Difference-in-Means estimator, which we can define in terms of observable quantities as follows:

$$\hat{\tau}(Z, Y) = \frac{Z'Y}{Z'Z} - \frac{(1-Z)'Y}{(1-Z)'(1-Z)} = \left(\frac{1}{\sum_{i=1}^N Z_i}\right) \sum_{i=1}^N Z_i Y_i - \left(\frac{1}{\sum_{i=1}^N (1-Z_i)}\right) \sum_{i=1}^N (1-Z_i) Y_i \tag{3}$$

In the example from Table 41.1, the random vectors<sup>3</sup> of  $\mathbf{Z}$  and  $\mathbf{Y}$  can take on any of the possible values,  $(\mathbf{z}_1, \mathbf{y}_1), \dots, (\mathbf{z}_{20}, \mathbf{y}_{20})$ , given in Table 41.2. If we apply the estimator in Equation (3) to the possible realizations of data in Table 41.2, then there are 20 possible estimates that correspond to each of the 20 possible realizations of data:

$$\hat{\tau}(\mathbf{z}_1, \mathbf{y}_1) = 6.6667, \dots, \hat{\tau}(\mathbf{z}_{20}, \mathbf{y}_{20}) = -0.6667.$$

The researcher can observe only one of these 20 possible estimates and this single estimate should be generated by an estimator that is ‘good’. More specifically, three ‘good’ properties of an estimator are *unbiasedness*, *consistency* and *precision*. An unbiased estimator is one in which, although any single estimate may be close to or far from the true value of the estimand, the *expected value of the estimator* – i.e., the probability-weighted mean of all possible estimates – is equal to the value of the estimand. Consistency states that as the number of units in the study increases asymptotically, holding all other factors constant, the probability distribution of an estimator concentrates increasingly around the truth. (For any fixed  $\varepsilon > 0$ , the probability that the estimate and its target differ by no more than  $\varepsilon$  tends to 1.) Lastly, a precise estimator is one in which the expected distance of an estimate from the true causal estimand is small.

In the following discussion, we show the role that research design plays in whether the Difference-in-Means estimator is unbiased, consistent and/or precise with respect to the estimand  $\bar{\tau} = \left(\frac{1}{N}\right) \sum_{i=1}^N \tau_i$ . We also show how designs can yield estimators that are more or less precise. Researchers may want to estimate quantities other than  $\bar{\tau}$ . Although we do not discuss such cases, the general principles for determining whether an estimator is unbiased, consistent and/or precise with respect to the causal quantity of interest is the same: researchers can define an estimand that they seek to infer, define an estimator by which they would estimate this quantity

under all possible realizations of data and subsequently assess whether this estimator is unbiased, consistent and/or precise based only on the probabilities with which possible data are realized.

### Unbiasedness

We now show that a ‘good’ estimator of the unknown estimand,  $\bar{\tau}$ , is the estimator given in Equation (3),  $\hat{\tau}(\mathbf{Z}, \mathbf{Y})$ . In particular, we will show that this estimator satisfies the criterion of no systematic error – i.e., unbiasedness – in a *uniform randomized experiment*, when the numbers of treatment and control units are both fixed.<sup>4</sup> Whether the Difference-in-Means estimator is unbiased with respect to  $\bar{\tau}$  depends solely on the known research design, i.e., whether or not there is a uniform probability distribution on the set of assignments.

Returning to the example from Table 41.1, recall that if we apply the estimator in Equation (3) to the possible realizations of data in Table 41.2, then there are 20 possible estimates that correspond to each of the 20 possible realizations of data:

$$\hat{\tau}(\mathbf{z}_1, \mathbf{y}_1) = 6.6667, \dots, \hat{\tau}(\mathbf{z}_{20}, \mathbf{y}_{20}) = -0.6667.$$

Informally, an unbiased estimator produces a guess about the estimand with no systematic error. Slightly more formally, an estimator is unbiased if the average of all possible estimates is equal to the true value of the estimand. This average of estimates, however, must be weighted by the probabilities of observing each possible estimate; we call this average the ‘expected value’ and denote the expected value of the Difference-in-Means estimator by  $\mathbb{E}[\hat{\tau}(\mathbf{Z}, \mathbf{Y})]$ .

To calculate the expected value of the Difference-in-Means estimator to assess properties like bias and consistency, we need to know the probability associated with each of these 20 possible estimates. We know that the estimator is a function of two random quantities,  $\mathbf{Z}$  and  $\mathbf{Y}$ ,



but  $\mathbf{Y}$  inherits randomness only from  $\mathbf{Z}$ , since  $Y_i = Z_i y_{1,i} + (1 - Z_i) y_{c,i}$  for all  $i \in \{1, \dots, N\}$  units. Therefore, each probability associated with its corresponding estimate depends only on  $\mathbf{Z}$ . So, we calculate the expected value of the estimator in general as follows:

$$\mathbb{E} \left[ \hat{\tau}(\mathbf{Z}, \mathbf{Y}) \right] = \hat{\tau}(\mathbf{z}_1, \mathbf{y}_1) \Pr(\mathbf{Z} = \mathbf{z}_1) \\ + \dots + \hat{\tau}(\mathbf{z}_{|\Omega|}, \mathbf{y}_{|\Omega|}) \Pr(\mathbf{Z} = \mathbf{z}_{|\Omega|}).$$

In the context of the running example, there are 20 possible estimates corresponding to each of the  $\mathbf{z}_1, \dots, \mathbf{z}_{20}$  possible assignments, and the probability of each of those possible assignments is  $\frac{1}{20}$ . Therefore, the expected value of the Difference-in-Means estimator is

$$\mathbb{E} \left[ \hat{\tau}(\mathbf{Z}, \mathbf{Y}) \right] = \hat{\tau}(\mathbf{z}_1, \mathbf{y}_1) \Pr(\mathbf{Z} = \mathbf{z}_1) \\ + \dots + \hat{\tau}(\mathbf{z}_{20}, \mathbf{y}_{20}) \Pr(\mathbf{Z} = \mathbf{z}_{20}) \\ = 6.6667 \left( \frac{1}{20} \right) \\ + \dots + -0.6667 \left( \frac{1}{20} \right) \\ = 3.$$

In this example, the expected value of the estimator,  $\mathbb{E}[\hat{\tau}(\mathbf{Z}, \mathbf{Y})]$ , is exactly equal to the true mean causal effect,  $\bar{\tau}$ . The estimator is unbiased given the design. If  $\Pr(\mathbf{Z} = \mathbf{z})$  did not equal  $\frac{1}{20}$  for all  $\mathbf{z}$  – i.e., if some assignments were more or less probable than others – then the Difference-in-Means estimator might not be unbiased. In general, the equality between  $\mathbb{E}[\hat{\tau}(\mathbf{Z}, \mathbf{Y})]$  and  $\bar{\tau}$  holds when (1) units are assigned to study conditions as individuals (not as groups, i.e., *clusters*), (2) there is always at least one unit in the treatment condition and one in the control condition and (3) each possible assignment has an identical probability of realization. In other words, the Difference-in-Means estimator is

unbiased with respect to the mean additive causal effect in a uniform randomized experiment under either complete individual assignment or simple individual assignment so long as there is always at least one unit in each of the study conditions.

Notice that the Difference-in-Means estimator did not require large numbers of units or assumptions about the distributions of potential outcomes to be unbiased. The potential outcomes could have been any set of values and the property of unbiasedness would still have held. However, the unbiasedness property did require that there be a uniform probability distribution on the set of possible assignments. But in a controlled research design, like a randomized experiment, the researcher knows whether or not this condition is true.

### Consistency

‘No systematic error’ is not the same as ‘close to the truth’. Achen (1982, 36) explains the need for another conception of a ‘good’ estimator when he writes ‘[u]nbiasedness is too weak a property, since it says nothing about approximating the truth’. While we know that an unbiased estimator yields estimates that are, on average, equal to the true value of the estimand, any single estimate might be far from the truth. In our running example, not one of the 20 possible estimates is actually equal to the true mean causal effect of  $\bar{\tau} = 3$ , even though the probability-weighted average of those 20 estimates is equal to 3. Another ‘good’ characteristic of an estimator is to produce values close to the truth as more information is supplied to the estimator from the design.

In contrast to unbiasedness, consistency states that as the number of units in the study grows asymptotically, holding all other factors constant, the probability of an estimate within an arbitrarily small distance,  $\epsilon$ , from the truth is equal to 1. More formally, we can define consistency as follows:

$$\lim_{h \rightarrow \infty} \Pr \left( \left| \hat{\tau}(\mathbf{Z}, \mathbf{Y}) - \bar{\tau} \right| < \varepsilon \right) = 1 \text{ for all } \varepsilon > 0 \quad (4)$$

or equivalently as

$$\lim_{h \rightarrow \infty} \Pr \left( \hat{\tau}(\mathbf{Z}, \mathbf{Y}) \in (\bar{\tau} - \varepsilon, \bar{\tau} + \varepsilon) \right) = 1 \text{ for all } \varepsilon > 0, \quad (5)$$

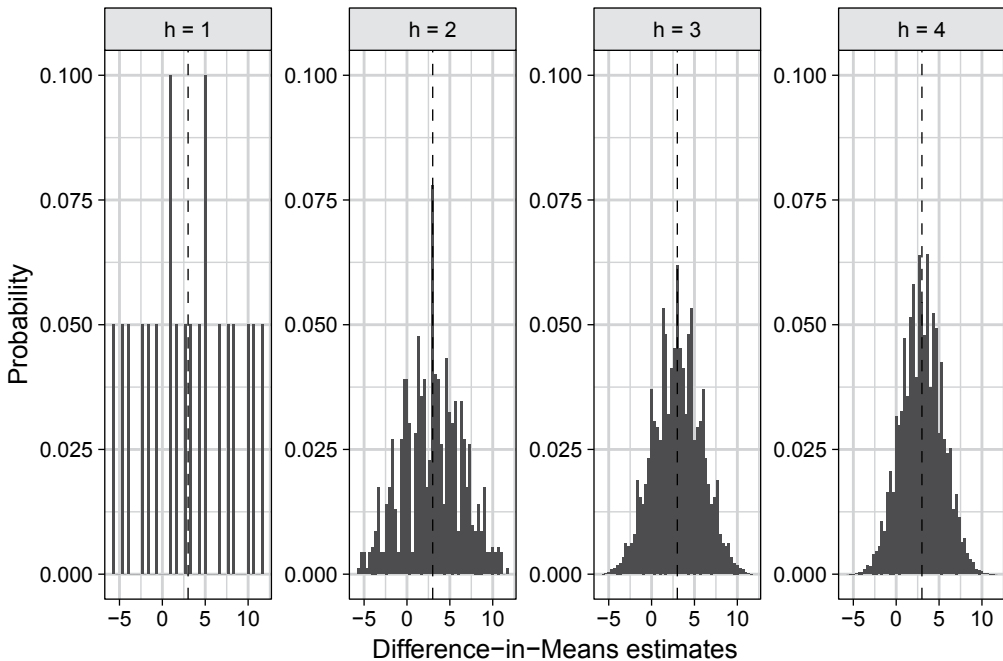
where, referring back to the conception of asymptotic growth in Table 41.3,  $h$  is the number of copies of the original finite population from Table 41.1.

To unpack Equations (4) and (5), Figure 41.1 shows what happens to the distribution of the Difference-in-Means estimator under complete random assignment as  $h \rightarrow \infty$ .

The general trend is that the probability of estimates close to the true mean causal effect,  $\bar{\tau} = 3$ , grows larger and larger and ultimately converges in probability (over the sequence of increasing finite populations) to 1. For example, following Equation (5), consider the probability that an estimate lies on

the interval  $(3 - \varepsilon, 3 + \varepsilon)$  and let  $\varepsilon = 1$ . The respective probabilities of an estimate on this interval for  $h \in \{1, 2, 3, 4\}$  are 0.1, 0.2359, 0.2559 and 0.2994, and as  $h \rightarrow \infty$ , that probability tends to 1. This property holds for  $\varepsilon = 1$  as well as for any positive value of  $\varepsilon$  that one could choose. For example, we could have let  $\varepsilon = 0.5$ , in which case the corresponding probabilities for  $h \in \{1, 2, 3, 4\}$  are 0.1, 0.1234, 0.1522 and 0.1643, and the limiting probability as  $h \rightarrow \infty$  is also 1. In general, it is not necessary that each probability be greater than its predecessor for  $h \in \{1, 2, 3, 4, \dots\}$ . Consistency states only that there exists some number in which, for any  $h$  greater than that number, the estimator will lie within an arbitrarily small interval (of distance  $\forall \varepsilon > 0$ ) around the true mean causal effect.

In this particular case, consistency follows (in part) from unbiasedness. As the size of  $N$  increases towards  $\infty$  while all other factors remain constant, the probability of an estimate



**Figure 41.1** Distribution of Difference-in-Means estimator as  $h \rightarrow \infty$ . The dashed lines denote the expected value of the estimator.

arbitrarily close to the estimator’s expected value is equal to 1. Unbiasedness ensures that the expected value of the estimator is equal to the true value of the estimand. Therefore, as the size of the study population grows towards  $\infty$ , the estimator produces a value arbitrarily close to the truth (not just to the estimator’s expected value) with a probability of 1. Both the unbiasedness and consistency of the Difference-in-Means estimator, moreover, arise solely from the research design. Even though the distribution of the estimator starts to look more and more normal as the size of the study population increases to  $\infty$ , we made no such distributional assumptions to show the estimator’s unbiasedness and consistency.

**Precision**

While unbiased and consistent estimators are desirable, such estimators may yield estimates far from the truth, with high probability in actual experiments with fixed study populations. One estimator is more precise than another estimator for a given study design if it produces estimates that are closer to the truth on average. In other words, a ‘good’ estimator also has a low variance; in our case, a more precise estimator than the Difference-in-Means estimator would make the 20 possible estimates in Table 41.2 closer to the true mean causal effect, on average. We now consider first the factors that make the Difference-in-Means estimator produce guesses with lower expected distance from the true mean causal effect and second the procedure one can use to conservatively estimate the variance of the Difference-in-Means estimator.

Neyman (1923) derived an exact analytic expression for the variance of the Difference-in-Means estimator based solely on the research design of a randomized experiment, as follows:

$$\sigma_{\bar{\tau}}^2 = \frac{1}{N-1} \left( \frac{n_t \sigma_{y_c}^2}{n_c} + \frac{n_c \sigma_{y_t}^2}{n_t} + 2\sigma_{y_c, y_t} \right), \quad (6)$$

where  $\sigma_{y_c}^2$  is the variance of control potential outcomes,  $\sigma_{y_t}^2$  is the variance of treatment potential outcomes and  $\sigma_{y_c, y_t}$  is the covariance of control and treatment potential outcomes.

Equation (6) suggests that one can increase the precision of the Difference-in-Means estimator (i.e., reduce the estimator’s variance) by increasing the number of treatment units and/or the number of control units. Precision can also be increased by decreasing the variances of treatment,  $\sigma_{y_t}^2$ , and control,  $\sigma_{y_c}^2$ , potential outcomes. For a simple and clear account of the factors that increase precision, see Gerber and Green (2012, section 3.2).

A standard design choice that researchers can make to increase precision is blocking. That is, a researcher can first construct blocks that are similar in terms of covariates related to potential outcomes and second assign units to study conditions within blocks. Blocked assignment works by excluding assignments that, on average, yield estimates far from the true mean effect. To see this point, we return to the example in Table 41.1 and introduce  $\mathbf{x}$ ,  $\mathbf{x} \in 0,1^N$  which is a vector of baseline covariate values for all  $i \in 1, \dots, N$  units. The vector  $\mathbf{x}$  is a fixed quantity that is measured for all units prior to assignment; hence,  $\mathbf{x}$  cannot change as a function of whichever assignment is realized.

From Table 41.4, we can see that  $\mathbf{x}$  is related to both  $y_c$  and  $y_t$ . The treatment potential outcomes are greater, on average, among units whose baseline covariate values are equal to 1 compared to units whose baseline covariate values are equal to 0. The same is

**Table 41.4 True values of  $y_c$ ,  $y_t$ ,  $\tau$  and the baseline covariate  $\mathbf{x}$**

$y_c$	$y_t$	$\tau$	$\mathbf{x}$
20	22	2	1
8	12	4	1
11	11	0	0
10	15	5	1
14	18	4	1
1	4	3	0

true for control potential outcomes. If the researcher puts units whose covariate values are equal to 1 in one block and units whose covariate values are equal to 0 in another, and then assigns half of the units to treatment and control within blocks, the set of possible assignments,  $\Omega_b$ , would be as follows:

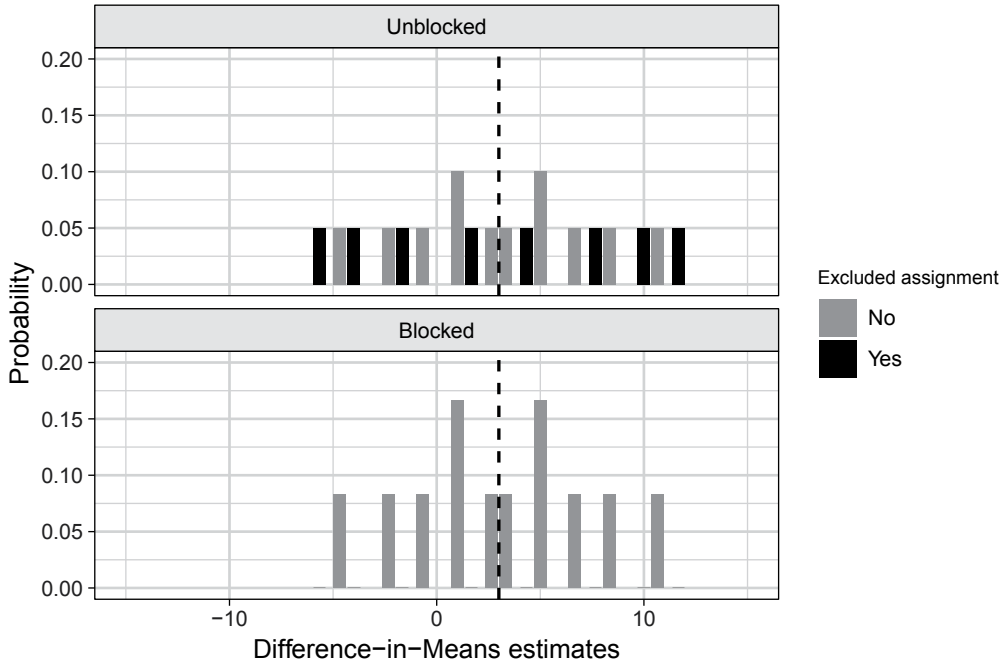
$$\Omega_b = \left\{ \begin{bmatrix} 1 \\ 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 1 \end{bmatrix}, \dots, \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 0 \end{bmatrix} \right\}. \quad (7)$$

The set  $\Omega_b$  above has only 12 possible assignments as opposed to the 20 possible assignments in Equation (2) under complete random assignment without blocks. In particular, the

assignments of  $\mathbf{z}_2, \mathbf{z}_3, \mathbf{z}_7, \mathbf{z}_8, \mathbf{z}_{13}, \mathbf{z}_{14}, \mathbf{z}_{18}, \mathbf{z}_{19} \in \Omega$  are excluded from  $\Omega_b$ .

Figure 41.2 shows the eight estimates corresponding to the eight assignments that were included in unblocked assignment but excluded in blocked assignment. On average, these eight estimates are farther from the true mean effect than are the other 12 estimates. More concretely, the average squared distance of the eight excluded estimates from the truth is 36.91667 and the same average squared distance of the 12 included assignments is 18.12963. Hence, this blocked design increases precision by reducing the probability (to 0) of estimates that are, on average, far from the truth and by increasing the probability of estimates that are, on average, closer to the truth.<sup>5</sup>

Thus far, we have focused on design choices that can decrease the variance of the Difference-in-Means estimator. But, note



**Figure 41.2** Distribution of Difference-in-Means estimates under (a) unblocked and (b) blocked assignment. The dashed lines denote the expected value of the estimator.

that the variance of the Difference-in-Means estimator is, like the true mean causal effect, a fixed, unobservable quantity. As we can see from Equation (6), the variance of the Difference-in-Means estimator depends on the variance of treatment and control potential outcomes as well as their covariance, none of which can be directly observed. We need a ‘good’ procedure by which we can estimate the variance of the Difference-in-Means estimator if we are to reasonably infer its precision. We will use such a variance estimator in the context not only of evaluating estimators but also hypothesis testing.

One can unbiasedly and consistently estimate two of the three unknown quantities in Equation (6). Following Cochran (1977), unbiased and consistent estimators of  $\sigma_{y_c}^2$  and  $\sigma_{y_t}^2$ , respectively, are:

$$\hat{\sigma}_{y_c}^2 = \left( \frac{N-1}{N(n_c-1)} \right) \sum_{i:Z_i=0}^N (y_{c,i} - \hat{\mu}_{y_c})^2$$

$$\text{and } \hat{\sigma}_{y_t}^2 = \left( \frac{N-1}{N(n_t-1)} \right) \sum_{i:Z_i=1}^N (y_{t,i} - \hat{\mu}_{y_t})^2, \text{ where}$$

$$\hat{\mu}_{y_c} = \left( \frac{1}{n_c} \right) \sum_{i=1}^N (1-Z_i)y_{c,i} \text{ and } \hat{\mu}_{y_t} = \left( \frac{1}{n_t} \right) \sum_{i=1}^N Z_i y_{t,i}$$

We cannot write an unbiased and consistent estimator for  $\sigma_{y_c, y_t}$  since no two potential outcomes for any unit can be jointly observed. Neyman (1923) noted, however, that one could use a conservative procedure for estimating the quantity in Equation (6) by assuming the largest possible value of  $2\sigma_{y_c, y_t}$ , which, by the Cauchy–Schwarz inequality and the AM–GM inequality (i.e., inequality of arithmetic and geometric means), is  $\sigma_{y_c}^2 + \sigma_{y_t}^2$ .

After substituting  $\sigma_{y_c}^2 + \sigma_{y_t}^2$  for  $2\sigma_{y_c, y_t}$ , the analytic expression for the variance of the Difference-in-Means estimator (assuming  $2\sigma_{y_c, y_t} = \sigma_{y_c}^2 + \sigma_{y_t}^2$ ) is

$$\frac{1}{N-1} \left( \frac{n_t \sigma_{y_c}^2}{n_c} + \frac{n_c \sigma_{y_t}^2}{n_t} + \sigma_{y_c}^2 + \sigma_{y_t}^2 \right),$$

which can be simplified to

$$\frac{N}{N-1} \left( \frac{\sigma_{y_c}^2}{n_c} + \frac{\sigma_{y_t}^2}{n_t} \right). \tag{8}$$

Now there are only two unknown quantities in Equation (8), each of which can be unbiasedly estimated. Hence, one can now unbiasedly estimate the quantity in (8) via the conservative variance estimator of

$$\hat{\sigma}_{\hat{\tau}}^2 = \frac{N}{N-1} \left( \frac{\hat{\sigma}_{y_c}^2}{n_c} + \frac{\hat{\sigma}_{y_t}^2}{n_t} \right). \tag{9}$$

This estimator is conservative because, since it unbiasedly estimates the quantity in (8), its expected value is equal to or greater than the true variance of the estimator given in (6).<sup>6</sup>

Thus far, we have explained the role that research design – i.e., the probability distribution on the set of assignments,  $\Omega$  – plays in determining whether estimators are unbiased, consistent and precise. We have also explained how one can infer the variance of an estimator via a conservative procedure. We have used a simple example of complete uniform assignment to illustrate these points; yet an estimator that is unbiased, consistent and/or relatively precise in this design may not be so in another design. For example, the Difference-in-Means estimator is not necessarily unbiased when there is a non-uniform probability distribution on  $\Omega$ ; however, the Horvitz–Thompson (i.e., inverse probability weighted) estimator (Horvitz and Thompson, 1952) is unbiased in such a design (see Aronow and Middleton, 2013). Such design-based inference differs from model-based inferences in that the former remains reliable without the need to impose a probability model on potential outcomes or to model the functional form (e.g., a linear model) that links the treatment variable to potential outcomes. The only probability model in design-based inference is the assignment process itself, which, in the case of a randomized experiment, is known to be the true model of the data-generating process.

## HYPOTHESIS TESTING

Focusing on  $\bar{\tau}$  engages with the fundamental problem of causal inference by aggregation across units in the study and  $\hat{\bar{\tau}}$  can be shown to be an unbiased, consistent and, depending on the study design, relatively precise estimator. An alternative approach begins with claims about causal effects rather than guesses about them, stating hypotheses about  $\bar{\tau}$  or even about individual effects,  $\tau$ . This approach then tests those claims, and so the procedures that we will assess in this section are properties of tests rather than of estimators. A researcher who performs a hypothesis test for causal inference first states a null hypothesis about a relationship between unobserved potential outcomes and a (usually composite) alternative hypothesis, which we will define more precisely below. The researcher can then assess the probability that the research design generates data under the null hypothesis that are more extreme than the actually observed data. The observed data and data under the null hypothesis are summarized via a test statistic that maps the data — the observed outcome and treatment assignment — to a single number. For example, the test statistic calculated on whichever data are generated by the design,  $t(\mathbf{Z}, \mathbf{Y})$ , could be

$$\left( \frac{1}{\sum_{i=1}^N Z_i} \right) \sum_{i=1}^N Z_i Y_i - \left( \frac{1}{\sum_{i=1}^N (1 - Z_i)} \right) \sum_{i=1}^N (1 - Z_i) Y_i,$$

which is the same formula that we labeled  $\hat{\bar{\tau}}$  in Equation (3) but, in the context of hypothesis testing, is not an estimator but a data summary. We call the probability of a test statistic under the null more extreme than the observed test statistic a probability-value or  $p$ -value. Typically, when the  $p$ -value is less than or equal to a pre-specified ‘significance level’ of the test, which we denote by  $\alpha \in (0,1)$ , a researcher *rejects* the null hypothesis, meaning that the researcher declares that the observed data are not consistent with the null

hypothesis relative to an alternative hypothesis. When the  $p$ -value is greater than the significance level of the test, the researcher *fails to reject* the null hypothesis relative to an alternative hypothesis — meaning that the researcher declares that there is not enough information to state that the observed data are inconsistent with the hypothesized state of the world. Sometimes, we talk about hypothesis testing as an attempt to distinguish signal from noise; a high  $p$ -value tells us that we cannot distinguish signal from noise, and a low  $p$ -values tells us that we can do so.

Hypothesis tests are subject to at least two types of errors: first, one could reject the null hypothesis when it is true (a type I error) or, second, fail to reject the null hypothesis when it is false (a type II error). Two features of hypothesis tests related to these two potential errors are the  $\alpha$  size of the test and the *power* of the test. We now define the  $\alpha$  size (distinct from the  $\alpha$  level) and power-of-hypothesis tests.

A test’s  $\alpha$  level is, in the words of Rosenbaum (2010), that test’s ‘promise’ that the probability of a Type I error (i.e., the probability of a  $p$ -value that is less than  $\alpha$  when the null hypothesis is true) is less than or equal to the  $\alpha$  level. The test’s  $\alpha$  size, on the other hand, is the test’s true probability of a Type I error, which, in general, can be greater than, equal to or less than the  $\alpha$  level ‘promised’ by the test. In contrast to the  $\alpha$  level and size of a test, a test’s power is the probability of a  $p$ -value that is less than the  $\alpha$  level when the null hypothesis is false. In other words, power is 1 minus the Type II error probability; hence, as the power of a test increases, the Type II error probability decreases.

In the subsections to follow, we first define ‘good’ properties of hypothesis tests. We then describe tests of causal hypotheses in two distinct traditions traceable to Fisher (1935) (subsequently developed most extensively by Rosenbaum, 2002, 2010), and Neyman and Pearson (1933). We then explain the role that research design plays in justifying whether tests in either of these two traditions have good properties.

### **What Makes a Hypothesis Test a 'Good' Test?**

We have already discussed three properties of good estimators – namely, unbiasedness, consistency and precision – but what makes a test of a null hypothesis relative to an alternative hypothesis a 'good' test? Just as we did for estimation, we describe 'good' features of hypothesis tests in the context of a fixed, finite population and in the context of a hypothetical scenario in which a study's size increases towards  $\infty$  by increasing the number of copies of the study. The first two 'good' properties (a Type I error probability less than the  $\alpha$  level and an unbiased test) refer to the former context, and the third property (a consistent test) refers to the latter context. Informally, a good hypothesis test should rarely mislead us: it should rarely encourage us to declare that we have discovered a signal in the noise when no signal exists and it should often find signals when they do exist.

Regardless of the size of a given study population, a hypothesis test first ought to control its  $\alpha$  size (true Type I error probability) such that it is less than or equal to the test's  $\alpha$  level. Second, a hypothesis test ought to be an *unbiased test* (not to be confused with an *unbiased estimator*), i.e., the probability of rejecting the null hypothesis when it is false and the alternative hypothesis is true should be at least as great as the probability of rejecting the null hypothesis when it is true and the alternative hypothesis is false (see Lehmann and Romano, 2005, chapter 4). Intuitively, we want to *reject* something that is *false* and we want to *not reject* something that is *true*. A test that leads us to reject true nulls with greater probability than we reject false nulls does not yield inferences that track the true causal effect. A good hypothesis test ought to be an unbiased test in this sense.

Turning now to the asymptotic context described in the section 'An Illustrative Example', a hypothesis test ought to be a

*consistent test* (not the same as a *consistent estimator*); that is, as the size of the study population increases asymptotically while all other relevant factors remain constant, the probability of rejecting the null hypothesis when it is false and the alternative is true should tend to 1 (see Lehmann and Romano, 2005, chapter 11). We now show the role that research design plays in enabling 'good' hypothesis tests in two different design-based traditions: Fisherian (Fisher, 1935) and Neymanian (Neyman and Pearson, 1933).

### **Fisherian Hypothesis Testing**

Hypothesis testing in the tradition of Fisher (1935), later developed most extensively by Rosenbaum (2002, 2010), assesses the consistency of the observed data with a null hypothesis vis-à-vis an alternative hypothesis. A strong null hypothesis,<sup>7</sup> which we denote by  $\tau_0$ , postulates an individual treatment effect for all  $i \in \{1, \dots, N\}$  units in a given study population. For example, one strong null hypothesis is  $\tau'_0 = [5 \ 5 \ \dots \ 5 \ 5]$  and another might be  $\tau'_0 = [0.5 \ -10 \ \dots \ 200 \ -74.25]$ . The *strong null hypothesis of no effect* (which we henceforth refer to as 'the strong null of no effect') specifically postulates that  $\tau_i = 0$  for all  $i \in \{1, \dots, N\}$  units – i.e., that  $\tau'_0 = [0 \ 0 \ \dots \ 0 \ 0]$ .

The consistency of the observed data with a strong null hypothesis vis-a-vis an alternative is typically assessed via  $p$ -values. To reiterate, a  $p$ -value is the probability of a test statistic at least as extreme as the observed test statistic from the perspective of the null hypothesis: as we will show below, the hypothetical world of the null generates, along with the known research design, the probability distribution that we compare against our single observed test statistic. In the context of Fisherian hypothesis tests, we can formally represent upper ( $p_u$ ), lower ( $p_l$ ) and two-sided ( $p_t$ )  $p$ -values as follows:

$$\begin{aligned}
 p_u &= \sum_{m=1}^{|\Omega|} \mathbb{1} \left[ t(\mathbf{z}_m, \mathbf{y}_{0_j}) \geq T \right] \Pr(\mathbf{Z} = \mathbf{z}_j) \\
 p_l &= \sum_{m=1}^{|\Omega|} \mathbb{1} \left[ t(\mathbf{z}_m, \mathbf{y}_{0_j}) \leq T \right] \Pr(\mathbf{Z} = \mathbf{z}_j) \quad (10) \\
 p_t &= \min\{1, 2 \min\{p_u, p_l\}\},
 \end{aligned}$$

where the index  $m \in \{1, \dots, |\Omega|\}$  runs over all possible assignments in the set of assignments  $\Omega$ ,  $\mathbb{1}$  is an indicator function that is 1 if the argument  $[\cdot]$  is true and 0 if false,  $t(\mathbf{z}_m, \mathbf{y}_{0_m})$  is the null test statistic (using  $\mathbf{y}_{0_m}$  to refer to the vector of observed outcomes for the  $m$ th assignment implied by the null hypothesis,  $H_0$ ) and  $T$  is the observed test statistic.<sup>8</sup>

To provide an illustration of Fisherian  $p$ -values, we return to the example in Table 41.1 and imagine that the assignment  $\mathbf{z}'_8 = [1 \ 0 \ 0 \ 1 \ 1 \ 0]$  happened to be the one randomly selected. In this case, the realization of data would be as in Table 41.5.

If the researcher uses the Difference-in-Means test statistic to provide a single, numerical summary of the observed data in Table 41.5, then the observed test statistic would be  $t(\mathbf{z}_8, \mathbf{y}_8) = 11.6667$ . Let's assume that the researcher wants to assess the consistency of this observed test statistic with the strong null of no effect – i.e., that  $\tau_i = 0$  for all  $i$  – relative to the alternative hypothesis of a positive effect – i.e., that  $\tau_i$  is non-negative for all  $i$  and positive for at least one  $i$ .

Potential outcomes are only partially observed, but the researcher can ‘fill in’ the

missing potential outcomes according to the strong null hypothesis  $H_0 : y_{t,i} = y_{c,i}$  for all  $i$ . Below, we can also show what this hypothesis implies for the observed outcomes  $y_t$ , recalling that  $y_i = Z_i y_{t,i} + (1 - Z_i) y_{c,i}$  and writing  $y_{c0,i}$  and  $y_{t0,i}$  to mean ‘value of  $y_{c,i}$  and  $y_{t,i}$  under a test of  $H_0$ ’:

$$\begin{aligned}
 y_{c0,i} &= y_i - z_i \tau_{0_i}, \\
 y_{t0,i} &= y_i + (1 - z_i) \tau_{0_i},
 \end{aligned} \quad (11)$$

which in the case of the strong null of no effect implies that units’ null potential outcomes and observed outcomes are as they appear in Table 41.6.

Assuming for the sake of argument that the strong null of no effect is true, the researcher knows exactly what all other possible realizations of data would look like under each possible assignment in  $\Omega$ . Hence, the researcher can summarize all other possible realizations of data under the null with the same Difference-in-Means test statistic – generating a probability distribution of those null test statistics – and then calculate the probability of a null test statistic greater than or equal to the observed test statistic of 11.6667 (see Figure 41.3).

In this case, the upper  $p$ -value –  $p_u$  from Equation (10) – is 0.05. Assume that the value of  $\alpha$  has been pre-set to a level greater than 0.05, which leads the researcher to reject the strong null of no effect in favor of the alternative of a positive effect. Table 41.1 shows that the strong null of no effect is

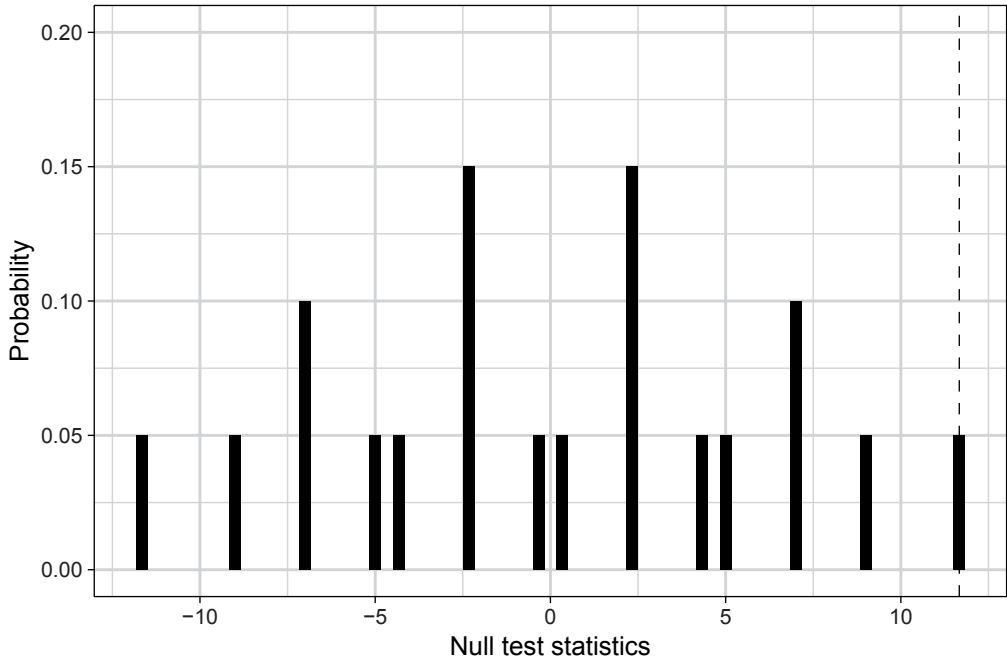
**Table 41.5 Realization of Data if  $\mathbf{z}_8$  were the randomly drawn assignment**

$\mathbf{z}_8$	$\mathbf{y}_c$	$\mathbf{y}_t$	$\mathbf{y}_8$
1	?	22	22
0	8	?	8
0	11	?	11
1	?	15	15
1	?	18	18
0	1	?	1

**Table 41.6 Null potential outcomes if  $\mathbf{z}_8$  were the realized assignment and under a test of the null hypothesis that  $\mathbf{y}_{t,i} = \mathbf{y}_{c,i}$  for all  $i$**

$\mathbf{z}_8$	$\mathbf{y}_{c08}$	$\mathbf{y}_{t08}$	$\mathbf{y}_8$
1	22	22	22
0	8	8	8
0	11	11	11
1	15	15	15
1	18	18	18
0	1	1	1





**Figure 41.3** Distribution of Difference-in-Means test statistic under test of strong null of no effect when  $z_g$  is the realized assignment. The dashed line denotes the value of the observed test statistic.

false ( $y_{t,i} \neq y_{c,i}$  for all  $i$ ) and the alternative of a positive effect is true; hence, this particular choice to reject the strong null of no effect relative to the alternative of a positive effect happened to be a good one.

To assess whether the hypothesis-testing procedure is a good one overall, we want to establish that the test (1) has a true Type I error probability less than the  $\alpha$  level, (2) is unbiased and (3) is consistent. We will now illustrate these three properties in turn.

Let's return to the example given in Table 41.1, where the true vector of individual causal effects is  $\tau' = [2\ 4\ 0\ 5\ 4\ 3]$ , and Table 41.2 describes the 20 possible realizations of data. When the null hypothesis,  $\tau_0$ , is false, the potential outcomes implied by the null hypothesis vary depending on which data are realized. However, when the null hypothesis,  $\tau_0$ , is true, i.e.,

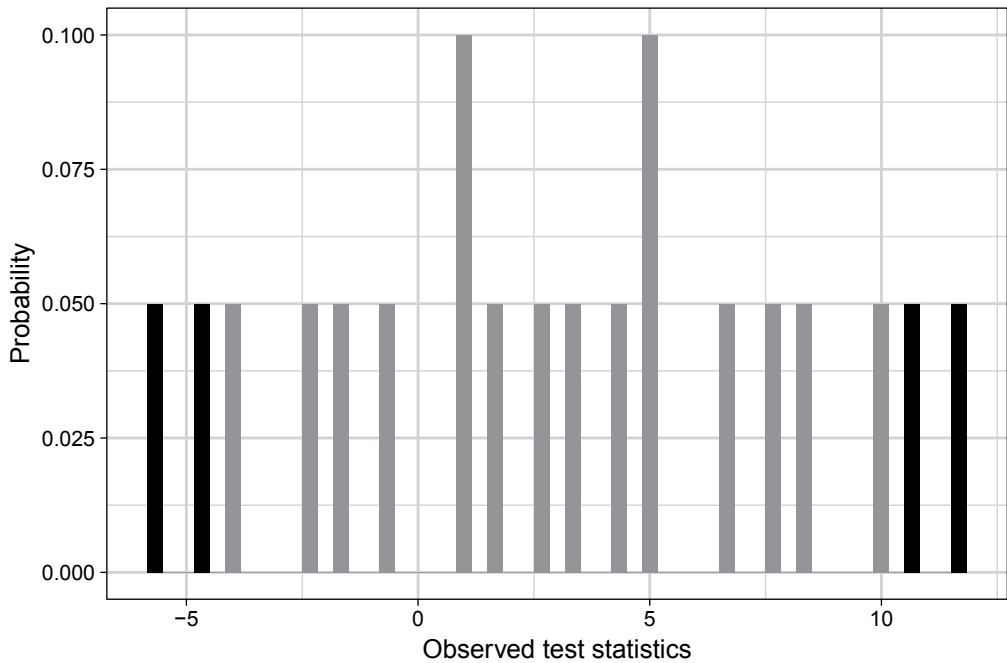
when  $\tau_0 = \tau$ , the potential outcomes implied by the null hypothesis are fixed across all possible realizations of data, as shown by Table 41.7.

If we set the significance level of the test to, say,  $\alpha = 0.10$ , then it is true by definition that the probability that an observed test statistic lies in the lower tail of its distribution is less than or equal to 0.10, and the same is true for the probability that an observed test statistic lies in the upper tail of its distribution. Figure 41.4 shows the distribution of this test statistic, in which both the lower and upper tails according to  $\alpha = 0.10$  are shaded darker.

Notice that no matter which of the 20 possible realizations of data is actually realized, the null potential outcomes when the null hypothesis is true are identical to the true potential outcomes in Table 41.1, and hence each of the 20 possible null distributions of

**Table 41.7 Null potential outcomes for all possible realizations of data when the null hypothesis is true.**

$z_1$	$y_{c0}$	$y_{t0}$	$y_1$	$z_2$	$y_{c0}$	$y_{t0}$	$y_2$	$z_{19}$	$y_{c0}$	$y_{t0}$	$y_{19}$	$z_{20}$	$y_{c0}$	$y_{t0}$	$y_{20}$
1	20	22	22	1	20	22	22	0	20	22	20	0	20	22	20
1	8	12	12	1	8	12	12	0	8	12	8	0	8	12	8
1	11	11	11	0	11	11	11	...	1	11	11	11	0	11	11
0	10	15	10	1	10	15	15	0	10	15	10	1	10	15	15
0	14	18	14	0	14	18	14	1	14	18	18	1	14	18	18
0	1	4	1	0	1	4	1	1	1	4	4	1	1	4	4



**Figure 41.4 Distribution of observed test statistic. The upper and lower tails containing 10% of the of the probability mass are in the darker shade.**

the test statistic are identical to the distribution of the observed test statistic. Since all 20 possible null distributions are identical to the distribution of the observed test statistic, the probability that an observed test statistic lies in one of the tails of the null distribution must also be less than or equal to 0.10. This means that the probability of a null test statistic that is more extreme than the observed test statistic must be less than or equal to  $\alpha$ .

Fisherian hypothesis tests possess the property that the Type I error probability is less than or equal to the test's  $\alpha$  level. Yet such a property does *not* imply that the power of the test is greater than the test's Type I error probability (i.e., that the test is unbiased). To show that a test is unbiased relative to a specific class of alternative hypotheses, we now need to more precisely define the alternative to the null hypothesis.

In the previous sections, we referred to causal effects by the vector  $\tau$ , but now we define the size of causal effects using the vectors of control and treatment potential outcomes:  $y_c$  and  $y_t$ , respectively. More specifically, following Rosenbaum (2002), we define a treatment effect that is ‘larger’ than another treatment effect as follows:

**Definition 1.** *One treatment effect  $(y_c^*, y_t^*)$  has a larger effect than another treatment effect  $(y_c, y_t)$  if and only if  $y_{t,i}^* \geq y_{t,i}$  and  $y_{c,i}^* \leq y_{c,i}$  for all  $i \in \{1, \dots, N\}$  units, where  $y_t^* \neq y_t$  or  $y_c^* \neq y_c$ .*

Such an ordering of causal effects is consistent with many models of treatment effects, such as additive, multiplicative, tobit and dilated effects (Rosenbaum, 1999, 2002, 2010), not solely with the model of a constant, additive effect.

We now use this formal definition of a ‘larger effect’ in terms of potential outcomes to define a desirable property of a test statistic: ‘larger’ effects yield test statistic values greater than those produced by ‘smaller’ effects. Rosenbaum (2002, chapter 2) shows that an ‘effect-increasing’ test statistic with respect to two possible realizations of data,  $(z, y)$  and  $(z, y^*)$ , satisfies this property. Following Rosenbaum (2002), we define an effect-increasing test statistic as follows:

**Definition 2.** *A test statistic,  $t(\cdot, \cdot)$ , is effect increasing when  $t(z, y) \leq t(z, y^*)$  whenever  $y_i \leq y_i^*$  for all  $i \in \{1, \dots, N\} : z_i = 1$  and  $y_i^* \leq y_i$  for all  $i \in \{1, \dots, N\} : z_i = 0$ .*

An effect-increasing test statistic ensures that, when the null hypothesis is false and the alternative of a larger effect is true, each possible realization of data yields a test statistic value that is greater than or equal to the corresponding test statistic value when the null hypothesis is true and the alternative of a larger effect is false. To understand this property, consider the following example in Table 41.8 of two possible causal effects,  $(y_c, y_t)$  and  $(y_c^*, y_t^*)$ , in which the former is a null effect and the latter is a larger, positive causal effect (see definition 1).

**Table 41.8** No-effects,  $(y_c, y_t)$ , and positive-effects,  $(y_c^*, y_t^*)$

$y_c$	$y_t$	$y_c^*$	$y_t^*$
20	20	20	22
8	8	3	12
11	11	10	11
10	10	10	15
14	14	9	19
1	1	1	4

For any  $z$ , regardless of whichever three out of the six units are assigned to treatment, the larger causal effect,  $(y_c^*, y_t^*)$ , will always yield a value of the observed outcome for all treatment units that is greater than or equal to the outcome we would see with the same  $z$  in the no-causal-effect state, and we would also see a value of the observed outcome for all control units that is less than or equal to what we would see in the no-causal-effect state. Table 41.9 shows that an effect-increasing test statistic ensures that the observed test statistic of a larger effect is always greater than or equal to the observed test statistic of a smaller effect.

In addition to the property that increasing causal effects map monotonically onto increasing test statistics, we also want the  $p$ -values for tests of the null hypothesis when the null is false and the alternative is true to be smaller compared to the  $p$ -values when the null is true and the alternative is false. An effect-increasing test statistic also suffices for this property (for a formal proof, see Rosenbaum, 2002, chapter 2).

Table 41.10 shows that for all  $z_1, \dots, z_{20}$  possible assignments the  $p$ -value of the strong null of no effect when the positive effect,  $(y_c^*, y_t^*)$ , is true is less than or equal to the strong null’s  $p$ -value when no-effect,  $(y_c, y_t)$ , is true.

We have just shown that Fisherian tests using effect-increasing test statistics are unbiased: they provide more evidence against false claims than against true claims. In addition to being unbiased, we would also like our tests

**Table 41.9 Observed difference-in-means test statistics under all possible assignments for both a no-effects and positive-effects true causal state**

Z	No-effects $t(Z, Y)$	Positive-effect $t(Z, Y^*)$
$z_1$	4.67	8.33
$z_2$	4.00	9.67
$z_3$	6.67	10.67
$z_4$	-2.00	3.00
$z_5$	6.00	11.67
$z_6$	8.67	12.67
$z_7$	0.00	5.00
$z_8$	8.00	14.00
$z_9$	-0.67	6.33
$z_{10}$	2.00	7.33
$z_{11}$	-2.00	2.67
$z_{12}$	0.67	3.67
$z_{13}$	-8.00	-4.00
$z_{14}$	0.00	5.00
$z_{15}$	-8.67	-2.67
$z_{16}$	-6.00	-1.67
$z_{17}$	2.00	7.00
$z_{18}$	-6.67	-0.67
$z_{19}$	-4.00	0.33
$z_{20}$	-4.67	1.67

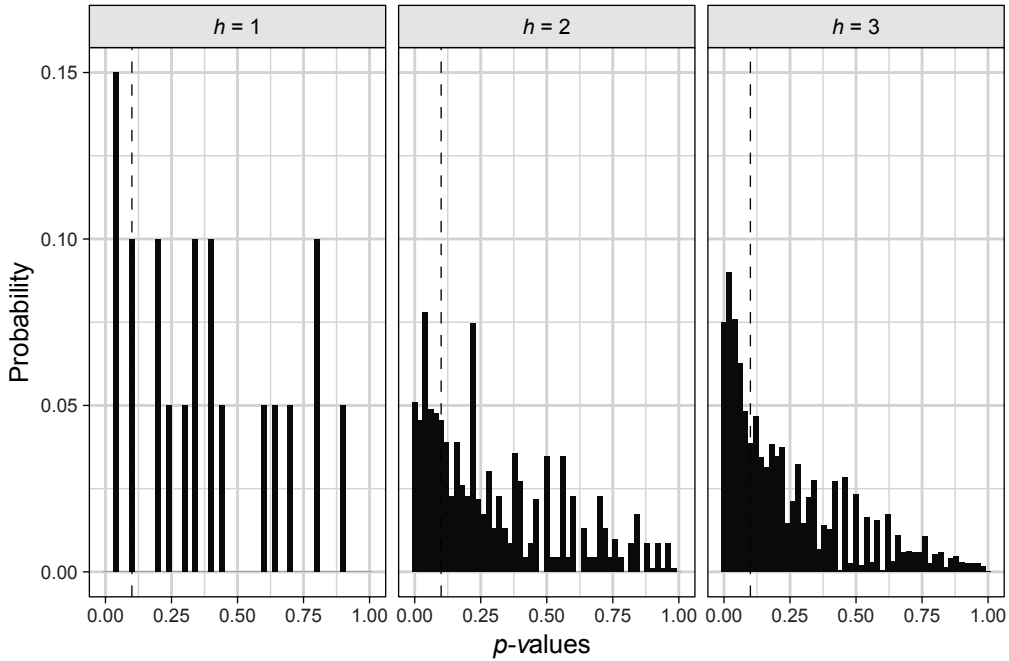
**Table 41.10 Comparing  $p$ -values for tests of the strong null of no effect under all possible assignments for both a no-effects and positive-effects true causal state**

Z	No-effects $p$ -value for $(y_c, y_t)$	Positive-effects $P$ -value for $(y_c^*, y_t^*)$
$z_1$	0.25	0.20
$z_2$	0.30	0.10
$z_3$	0.15	0.05
$z_4$	0.70	0.40
$z_5$	0.20	0.10
$z_6$	0.05	0.05
$z_7$	0.55	0.40
$z_8$	0.10	0.05
$z_9$	0.60	0.35
$z_{10}$	0.40	0.20
$z_{11}$	0.70	0.45
$z_{12}$	0.45	0.30
$z_{13}$	0.95	0.90
$z_{14}$	0.55	0.30
$z_{15}$	1.00	0.80
$z_{16}$	0.85	0.65
$z_{17}$	0.40	0.25
$z_{18}$	0.90	0.80
$z_{19}$	0.75	0.65
$z_{20}$	0.80	0.60

to be consistent, i.e., as the size of the study population grows towards  $\infty$ , while other factors remain constant, the power of the test tends to 1. Returning to the example in Table 41.1, notice that the probability of a  $p$ -value less than  $\alpha$  increases along the sequence of finite populations of increasing size, given in Table 41.3. For example, with an  $\alpha$  level of  $\alpha = 0.10$ , Figure 41.5 shows that the probability of a  $p$ -value less than  $\alpha = 0.10$  grows greater and greater.

Figure 41.5 shows that when the study population size grows from 6 to 12 and 18 units while all other factors are held constant, the power increases from 0.15 to roughly 0.2933 and 0.3741, respectively. As the population size increases along the sequence given in Table 41.3, the power of the test of the strong null of no effect will tend to 1. In other words, as we draw upon more and more data, we *reject* a *false* null with greater and greater probability.

We used the Difference-in-Means estimator as a test statistic for our example data in order to demonstrate the properties of Fisherian hypothesis tests in randomized experiments. However, we could have also used rank-based test statistics, standardized mean differences or many other functions of treatment and outcomes, most of which are effect increasing like the Difference-in-Means test statistic. In fact, one of the challenges of Fisherian testing is its flexibility in terms of test statistics. We do not engage with those decisions here but encourage interested readers to see (Rosenbaum, 2010, chapter 2), as well as Caughey et al. (2018), for some discussion on using this flexibility to assess substantively interesting hypotheses about pareto optimal causal effects, and Bowers et al. (2013, 2016) for examples on the propagation of causal effects on networks.



**Figure 41.5** Distribution of Fisherian  $p$ -values for test of strong null of no effect under all realizations of data as the size of the experimental pool grows from one copy of the study,  $h = 1$  ( $n = 6$ ); to two copies of the study,  $h = 2$  ( $n = 12$ ); to three copies of the study,  $h = 3$  ( $n = 18$ ). The dashed lines denote the significance level of  $\alpha = 0.10$ .

### *Hypothesis Testing in the Neymanian Tradition*

While Fisherian tests allow the use of many different kinds of test statistics, Neymanian hypothesis tests are much more closely related to the estimation of mean causal effects (see the third section). In any given study, one can observe only a single estimate, now interpreted as a test statistic. However, a researcher can postulate (provisionally, for the sake of argument) a weak null hypothesis,  $H_0 : \bar{\tau} = \bar{\tau}_0$ , relative to some alternative hypothesis, such as  $H_a : \bar{\tau} > \bar{\tau}_0$ , and subsequently assess the probability of a test statistic under the weak null more extreme than the test statistic the researcher actually observed. Such Neymanian hypothesis tests differ from Fisherian tests in several fundamental ways. Neymanian hypothesis tests

(1) hypothesize about the mean causal effect, not the individual causal effect for each unit in the study, (2) require that the Difference-in-Means estimator be unbiased such that a hypothesis about the mean causal effect implies the same value for the mean (i.e., expected value) of the estimator, (3) require that the researcher estimate the variance of the Difference-in-Means estimator (recall that the distribution of the test statistic in the Fisherian test is known under random assignment and a strong null hypothesis) and (4) draw upon the finite population central limit theorem (Erdős and Rényi, 1959; Hájek, 1960; Li and Ding, 2017), which implies that the product of  $\sqrt{N}$  multiplied by the difference between the estimator and its expected value converges to a normal distribution with mean equal to 0 and variance equal to  $\sigma_{\bar{\tau}}^2$ , which, Due to Slutsky's theorem, one can

equivalently state this property as . . .  $\frac{\hat{\tau} - \mathbb{E}[\hat{\tau}]}{\sqrt{\hat{\sigma}_{\hat{\tau}}^2}}$ , which is known as the z-score, converges in distribution to a standard normal (i.e., normal distribution with mean 0 and variance equal to 1).

We can see points (1)–(4) by looking at the common expressions for Neymanian  $p$ -values:

$$\begin{aligned} p_u &= 1 - \Phi\left(\frac{\hat{\tau} - \bar{\tau}_0}{\sqrt{\hat{\sigma}_{\hat{\tau}}^2}}\right) \\ p_l &= \Phi\left(\frac{\hat{\tau} - \bar{\tau}_0}{\sqrt{\hat{\sigma}_{\hat{\tau}}^2}}\right) \\ p_t &= 2\left(1 - \Phi\left(\frac{|\hat{\tau} - \bar{\tau}_0|}{\sqrt{\hat{\sigma}_{\hat{\tau}}^2}}\right)\right), \end{aligned} \tag{12}$$

where  $\hat{\tau}$  is the familiar Difference-in-Means estimator (from the third section) now interpreted as a test statistic, not an estimator,  $\hat{\sigma}_{\hat{\tau}}^2$  is the conservative variance estimator (also from the third section) now used to describe the distribution of the Difference-in-Means test statistic under the weak null hypothesis rather than the precision of an estimator,  $\bar{\tau}_0$  is a weak null hypothesis and  $\Phi(\cdot)$  is the standard normal cumulative distribution function (CDF).

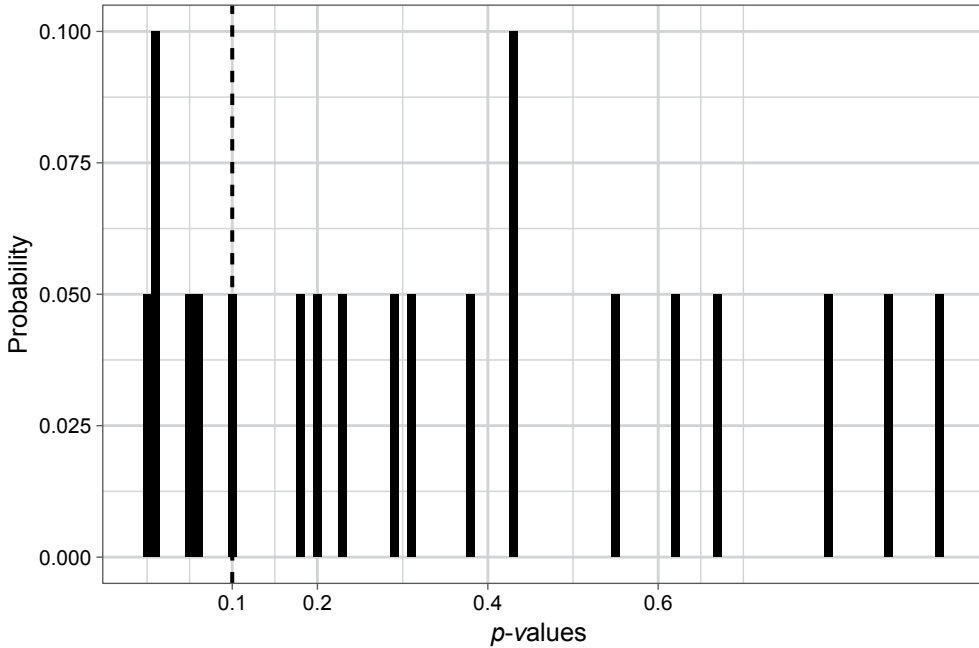
The expressions in Equation (12) return the probability of Difference-in-Means test statistic at least as extreme as the observed test statistic if the weak null hypothesis,  $\bar{\tau}_0$ , were true. Notice, though, that the weak null hypothesis,  $\bar{\tau}_0$ , is technically a claim about the mean of the Difference-in-Means test statistic, which we can denote by  $\mathbb{E}_0[\hat{\tau}]$ . However, because the Difference-in-Means estimator is unbiased, its expected value is always equal to the mean causal effect; this is why we use  $\bar{\tau}_0$  in Equation (12) rather than  $\mathbb{E}_0[\hat{\tau}]$ . Finally, note that the true variance of the Difference-in-Means test

statistic is unknown, but the normal CDF requires two arguments – a value for the mean and a value for the variance – to assign a probability to null test statistics as least as extreme as the observed test statistic. Rather than postulate a hypothesis about the variance of the estimator, like one does for the mean of the estimator, Neymanian tests use an estimate from the conservative variance estimator in Equation (9) to calculate a  $p$ -value via the standard normal CDF.

In many situations, we can easily justify the assumption that  $\frac{\hat{\tau} - \bar{\tau}_0}{\sqrt{\hat{\sigma}_{\hat{\tau}}^2}}$  is well approximated by a standard normal distribution by appealing to the aforementioned finite population Central Limit Theorem (CLT) and associated theory (see, e.g., Hogg, 1978). The Difference-in-Means test statistic scaled by  $\sqrt{N}$  is indeed asymptotically normal and the  $p$ -values in Equation (12) are all asymptotically valid – roughly meaning that, as . . .  $N$  grows towards  $\infty$ , the probability of a Type I error is less than or equal to the  $\alpha$  level of the test.

However, in small experiments in which the normal approximation is poor, tests of a weak null hypothesis relative to an alternative may have either a Type I error probability greater than the test’s  $\alpha$  level (when the null is true) or low power (when the null is false). Table 41.5 and Figure 41.3 demonstrate Neymanian  $p$ -values, in which the assignment  $\mathbf{z}_8$  happened to be the one randomly drawn by the researcher. In this example, the observed Difference-in-Means test statistic is 11.6667 and the conservatively estimated variance is 12.8889. If we were to test the weak null hypothesis of no effect, i.e., that  $\bar{\tau}_0 = 0$ , against the alternative hypothesis that  $\bar{\tau}_a > 0$ , then the upper one-tailed  $p$ -value would be as follows:

$$\left(1 - \Phi\left(\frac{11.6667 - 0}{\sqrt{12.8889}}\right)\right) \approx 0.0006, \tag{13}$$



**Figure 41.6** Distribution of Neymanian  $p$ -values under all realizations of data. The dashed line is the observed  $p$ -value

which yields a smaller  $p$ -value than the upper  $p$ -value we calculated via the Fisherian test of the strong null of no effects (which was  $p = 0.05$ ). Figure 41.6 illustrates all upper  $p$ -values for a test of the weak null over all 20 possible realizations of data when the true effects are as given in Table 41.1.

In this particular case in which the weak null hypothesis is false, we can see that the Neymanian hypothesis test has high power. But in general, Neymanian tests can have bad properties in small experiments, such as a Type I error probability greater than  $\alpha$ . For example, imagine that the weak null hypothesis of no effect were true as is depicted in Table 41.11.

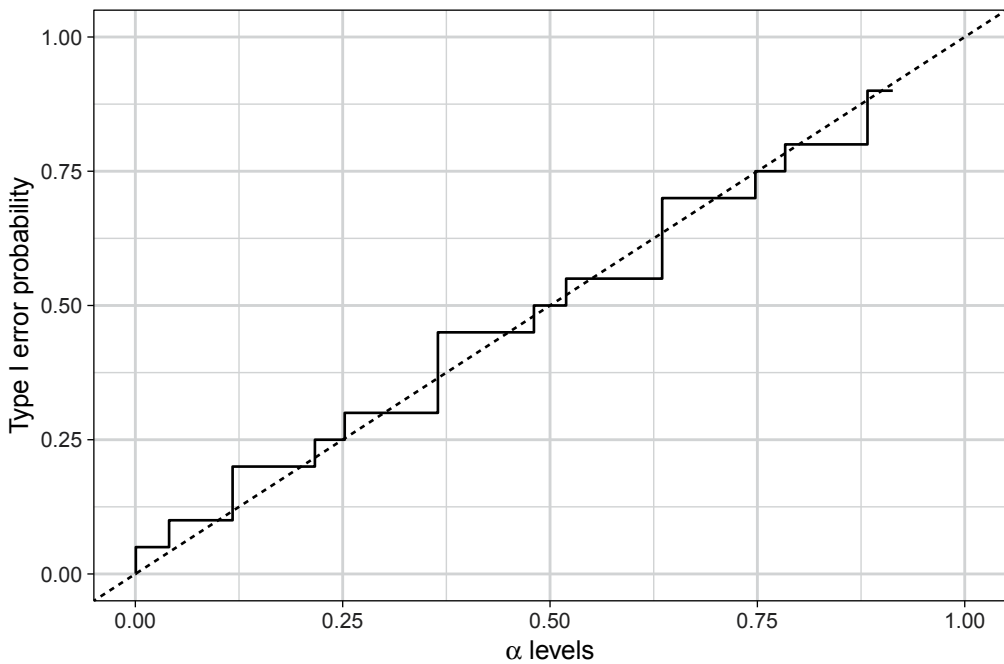
In this hypothetical experiment in which the weak null is true (and the strong null also happens to be true but need not be), Figure 41.7 shows that for some  $\alpha$  levels like  $\alpha = 0.05$ , the type I error probability is greater than the test's  $\alpha$  level (the solid line is above the dotted line). For other  $\alpha$  levels, the

**Table 41.11** Values of  $y_c$ ,  $y_t$  and  $\tau$  when weak null hypothesis is true

$y_c$	$y_t$	$\tau$
22	22	0
8	8	0
11	11	0
15	15	0
18	18	0
1	1	0

type I error probability is less than the  $\alpha$  level (the solid line is below the dotted line), which makes the test at that level a conservative test. In short, although Neymanian hypothesis tests are asymptotically valid, such tests (particularly in small experiments) may yield type I error probabilities that do not fulfill the 'promise' made by a given  $\alpha$  level.

An additional implication of the differences between Neymanian and Fisherian hypothesis tests is that, although the former is consistent,<sup>9</sup> it is not necessarily unbiased in finite contexts,



**Figure 41.7** Distribution of type I error probabilities for different  $\alpha$  levels

even when the Difference-in-Means test statistic is well approximated by a normal distribution. For intuition on this point, note that when the alternative of a larger effect is true and the null of no effects is false, the Difference-in-Means test statistic will yield values that are systematically larger than the values it would produce if the null of no mean effect were true – larger causal effects lead to larger test statistic values. Yet the alternative of a larger effect could be such that when it is true, the *variance estimates* are systematically larger than what the same variance estimates would be if the null were true – e.g., if a positive mean causal effect is caused by a few outliers that react strongly to treatment. Since the  $z$ -score scales the Difference-in-Means test statistics by the variance estimates, the systematically greater variance estimates when the alternative is true could yield *smaller*  $z$ -scores compared to when the null is true. Smaller  $z$ -scores yield larger  $p$ -values; hence, Neymanian hypothesis tests of weak causal hypotheses are not necessarily unbiased in finite contexts, even when the

assumption of normality approximately holds. For more on Fisherian versus Neymanian hypothesis tests, see Ding (2017), as well as the discussion from Aronow and Offer-Westort (2017), Chung (2017), Bailey (2017) and Loh et al. (2017).

Up to this point, we have discussed the role that research design plays in the quality of procedures for causal inference: estimation and testing. To simplify the exposition, we have been referring to situations in which the researcher completely controls and thus knows the research design. We now consider situations the researcher does not control or only partially controls the research design.

### **PARTIALLY CONTROLLED RESEARCH DESIGNS: NON-COMPLIANCE AND ATTRITION**

We refer to designs with imperfect compliance and/or attrition as partially controlled



designs, in that the researcher *does* control the probability distribution on the set of possible assignments but does *not* control whether units actually comply with the assigned treatment or report their outcomes. Causal inferences in such partially controlled designs often require more assumptions to make causal inferences, which are typically defined no longer on the whole study population but a specific stratum of units in the study.

**Non-compliance**

Non-compliance occurs when units who are assigned to receive treatment or control do not actually receive it, where the random variable  $D_i \in \{0,1\}$  is an indicator variable for whether unit  $i$  has received or not received the assigned treatment. We are often substantively interested in the causal effect of actual receipt of treatment and not its mere assignment. Under complete uniform random assignment, we have shown above that the probability  $Z_i = 1$  is identical for all  $i = 1, \dots, N$  units. Yet, we're interested in the causal effect of  $D_i$ , and since  $D_i$  is an outcome of  $Z_i$  (i.e., measured after  $Z_i$ ), the probability that  $D_i = 1$  is not necessarily identical for all units. Hence, a naive estimator of the difference in observed outcomes between those who did and did not receive the treatment (i.e., the per-protocol estimator) is not necessarily unbiased.

Since whether or not units actually receive (or comply with) the treatment is an outcome variable measured after assignment, we can define units' compliance status in terms of their unobservable potential outcomes (Table 41.12).

**Table 41.12 Compliance strata**

$z_i = 0$	$z_i = 1$	Stratum
$d_{c,i} = 1$	$d_{t,i} = 1$	<i>Always Taker</i>
$d_{c,i} = 0$	$d_{t,i} = 1$	<i>Complier</i>
$d_{c,i} = 1$	$d_{t,i} = 0$	<i>Defier</i>
$d_{c,i} = 0$	$d_{t,i} = 0$	<i>Never Taker</i>

Notice that the probability that  $D_i = 1$  for an *Always Taker* is 1 and the probability that  $D_i = 1$  for a *Never Taker* is 0. The only strata within which the probabilities of receiving treatment are identical and on the interval  $(0, 1)$  are *Compliers* and *Defiers*.

Angrist et al. (1996) show that scholars can reliably infer the mean causal effect among *Compliers* under three further assumptions in addition to SUTVA and a uniform probability distribution on the set of assignments,  $\Omega$ , which is sometimes referred to as the exogeneity of the instrument,  $\mathbf{Z}$ . These three assumptions, in addition to SUTVA and uniform random assignment, are:

- 1 Exclusion restriction – i.e., the instrument affects the outcome only through the receipt of treatment.
- 2 No Defiers – i.e., that there are no units for which  $d_{c,i} = 1$  and  $d_{t,i} = 0$ .
- 3 At least one Complier – i.e., there exists at least one unit in the study for which  $d_{c,i} = 0$  and  $d_{t,i} = 1$ .

To see this point, note that one can write the mean causal effect as a sum of the mean causal effects among the four strata (*Always-Takers*, *Compliers*, *Defiers* and *Never-Takers*), weighted by the proportion of units in each stratum:

$$\bar{\tau} = \delta_{AT}\pi_{AT} + \delta_C\pi_C + \delta_D\pi_D + \delta_{NT}\pi_{NT},$$

where  $\delta_s, \pi_s: s \in \{AT, C, D, NT\}$  represent the mean causal effect in each stratum and the proportion of units in that stratum, respectively.

By the exclusion restriction assumption, the causal effect of  $Z$  on  $Y$  must be 0 for *Always Takers* and *Never Takers* (i.e.,  $\delta_{AT} = 0$  and  $\delta_{NT} = 0$ ). By the assumption of no *Defiers*, the proportion of *Defiers*,  $\pi_D$ , is 0. Hence, the mean causal effect among *Compliers* is

$$\bar{\tau} = \delta_C\pi_C$$

$$\frac{\bar{\tau}}{\pi_C} = \delta_C,$$

i.e., the mean causal effect on all units scaled by the proportion of Compliers is equal to the mean causal effect among Compliers.

We showed in the section on unbiasedness that under complete uniform random assignment, the Difference-in-Means estimator is unbiased regardless of the distributions of potential outcomes. Both  $\mathbf{Y}$  and  $\mathbf{D}$  are outcomes of  $\mathbf{Z}$ ; hence, the Difference-in-Means estimators of  $\hat{\tau}(\mathbf{Z}, \mathbf{Y})$  and  $\hat{\tau}(\mathbf{Z}, \mathbf{D})$  unbiasedly estimate  $\bar{\tau}$  and  $\pi_c$ , respectively. The Wald (or IV) estimator (Wald, 1940), which is also sometimes referred to as the Bloom estimator (Bloom, 1984) or the CACE (Complier Average Causal Effect) estimator (Gerber and Green, 2012), is defined as  $\frac{\hat{\tau}(\mathbf{Z}, \mathbf{Y})}{\hat{\tau}(\mathbf{Z}, \mathbf{D})}$ . In

short, it is the ratio of these two unbiased estimators. This ratio estimator consistently, though not necessarily unbiasedly, estimates  $\frac{\bar{\tau}}{\pi_c}$  (see Angrist and Pischke, 2008, chapters 4.6 and 4.7).

A design-based test of the hypothesis of no mean causal effect among *Compliers* is difficult due to the absence of an analytic expression for the variance of the CACE ratio estimator (for more on this topic, see Imbens and Rosenbaum, 2005; Kang et al., 2018). On the other hand, tests of the null hypothesis that the causal effect is 0 for all Compliers is relatively straightforward.

The null hypothesis of no causal effect among Compliers implies the strong null of no effect for all units under the assumptions of the exclusion restriction and no Defiers. Since these two assumptions jointly imply that the individual causal effect is 0 for all Always Takers and Never Takers, and that there are no Defiers, then no causal effect among Compliers implies no causal effect among all units. For example, let's imagine that (like in Table 41.5)  $\mathbf{z}_8$  was the assignment that the researcher happened to randomly

**Table 41.13 Realization of data if  $\mathbf{z}_8$  were the randomly drawn assignment**

$z_8$	$y_c$	$y_t$	$y_8$	$d_8$	$d_c$	$d_t$
1	?	22	22	0	?	0
0	8	?	8	0	0	?
0	11	?	11	0	0	?
1	?	15	15	1	?	1
1	?	18	18	0	?	0
0	1	?	1	1	1	?

draw, except that now we observe the following imperfect compliance, as shown in Table 41.13.

We know that if  $d_{t,i} = 0$ , then, by the no-Defiers assumption, that unit must be a Never Taker, and if  $d_{c,i} = 1$ , then, by the same assumption, that unit must be an Always Taker. We don't know, however, whether a unit  $i$  such that  $z_i = 1$  and  $d_{\{t,i\}} = 1$  is a Complier or Always Taker, and we don't know whether a unit  $i$  such that  $z_i = 0$  and  $d_{\{c,i\}} = 1$  is a Complier or Never Taker. Yet under the null hypothesis of no causal effect among Compliers, the individual causal effect is 0 regardless of whether a given unit is a Complier, Always Taker or Never Taker. Hence, we can fill in missing potential outcomes without knowing those units' compliance strata, as in Table 41.14.

One can now use an effect-increasing test statistic, such as the Difference-in-Means test statistic, to test the hypothesis of no Complier causal effect against either the alternative of a positive Complier causal effect or a negative Complier causal effect. Hansen and Bowers

**Table 41.14 Null potential outcomes for a test of the strong null of no effect if  $\mathbf{z}_8$  were realized assignment**

$z_8$	$y_{c,0,8}$	$y_{t,0,8}$	$y_8$	$d_8$	$d_c$	$d_t$
1	22	22	22	0	0	0
0	8	8	8	0	0	?
0	11	11	11	0	0	?
1	15	15	15	1	?	1
1	18	18	18	0	0	0
0	1	1	1	1	1	1

(2009) present an application of this idea in the context of the one-sided compliance in a cluster-randomized get-out-the-vote campaign with a binary outcome, replacing what would be a complex two-stage logistic multilevel model with a relatively simple Fisherian hypothesis test. Imbens and Rosenbaum (2005) show how the Fisherian approach produces valid confidence intervals, where the Neyman-style approach using two-stage least squares fails to control the type I error probability when the instrument is weak (i.e., there are few Compliers).

Under the assumptions of the exclusion restriction and no Defiers, we do not need to know which units are Compliers in order to assert the hypothesis of no Complier causal effect. However, if one were to posit a hypothesis other than no Complier causal effect, one would also need to posit a hypothesis about which units are Compliers and which are not. Hypothesis testing with imperfect compliance is therefore more complicated when testing hypotheses other than that of no Complier causal effect (for more on Fisherian approaches to instrumental variable analysis, see Kang et al., 2018; Rosenbaum, 1996, among others).

**Attrition or Missing Outcomes**

Our second step away from the ideal case of complete control over the research design is to allow for the possibility of missing outcomes. Let  $r_{t,i}$  be an indicator, i.e.,  $r_{t,i} \in \{0,1\}$ , for whether subject  $i$  would respond to an attempt to measure an outcome if assigned to treatment, and let  $r_{c,i} \in \{0,1\}$  be an indicator for whether subject  $i$  would respond if assigned to control. We can represent whether an individual’s outcomes are missing or not, based on Equation (14):

$$Y_i = \begin{cases} y_{c,i} + [y_{t,i} - y_{c,i}]Z_i & \text{if } R_i = 1 \\ \text{NA} & \text{if } R_i = 0, \end{cases} \quad (14)$$

where  $R_i = Z_i r_{t,i} + (1 - Z_i)r_{c,i}$ .

From Equation (14), we can see that if  $R_i = 1$ , then the researcher will observe  $y_{c,i}$  for unit  $i$  if  $Z_i = 0$  and  $y_{t,i}$  for unit  $i$  if  $Z_i = 1$ . By contrast, if  $R_i = 0$ , then  $Y_i$  will be unobserved – i.e., NA.

We can define four distinct strata of subjects (see Table 41.15) with regard to attrition in order to help us understand how attrition can affect the properties of estimators and hypothesis tests. Just as in the case of imperfect compliance, we can only infer the causal effects on certain subgroups when outcomes are missing (even when assignments are randomized).

In the context of attrition, can we define a set of assumptions, as we did under imperfect compliance, in which estimators and tests satisfy good’ properties on some subset of the experimental data? It turns out that we can only do so if the question of whether a unit attrits or not is independent of treatment assignment. Without further assumptions, we know only that the random variable  $R_i$  is independent of the random variable  $Z_i$  among only *Always Reporters* and *Never Reporters*.<sup>10</sup> Therefore, if we assume that all experimental units belong to one of these two types, then our estimators and tests maintain the properties they should among the set of *Always Reporters*.<sup>11</sup> (We cannot observe outcomes for *Never Reporters* and hence cannot estimate or test hypotheses about causal effects on units that are *Never Reporters*.) On this set of *Always Reporters*, one can estimate causal effects and test strong or weak causal hypotheses, as we did above.

**Table 41.15 Attrition strata**

$Z_i = 0$	$Z_i = 1$	Stratum
$r_{c,i} = 1$	$r_{t,i} = 1$	<i>Always Reporter</i>
$r_{c,i} = 0$	$r_{t,i} = 1$	<i>If Treated Reporter</i>
$r_{c,i} = 1$	$r_{t,i} = 0$	<i>If Untreated Reporter</i>
$r_{c,i} = 0$	$r_{t,i} = 0$	<i>Never Reporter</i>

## UNCONTROLLED RESEARCH DESIGNS: OBSERVATIONAL STUDIES

In this section, we finally relax the assumption that the researcher has control over how the treatment variable is assigned. The key distinction between experimental and observational studies is that in a randomized experiment, the researcher knows the probabilities with which units are *assigned* to treatment and control conditions; however, in an observational study, the researcher observes units only after they have *selected* into study conditions with unknown probabilities. How, then, is one to generate statistical inferences using estimators and tests focusing on causal effects when the probability distribution on the set of assignments,  $\Omega$ , is unknown? A common design-based approach to this problem is to define units' treatment assignment probabilities as an unknown function of a set of baseline covariates. In the ideal (and often unattainable) case, by appropriate conditioning on these baseline covariates, the researcher can estimate and test hypotheses about causal effects via procedures that meet the desirable properties described at the outset of this chapter. Much of the work on observational studies emphasizes appropriate conditioning on baseline covariates, as well as methods to diagnose the success of such conditioning strategies (e.g., Hansen and Bowers, 2008; Hartman and Hidalgo, 2018, among others). Realistically, the design-based approach in observational studies might be called an 'as-if-randomized' approach, e.g., a researcher might make choices about comparison groups such that within a group, treatment selection *appears* random.

The model of an observational study states that units are individually assigned to treatment or control by  $N$  *independent* (but not necessarily *identically distributed*) coin tosses. More specifically, for all

$i \in \{1, \dots, N\}$  units, we let  $\Pr(Z_i = 1)$  be equal to  $\lambda(\mathbf{x}_i)$ , where  $\lambda(\cdot)$  is an unknown function and  $\mathbf{x}_i$  is unit  $i$ 's fixed vector of baseline covariate values. Even if we don't know  $\lambda(\cdot)$ , if  $\mathbf{x}_i = \mathbf{x}_j$  for any two units  $i$  and  $j \neq i$ , then it follows that  $\Pr(Z_i = 1) = \Pr(Z_j = 1)$ . Of course, we still don't know the function  $\lambda(\cdot)$  and hence don't know the actual values of  $\Pr(Z_i = 1)$  and  $\Pr(Z_j = 1)$ ; we know only that these two values are equal. Therefore, if we construct a block,  $b$ , that consists of treatment unit  $i$  and control unit  $j \neq i$  (or vice versa) then each possible assignment within that block has an equal probability of realization. If each possible assignment has an equal probability, then an observational study can be analyzed as if it is a uniform, block randomized experiment (for more on the analysis of block, randomized experiments, see Gerber and Green, 2012, chapter 4). This approach does not directly estimate  $\lambda(\cdot)$ , although there are alternative approaches that do so and subsequently use these estimated values (known as estimated propensity scores) as a basis for inference (see, e.g., Robins et al., 2000).

Scholars can therefore attempt to make an observational study as experiment-like as possible by creating matched blocks on the basis of observed covariates that the researcher believes to determine units' treatment assignment probabilities. A range of matching algorithms exist to improve covariate balance and hence make observational studies like experiments as much as possible (at least on the basis of observed covariates). For more on this topic, see Hansen (2004), Diamond and Sekhon (2013), Sävje et al. (2017) and Zubizarreta (2012), among others. After matching (or some other form of covariance adjustment) and favorable comparisons with randomized designs, researchers then must confront the fact that their observational studies are not randomized studies. This leads directly to the topic of sensitivity analyses.

**SENSITIVITY TO ASSUMPTIONS  
IN UNCONTROLLED RESEARCH  
DESIGNS**

Thus far, we have considered cases in which either the probability distribution on  $\Omega$  is known by random assignment or units' assignment probabilities are a function of only *observed* covariates. But in an observational study, we rarely know – let alone observe – all relevant covariates. We now consider deviations from the assumption that units' assignment probabilities are determined by only observed covariates and subsequently assess how one's inferences would change under violations of this assumption.

A powerful, design-based framework for such a sensitivity analysis is given by Rosenbaum (2002, chapter 4). Before explaining this framework, we need to define a few additional terms. First, the *treatment odds* for unit  $i \in \{1, \dots, N\}$  is  $\frac{\pi_i}{(1 - \pi_i)}$ , which is simply the

$i$ th unit's probability of assignment to treatment divided by that unit's probability of assignment to control. The *treatment odds ratio* for any two units  $i$  and  $j \neq i$  is simply the ratio of the  $i$ th unit's treatment odds and the  $j$ th unit's treatment odds. If units' treatment odds are a function of only observed covariates and the researcher is able to obtain balance on all of these observed covariates, then the treatment odds for units  $i, j \neq i : \mathbf{x}_i = \mathbf{x}_j$  is identical and their treatment odds ratio is 1.

Rosenbaum (2002, chapter 4) considers what happens when units' treatment odds are a function not only of observed covariates,  $\mathbf{x}$ , but also an unobserved covariate,  $\mathbf{u}$ . Under the assumption of a logistic functional form between all units' treatment odds and baseline covariates, as well as the constraint that  $0 \leq u_i \leq 1$  for all  $i$ , one can write the treatment odds of the  $i$ th unit as follows:

$$\frac{\pi_i}{(1 - \pi_i)} = \exp\{\kappa(\mathbf{x}_i) + \gamma u_i\}$$

$$\log\left(\frac{\pi_i}{(1 - \pi_i)}\right) = \kappa(\mathbf{x}_i) + \gamma u_i,$$

where  $\kappa(\cdot)$  is an unknown function and  $\gamma$  is an unknown parameter, and the treatment odds ratio for units  $i$  and  $j$  is  $j \neq i$ :

$$\frac{\left(\frac{\pi_i}{1 - \pi_i}\right)}{\left(\frac{\pi_j}{1 - \pi_j}\right)} = \frac{\exp\{\kappa(\mathbf{x}_i) + \gamma u_i\}}{\exp\{\kappa(\mathbf{x}_j) + \gamma u_j\}}$$

$$= \exp\left\{\left(\kappa(\mathbf{x}_i) + \gamma u_i\right) - \left(\kappa(\mathbf{x}_j) + \gamma u_j\right)\right\}.$$

If  $\mathbf{x}_i = \mathbf{x}_j$ , then  $\kappa(\mathbf{x}_i) = \kappa(\mathbf{x}_j)$  and, hence, the treatment odds ratio is simply:

$$\exp\left\{\gamma(u_i - u_j)\right\}.$$

Since  $u_i, u_j \in [0, 1]$ , the minimum and maximum possible values of  $(u_i - u_j)$  are  $-1$  and  $1$ . Therefore, the minimum and maximum possible values of the treatment odds ratio are  $\exp\{-\gamma\}$  and  $\exp\{\gamma\}$ . After noting that  $\exp\{-\gamma\} = \frac{1}{\exp(\gamma)}$ , we can bound the treatment

odds ratio between  $i$  and  $j$  as follows:

$$\frac{1}{\exp(\gamma)} \leq \frac{\left(\frac{\pi_i}{1 - \pi_i}\right)}{\left(\frac{\pi_j}{1 - \pi_j}\right)} \leq \exp\{\gamma\}. \quad (15)$$

We can denote  $\exp\{\gamma\}$  by  $\Gamma$  and subsequently consider how one's inferences would change for various values of  $\Gamma$ .

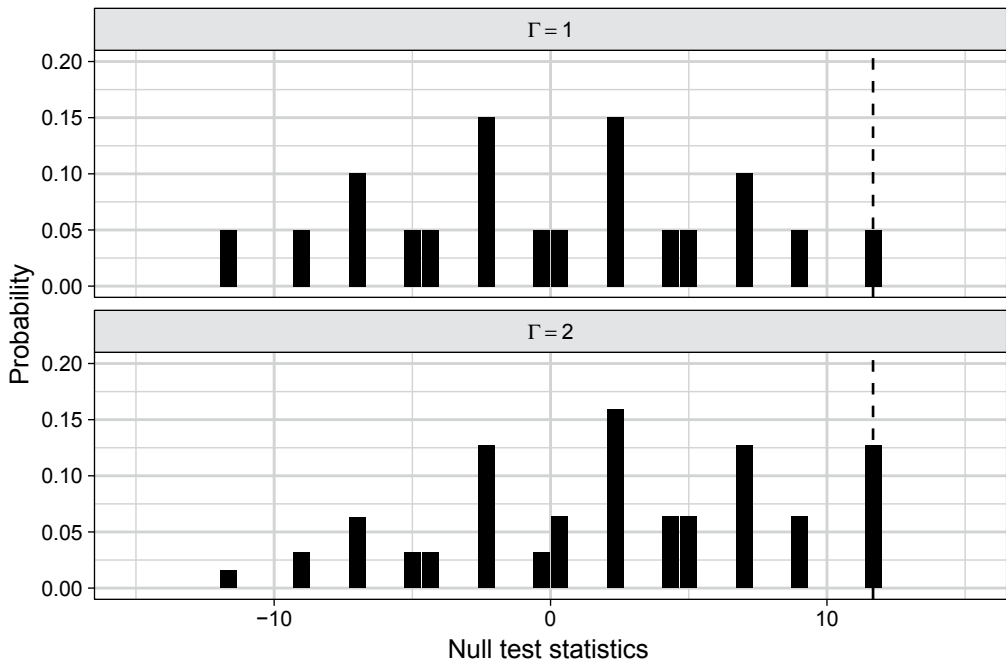
For example, let's say that a researcher obtains balance via stratification on all observed covariates – such that the

design closely resembles a uniform, block randomized experiment – and subsequently tests a strong null hypothesis under the assumption that all units’ treatment odds are identical. Now the researcher considers deviations from this assumption. Different assumptions about  $u$  and  $\gamma$  imply differing probabilities of possible assignments, which, as Rosenbaum (2002, chapter 4) shows, can be represented by

$$\Pr(\mathbf{Z} = \mathbf{z}) = \frac{\exp\{\gamma \mathbf{z}' \mathbf{u}\}}{\sum_{\mathbf{z} \in \Omega} \exp\{\gamma \mathbf{z}' \mathbf{u}\}}. \quad (16)$$

To return to the working example from Table 41.1, let’s imagine that the realized assignment was  $\mathbf{z}_8$ , which yielded an observed test statistic of 11.6667 and a  $p$ -value of 0.05 (see Figure 41.3). We calculated that

$p$ -value under the assumption that all units had identical probabilities of assignment. However, in an observational study, it could be the case that the strong null of no effects is true and that there is an unobserved covariate,  $u$ , such that units do not have identical assignment probabilities. A sensitivity analysis permits us to directly assess how our inferences would change under such a scenario. For example, we might first assume that  $\Gamma = 2$ , which implies that  $\gamma = \log(2) \approx 0.6931$ . The quantity  $\gamma$  is the coefficient of a unit’s unobserved baseline covariate  $u_i$ . If all units have the same value of this unobserved covariate — i.e.,  $u_i = u_j$  for all  $i \neq j$  — then all units’ assignment probabilities will remain identical. A conservative approach therefore might instead find a vector  $\mathbf{u}$  that, given a value of  $\gamma$ , will maximize the  $p$ -value of a test



**Figure 41.8** Distributions of Difference-in-Means test statistic under strong null of no effect when  $\mathbf{z}_8$  is the realized assignment under (a) a model where an unobserved covariate has no effect on odds of treatment,  $\Gamma = 1$ , and (b) a model where an observed covariate doubles the odds of treatment,  $\Gamma = 2$ . The dashed lines denote the value of the observed test statistic.

of the strong null hypothesis of no effect. In this particular example,  $\mathbf{u}' = [1\ 0\ 0\ 1\ 1\ 0]$  maximizes the  $p$ -value for an upper one-sided test of the strong null hypothesis of no effect. In general, different procedures exist for finding the vector  $\mathbf{u}$  that maximizes (or minimizes) the  $p$ -value for a given value of  $\Gamma$  (see Gastwirth et al., 2000; Rosenbaum, 2018; Rosenbaum and Krieger, 1990). Figure 41.8 illustrates how the respective null distributions would differ under  $\Gamma = 1$  and  $\Gamma = 2$  in which  $\mathbf{u}' = [1\ 0\ 0\ 1\ 1\ 0]$ .

Notice that under  $\Gamma = 1$  and  $\Gamma = 2$ , the observed test statistic remains fixed at 11.6667. The set of 20 null test statistic also remains fixed for  $\Gamma = 1$  and  $\Gamma = 2$ . The value of  $\Gamma$  changes only the probability associated with each of the null test statistics. As we can see, when  $\Gamma = 1$ , the  $p$ -value is 0.05, but when  $\Gamma = 2$ , the  $p$ -value increases to approximately 0.1270. The researcher is no longer able to reject the strong null of no effect at a level of  $\alpha = 0.10$  when  $\Gamma = 2$  compared to when  $\Gamma = 1$ , and this is one way in which researchers can assess the sensitivity of their inferences to assumptions about the research design.

This approach is not the only way to formalize the impact of the assumptions underlying the ‘as-if randomized’ research designs used for causal inference when researchers have little to no control over the selection process of the main explanatory variable. See also Hosman et al. (2010) for an approach focusing on regression coefficients and regression-based adjustment, as well as an application by Chaudoin et al. (2018) of ideas like these to problems in international relations.

## CONCLUSION

Statistics and research design help us learn about general and abstract social science theory using concrete and specific observations. Observation helps us learn about

theory, but observation occurs using the tools of research design and they are summarized, described and interpreted using the tools of statistics. In fact, we have shown that certain counterfactual causal quantities can never be directly observed, and that our ability to report with confidence about such unobserved causal effects depends heavily on statistical tools, which, in turn, depend on research design for their operation. The persuasiveness of claims about links from an estimate or  $p$ -value to an unobserved causal effect and general theory depends on the clarity of each step. At the most nitty-gritty level, we want our tools to work well – our estimators and tests should err in known and controlled manners and should err rarely.<sup>12</sup> We have asked, ‘How can we know that tools do the work that we want them to do?’ Additionally, we have shown how to use the facts of research design to answer that question and to justify use of these common tools. This means that when we want to persuade an audience that our findings support a given theory (or urge modification of it), we do not need our audiences to believe that (1) we have a random sample from a well-defined population, (2) that the outcome arises from some known probability process (like a normal or zero-inflated Poisson distribution) or (3) that the treatment or selection process relates to background covariates in some known (often linear and additive) fashion. Instead, in a randomized experiment, we ask that a reader believe that our research design correctly describes the physical processes that occurred in the research (a request that is not hard to verify and assess). In an observational study, we ask readers to agree that the as-if randomized approach is reasonable, and we present direct evidence comparing observational designs to randomized experiments in order to make the provisional as-if randomized approach easier to assess.

In explaining reliable procedures, we have used a very simple set of examples.

Our simplifications include the use of only two experimental conditions (treatment and control); yet the general modes of inference described in this chapter can be straightforwardly applied to factorial experiments and other contexts with multiple treatments (see Dasgupta et al., 2015). A second such simplification has been the Stable Unit Treatment Value Assumption (SUTVA) (Cox, 1958; Rubin, 1980, 1986), which, in the case of a binary treatment variable, implies that all units have only two potential outcomes. Yet both estimation (see Aronow and Samii, 2017; Hudgens and Halloran, 2008) and testing (see Bowers et al., 2013, 2016; Rosenbaum, 2007) are possible when units have more than two potential outcomes due to interference between units, for example, in the context of experiments on networks. Furthermore, we have focused on randomized experiments and only briefly pointed to the strategy of ‘as-if randomized’ approaches to estimation and testing in observational studies. We explained that such approaches, when paired with sensitivity analyses, can enable persuasive statistical inferences about causal effects when the researcher lacks control over the design. In other words, once an observational research design compares favorably to the standard of an equivalent experiment (the way that a matched design can be compared to a block-randomized experiment), statistical inference about causal effects can use the same procedures, provisionally justified in the same way, as in a randomized experiment if also followed by a sensitivity analysis.

This focus on the basics – on ensuring that our statistical tools do what they should – leaves larger questions unexplored. For example, some researchers would prefer to make inferences about not only counterfactual causal effects among units in a given study but also about future units in data contexts that differ from the one under study. One might desire unbiased and consistent estimators, as well as valid and powerful

tests, not only based on the research design generating the data collected here and now, but also for unknown future research designs guiding data collection elsewhere and at other times. Forecasting causal effects is an active research area (for only a few recent works on the topic, see Bisbee et al., 2017; Coppock et al., 2018; Dehejia et al., 2019; Pearl, 2015; Stuart et al., 2015). Whether or not a researcher or policy-maker would like a formal forecast of the causal effects of an intervention from one study to a new context (in time, space and/or units), information provided by a single study to the motivating theoretical question still depends on the reliability of the tools used to conduct and analyze the study. We have focused on showing how the reliability of such procedures is based on the research design itself and leave questions about the properties of procedures for forecasting causal effects as a separate, though important, topic.

Design-based causal inference emphasizes inferring a counterfactual quantity, not a quantity in a population or of an outcome-generating model. Such an emphasis arises naturally from a wide range of social science research contexts, such as (1) when units are not a random sample, e.g., when the units are administrative units like schools or countries or convenience samples arising without a known chance process, or (2) when probability models of outcomes – and their structural relationship to explanatory variables – are difficult to write or justify, e.g., when an outcome can be plausibly described by multiple different probability models. In such contexts, simple comparisons based on the research design can advance social science theory and avoid debates about data models. When strong theory generates clear probability and structural models, a model-based justification for statistical inference might be preferred, although we would want this mode of inferences to satisfy the same properties discussed in this chapter: tests should not mislead and estimators



should produce estimates close to the truth. The model-based approach to showing whether these characteristics hold is well described in most statistics textbooks, and we recommend Cox (2006) for an overview.

A general question nevertheless remains: are design-based procedures better or worse than model-based procedures for advancing social science theory or policy learning? Design-based inference is simple, easily interpretable and can ensure that estimators and tests are reliable based on few assumptions, where the assumptions tend to be easily defended in terms of the known features of the research design. But does design-based inference possess reliable properties only for narrowly defined research questions? To be sure, there is nothing intrinsic to design-based inference that requires scholars to use only specific estimators that focus only on specific estimands, such as mean causal effects, or to test only certain hypotheses about no causal effects instead of others. And, although we did not show it here, it is straightforward to assess the properties of non-standard estimators and tests by representing the research design and simulating from it (see Blair et al., 2019, for an example of a framework for simulation based assessment of estimators and tests). For only one example of the flexibility of design-based approaches, imagine that we wondered whether a certain non-linear structural model described well a relationship between a causal driver (like an experimental treatment) and an outcome. In this case, Bowers et al. (2013, 2016) show how the evidence against structural models of unobserved potential outcomes can be generated and hypothesis tests created, i.e., there is nothing about design-based approaches that precludes the use of structural models. That said, a clear difference of means can often teach enough about a complicated structural theory such that there is no need to complicate the research design or statistical inference tasks. In the end, all else equal, reliable procedures advance scientific knowledge

more than unreliable procedures do. For this reason, one of the benefits of engaging with design-based inference is that it brings clarity to the task of judging and choosing our statistical tools and provokes us to directly confront and grapple with the conditions under which evidence can be reliably interpreted as evidence for or against causal claims.

## Notes

- 1 Holland (1986, 947) refers to the inability to observe both potential outcomes for a single unit at the same moment in time as the ‘fundamental problem of causal inference’.
- 2 For simplicity, we consider studies in which there are two conditions – treatment and control – although the same general principles apply to studies with multiple conditions (see Dasgupta et al., 2015).
- 3 To distinguish between fixed quantities and quantities that can take on different values with certain probabilities (i.e., random quantities), we now use uppercase letters for random quantities and lowercase letters for fixed or realized quantities.
- 4 When the numbers of treatment and control units are *not* fixed, such as in simple, individual assignment, the Difference-in-Means estimator remains unbiased in a uniform randomized experiment so long as at least one unit is always in the treatment and control conditions, respectively. In general, when the numbers of treatment and control are *not* fixed, the Difference-in-Means estimator is not necessarily unbiased, such as in cluster uniform random assignment when clusters are of unequal sizes (see Middleton and Aronow, 2015).
- 5 Note that, if treatment assignment probabilities differ *across* blocks (but are uniform *within* blocks), then the standard Difference-in-Means estimator may be biased. In such cases, an unbiased estimator would be the Difference-in-Means estimator that generates an estimate within each block and subsequently weights each block-specific estimate by the proportion of units in that block.
- 6 Recent work has derived a consistent estimator for an upper bound on the term  $2\sigma_{y_c, y_t}$  that is always less than or equal to  $\sigma_{y_c}^2 + \sigma_{y_t}^2$  (Aronow et al., 2014). Such an estimator enables

researchers to more precisely estimate the variance of the Difference-in-Means estimator and, as we will discuss later, increase the power of hypothesis tests about the mean causal effect.

- 7 We use the term 'strong' instead of 'sharp', as used by Fisher (1935), to contrast *strong* null hypotheses with *weak* null hypotheses which we discuss below.
- 8 The expression for a two-sided  $p$ -value,  $p_t = \min\{1, 2\min\{p_u, p_l\}\}$ , comes from Rosenbaum (2010, 33) who states that '[i]n general, if you want a two-sided P-value, compute both one-sided P-values, double the smaller one, and take the minimum of this value and 1'. The rationale is that doubling a one-sided  $p$ -value compensates for, in essence, testing twice.
- 9 We can see this property indirectly from Wu and Ding (2018) and Lin (2013).
- 10 Note that  $R_i$  is independent of  $Z_i$  if and only if the probability that  $R_i$  takes on any value in its sample space does not vary conditional on any value that  $Z_i$  takes on in its sample space – i.e., that  $\Pr(R_i = 1 | Z_i = 1) = \Pr(R_i = 1 | Z_i = 0)$  and  $\Pr(R_i = 0 | Z_i = 1) = \Pr(R_i = 0 | Z_i = 0)$ . The only two types of subjects who satisfy such independence are Always Reporters and Never Reporters.
- 11 Other approaches to estimation and testing relax the assumption that outcome missingness is independent of treatment assignment and devise procedures that bound inferences about treatment effects under best-or worst-case scenarios as they pertain to the true values of missing outcomes, e.g., 'trimming bounds' (Lee, 2009) and 'extreme value bounds' (Manski, 1990).
- 12 We did not assess software packages to implement randomization or sampling nor did we provide much guidance on the choice of which among many possible ways to randomize a treatment or sample from a population. Nevertheless, these topics are also important in the effort to advance theory through observation.

## REFERENCES

- Achen, C. H. (1982). *Interpreting and Using Regression*. Number 07-029 in Quantitative Applications in the Social Sciences. Newbury Park, CA: Sage Publications.
- Angrist, J. D., G. W. Imbens, and D. B. Rubin (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association* 91(434), 444–455.
- Angrist, J. D. and J.-S. Pischke (2008). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton University Press.
- Aronow, P. M., D. P. Green, D. K. Lee, et al. (2014). Sharp bounds on the variance in randomized experiments. *The Annals of Statistics* 42(3), 850–871.
- Aronow, P. M. and J. A. Middleton (2013). A class of unbiased estimators of the average treatment effect in randomized experiments. *Journal of Causal Inference* 1(1), 135–154.
- Aronow, P. M. and M. R. Offer-Westort (2017). Understanding ding's apparent paradox. *Statistical Science* 32(3), 346–348.
- Aronow, P. M. and C. Samii (2017). Estimating average causal effects under general interference, with application to a social network experiment. *Annals of Applied Statistics* 11(4), 1912–1947.
- Bailey, R. A. (2017). Inference from randomized (factorial) experiments. *Statistical Science* 32(3), 352–355.
- Bisbee, J., R. Dehejia, C. Pop-Eleches, and C. Samii (2017). Local instruments, global extrapolation: External validity of the labor supply–fertility local average treatment effect. *Journal of Labor Economics* 35(S1), S99–S146.
- Blair, G., J. Cooper, A. Coppock, and M. Humphreys (2019). Declaring and diagnosing research designs. *The American Political Science Review* 113(3), 838–859.
- Bloom, H. S. (1984). Accounting for no-shows in experimental evaluation designs. *Evaluation Review* 8(2), 225–246.
- Bowers, J., M. Fredrickson, and C. Panagopoulos (2013). Reasoning about interference between units: A general framework. *Political Analysis* 21(1), 97–124.
- Bowers, J., M. M. Fredrickson, and P. M. Aronow (2016). Research note: A more powerful test statistic for reasoning about interference between units. *Political Analysis* 24(3), 395–403.
- Brewer, K. (1979). A class of robust sampling designs for large-scale surveys. *Journal of the*

- American Statistical Association* 74(368), 911–915.
- Caughey, D., A. Dafoe, and L. Miratrix (2018, June). Beyond the sharp null: Randomization inference, bounded null hypotheses, and confidence intervals for maximum effects. Working Paper.
- Chaudoin, S., J. Hays, and R. Hicks (2018). Do we really know the wto cures cancer? *British Journal of Political Science* 48(4), 903–928.
- Chung, E. (2017). Randomization-based tests for 'no treatment effects'. *Statistical Science* 32(3), 349–351.
- Cochran, W. G. (1977). *Sampling Techniques* (Third ed.). Hoboken, NJ: John Wiley & Sons.
- Coppock, A., T. J. Leeper, and K. J. Mullinix (2018). Generalizability of heterogeneous treatment effect estimates across samples. *Proceedings of the National Academy of Sciences of the United States of America* 115(49), 12441–12446.
- Cox, D. R. (1958). *Planning of Experiments*. New York, NY: Wiley.
- Cox, D. R. (2006). *Principles of Statistical Inference*. New York, NY: Cambridge University Press.
- Dasgupta, T., N. S. Pillai, and D. B. Rubin (2015). Causal inference from 2k factorial designs by using potential outcomes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 77(4), 727–753.
- Dehejia, R., C. Pop-Eleches, and C. Samii (2019). From local to global: External validity an a fertility natural experiment. *Journal of Business & Economic Statistics* (just-accepted), 1–48.
- Diamond, A. and J. S. Sekhon (2013). Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *Review of Economics and Statistics* 95(3), 932–945.
- Ding, P. (2017). A paradox from randomization-based causal inference. *Statistical Science* 32(3), 331–345.
- Erdős, P. and A. Rényi (1959). On the central limit theorem for samples from a finite population. *Publications of the Mathematics Institute of the Hungarian Academy of Science* 4, 49–61.
- Fisher, R. A. (1935). *The Design of Experiments*. Edinburgh: Oliver and Boyd.
- Gastwirth, J. L., A. M. Krieger, and P. R. Rosenbaum (2000). Asymptotic separability in sensitivity analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63(3), 545–555.
- Gerber, A. S. and D. P. Green (2012). *Field Experiments: Design, Analysis, and Interpretation*. New York, NY: W.W. Norton.
- Hájek, J. (1960). Limiting distributions in simple random sampling from a finite population. *Publications of the Mathematics Institute of the Hungarian Academy of Science* 5, 361–374.
- Hansen, B. B. (2004). Full matching in an observational study of coaching for the sat. *Journal of the American Statistical Association* 99(467), 609–618.
- Hansen, B. B. and J. Bowers (2008). Covariate balance in simple, stratified and clustered comparative studies. *Statistical Science* 23(2), 219–236.
- Hansen, B. B. and J. Bowers (2009). Attributing effects to a cluster-randomized get-out-the-vote campaign. *Journal of the American Statistical Association* 104(487), 873–885.
- Hartman, E. and F. D. Hidalgo (2018). An equivalence approach to balance and placebo tests. *American Journal of Political Science* 62(4), 1000–1013.
- Höglund, T. (1978). Sampling from a finite population. a remainder term estimate. *Scandinavian Journal of Statistics* 5(1), 69–71.
- Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association* 81(396), 945–960.
- Horvitz, D. G. and D. J. Thompson (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* 47 (260), 663–685.
- Hosman, C. A., B. B. Hansen, and P. W. Holland (2010). The sensitivity of linear regression coefficients' confidence limits to the omission of a confounder. *The Annals of Applied Statistics* 4(2), 849–870.
- Hudgens, M. G. and M. E. Halloran (2008). Toward causal inference with interference.

- Journal of the American Statistical Association* 103(482), 832–842.
- Imbens, G. W. and P. R. Rosenbaum (2005). Robust, accurate confidence intervals with a weak instrument: Quarter of birth and education. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 168(1), 109–126.
- Kang, H., L. Peck, and L. Keele (2018). Inference for instrumental variables: A randomization inference approach. *Journal of the Royal Statistical Society. Series A: Statistics in Society* 181(4), 1231–1254.
- Lee, D. S. (2009). Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *The Review of Economic Studies* 76(3), 1071–1102.
- Lehmann, E. L. and J. P. Romano (2005). *Testing Statistical Hypotheses* (Third ed.). Springer Texts in Statistics. New York, NY: Springer-Verlag.
- Li, X. and P. Ding (2017). General forms of finite population central limit theorems with applications to causal inference. *Journal of the American Statistical Association* 112(520), 1759–1769.
- Lin, W. (2013). Agnostic notes on regression adjustments to experimental data: Reexamining freedman's critique. *The Annals of Applied Statistics* 1(1), 295–318.
- Loh, W. W., T. S. Richardson, and J. M. Robins (2017). An apparent paradox explained. *Statistical Science* 32(3), 356–361.
- Manski, C. F. (1990). Nonparametric bounds on treatment effects. *The American Economic Review* 80(2), 319–323.
- Middleton, J. A. and P. M. Aronow (2015). Unbiased estimation of the average treatment effect in cluster-randomized experiments. *Statistics, Politics and Policy* 6(1–2), 39–75.
- Neyman, J. (1923). Sur les applications de la théorie des probabilités aux expériences agricoles: Essai des principes. *Roczniki Nauk Rolniczych* 10, 1–51.
- Neyman, J. and E. S. Pearson (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London* 231(694–706), 289–337.
- Pearl, J. (2015). Generalizing experimental findings. *Journal of Causal Inference* 3(2), 259–266.
- Robins, J. M., M. A. Hernán, and B. Brumback (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology* 11(5), 550–560.
- Rosenbaum, P. R. (1996). Identification of causal effects using instrumental variables: Comment. *Journal of the American Statistical Association* 91(434), 465–468.
- Rosenbaum, P. R. (1999). Reduced sensitivity to hidden bias at upper quantiles in observational studies with dilated treatment effects. *Biometrics* 55(2), 560–564.
- Rosenbaum, P. R. (2002). *Observational Studies* (Second ed.). New York, NY: Springer.
- Rosenbaum, P. R. (2007). Interference between units in randomized experiments. *Journal of the American Statistical Association* 102(477), 191–200.
- Rosenbaum, P. R. (2010). *Design of Observational Studies*. New York, NY: Springer.
- Rosenbaum, P. R. (2018). Sensitivity analysis for stratified comparisons in an observational study of the effect of smoking on homocysteine levels. *Annals of Applied Statistics* 12(4), 2312–2334.
- Rosenbaum, P. R. and A. M. Krieger (1990). Sensitivity of two-sample permutation inferences in observational studies. *Journal of the American Statistical Association* 85(410), 493–498.
- Rubin, D. B. (1980). Comment on 'randomization analysis of experimental data in the fisher randomization test' by basu, d. *Journal of the American Statistical Association* 75(371), 591–593.
- Rubin, D. B. (1986). Which ifs have causal answers? (comment on 'statistics and causal inference' by Paul W. Holland). *Journal of the American Statistical Association* 81, 961–962.
- Sävje, F., M. J. Higgins, and J. S. Sekhon (2017). Generalized full matching. Working Paper.
- Senn, S. (2004). Controversies concerning randomization and additivity in clinical trials. *Statistics in Medicine* 23(24), 3729–3753.
- Stuart, E. A., C. P. Bradshaw, and P. J. Leaf (2015). Assessing the generalizability of randomized trial results to target populations. *Prevention Science* 16(3), 475–485.

Wald, A. (1940). The fitting of straight lines if both variables are subject to error. *The Annals of Mathematical Statistics* 11(3), 284–300.

Wu, J. and P. Ding (2018, October). Randomization tests for weak null hypotheses. In *eprint arXiv:1809.07419v2*.

Zubizarreta, J. R. (2012). Using mixed integer programming for matching in an observational study of kidney failure after surgery. *Journal of the American Statistical Association* 107(500), 1360–1371.

# Statistical Matching with Time-Series Cross-Sectional Data: Magic, Malfeasance, or Something in between?

Richard A. Nielsen<sup>1</sup>

In the beginning, there was history. If scholars of International Relations (henceforth IR) wanted to know why something happened, they consulted history for patterns. Under what conditions did that something happen? What if conditions had been different? What if choices had been different? What if timing had been different?

Then, spurred by a combination of factors, some IR scholars discovered that the field of statistics offered new ways to explore history. By compiling history into databases, wrangling it into numerical matrices, and applying mathematical models, these scholars abstracted away from the fine details to see broad patterns that are difficult for even the most systematic historian to pick out. Historical and statistical modes of inquiry seem entirely different: the data rearranged in unfamiliar ways, the skill sets distinct. Yet, the logic of inference can be surprisingly similar if the language barrier between traditions can be crossed.

Matching is a family of quantitative procedures for creating and analyzing a sample

of cases that differ on one key factor (called a ‘treatment’) and are ‘matched’ to be similar on others. It puts historical counterfactual comparison at the center of statistical analysis, providing one ‘border crossing’ between the historical and statistical traditions in IR.

Matching emerges naturally as the statistical analog to the qualitative tradition of paired case comparison (Tarrow, 2010). In a qualitative study of a puzzling phenomenon, our first move might be to identify a positive case: one in which the outcome of interest happens. We might then try to infer *why* the outcome happened from the historical record by considering possible counterfactual histories: if various factors had been different, would the outcome have been different also?

Relying on subjective intuition about counterfactual histories for a single case can lead us to merely confirm our prior assumptions. Comparison cases provide more objective information for inferring how the positive case might have turned out if some factor had been different. Following the logic of Mill’s

methods of difference, we might seek a comparison case that is similar to our positive case but differs in the factor that we suspect matters most. If our cases are otherwise identical and our theories are deterministic, then this comparison can tell us all we need to know. Any difference between the case outcomes is the causal effect of the differing factor. However, our theories are not usually deterministic and our paired cases rarely match in every respect, so the difference in outcomes could be due to something else. To be more certain that the factor explains the outcome, we might do another comparison, and another, to see if they support the same conclusion. Soon, we have too many pairs of cases to easily summarize qualitatively, so we might give a numerical summary: perhaps the average difference in outcomes across pairs of cases. We have arrived at matching. Typically, an analyst creates a matched sample by trimming down a quantitative data set, but a researcher employing the method of paired comparison could build the same matched sample ‘from the ground up’, as I have just described. The power of matching comes from the properties of the sample, not how the analyst obtains it.

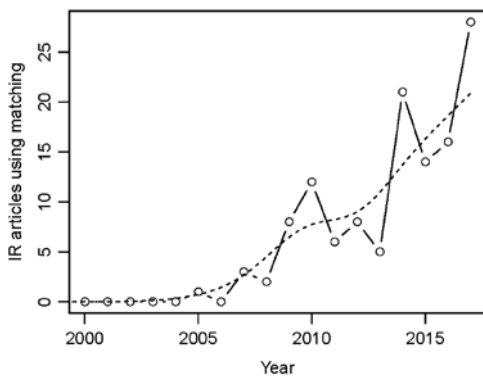
After constructing a matched sample, an analyst typically estimates the causal effect of treatment on the outcome using a regression model. Regression is a method for calculating

correlations that does not, by itself, necessarily produce credible estimates of causal effects. Matching is a method for selecting cases to use in the regression calculation so that the assumptions necessary to infer causation from correlation are more plausible. Matching is case selection for quantitative studies.<sup>2</sup>

In the mid 2000s, matching burst into IR and Political Science in a series of high-profile publications. To chronicle its dramatic rise, I collected a comprehensive list of articles applying matching to IR data in the 12 leading IR journals.<sup>3</sup> Following the first application by Simmons and Hopkins (2005), the number of applications has risen exponentially. In 2017, a record high of 28 articles used matching (see Figure 42.1), and by mid 2018 there were 20, with the year only half-way over (the latest for which I could collect data). Matching now features in at least 10% of the approximately 200 quantitative IR articles published each year.

The rise of matching has been controversial. Proponents tout the benefits for making causal inferences with non-experimental data. Detractors argue that matching is over-hyped as a magic bullet for causal inference, and that it offers unscrupulous researchers another tool to ‘p-hack’ their way to false discoveries (Arceneaux et al., 2006, Arceneaux et al., 2010, Miller, 2019). The time-series cross-sectional (TSCS) data sets common in IR are especially challenging for matching. With little guidance from methodologists, researchers have developed ad hoc approaches to matching with TSCS data, a fact that is equally troubling for the credibility of research findings despite receiving less attention.

The critics are right that matching is not magic. It is helpful for causal inference, but only with strong assumptions. Existing matching methods do not adequately accommodate the complex structure of many IR data sets, and matching does not stop researchers from p-hacking. Like any other method, it can be misused, oversold, and misunderstood. But matching is not just hype either. Matching has proven to be a reliably useful



**Figure 42.1** The number of articles per year using statistical matching in the 12 leading IR journals

method for causal inference with messy, non-experimental data. It belongs in the toolbox of IR scholars.

In this chapter, I introduce matching for readers with no prior experience and weigh its merits and weaknesses in IR applications using time-series cross-sectional data. There are already several excellent introductions to matching (Ho et al., 2007, Sekhon, 2009, Stuart, 2010), which I aim to complement rather than reproduce. I first introduce the philosophy of causal inference underlying matching. I then explain the mechanics of matching that arise from this philosophy. Matching mechanics are not always well-suited to the time-series cross-sectional data structures common in IR, so I discuss the challenges and attempts to surmount them. I conclude by returning briefly to the controversies.

## THE PHILOSOPHY OF MATCHING

Matching rests on a counterfactual approach to causation. A variable  $T$  (for ‘treatment’) is a cause of variable  $Y$  if  $Y$  would have been different had  $T$  been different. Imagine all possible values of  $Y$  for a unit  $i$  and call these the potential outcomes of unit  $i$ . For simplicity, consider the case where  $T$  is a binary indicator for whether the event represented by  $T$  happened ( $T = 1$ ) or not ( $T = 0$ ). The causal effect of  $T$  on  $Y$  for unit  $i$  is the potential outcome when  $T$  is one minus the potential outcome when  $T$  is zero, denoted  $Y_i(T = 1) - Y_i(T = 0)$ .

The fundamental problem of causal inference is that for any unit, we cannot observe both potential outcomes because we cannot re-run history (Holland, 1986). For any causal inference, at least half of the data for our calculation are missing. Thus, we *infer* causal effects rather than *measuring* them; we are always guessing about the counterfactual outcome that we would see if  $T$  were different.

Experimental manipulation is the best way to learn about cause and effect (Bowers and Leavitt, Chapter 41, this *Handbook*). In an

experiment, we randomly assign  $T$  to learn whether it causes changes in  $Y$ . Yet even in an experiment, we cannot observe both potential outcomes for any individual. To impute the missing potential outcomes, most analysts estimate average causal effects over many individuals. With enough units and random assignment of  $T$ , we can average  $Y$  for the  $T = 1$  group and subtract the average of  $Y$  for the  $T = 0$  group to get an unbiased estimate of the average treatment effect (abbreviated ATE). We can write this difference in means estimator as  $\bar{Y}_{\text{treated}} - \bar{Y}_{\text{control}}$ .

Despite efforts to bring experiments into IR (Findley et al., 2013), there are many situations where randomization is impossible, unethical, or both. We are forced to infer cause and effect from an *observational study* using data observed in the natural world. In observational studies, units select their level of treatment or have it assigned to them non-randomly (called ‘selection effects’, ‘endogeneity’, and ‘reverse causality’). This results in *imbalance*: systematic differences between treated and control groups. If selection into treatment is correlated with potential outcomes of  $Y$ , then the average control unit provides a biased estimate of the missing potential outcome for the average treated unit, making the estimator  $\bar{Y}_{\text{treated}} - \bar{Y}_{\text{control}}$  biased as well. These factors that are related to both treatment status and the potential outcomes are called *confounders*, and often denoted with the matrix  $X$ .

Some observational studies feature a natural experiment in which treatment is assigned ‘as if’ randomly by nature for a subset of the data. Even if treatment assignment in the overall data set is correlated with possible confounding variables, comparing the subset of units that received as if random treatment assignment can produce credible causal inferences. Techniques such as instrumental variables (Carter and Dunning, Chapter 40, this *Handbook*) and regression discontinuity designs (Cattaneo et al., Chapter 44, this *Handbook*) facilitate these comparisons, depending on the type of natural experiment that occurred.



Often, there is no natural experiment for a researcher to exploit. The next-best strategy is to compare outcomes for units with similar values of the confounding variables. This is called a *conditioning strategy* because we condition on the values of  $X$  to infer the causal effect of  $T$  on  $Y$ . Matching is a conditioning strategy proposed by Rubin and collaborators in a series of papers starting with Rubin (1973).<sup>4</sup> As developed by Rubin, matching identifies treated and control units that have similar values of the confounding variables  $X$  in a data set and then removes the other units, hoping to approximate the data that would have resulted from a randomized experiment. In this matched sample, the variables in  $X$  are no longer correlated with treatment assignment so we can obtain unbiased estimates of the ATE using the difference in means:  $\bar{Y}_{\text{treated}} - \bar{Y}_{\text{control}}$ .

The core insight of matching is that a subset of data may be useful for credible causal inference, even if all available data are not. However, subsetting the data may limit the inferences a researcher can make. In a natural experiment, only some units have treatment randomized by nature, so the estimated quantity of interest is a *local average treatment effect* for only those units (Imbens, 2009). Analogously, matching only allows credible causal inference for the types of units retained in the matched sample; all other inferences rely on extrapolation. If an analyst's quantity of interest is the *sample average treatment effect* on the treated units (abbreviated SATT), then they cannot discard any treated units. This quantity of interest answers the question 'what would have happened if all treated units in this sample had instead received control?' Answering an alternative question such as 'what if all control units received treatment?' requires a different matching scheme that retains all control units but may discard treated units. Sometimes, an analyst may wish to estimate the SATT, but there are treated units that do not have good matches and discarding them would greatly reduce imbalance. Analysts may discard these treated observations, but their quantity of

interest is now the *feasible sample average treatment effect on the treated* (FSATT): the ATE in the subsample for which causal inference is feasible (King et al., 2017: 475).

## Assumptions

The assumptions necessary for causal inference with a conditioning strategy are stringent. First, we require the *stable unit treatment value assumption*, abbreviated SUTVA, which states that the fixed (but generally unknown) potential outcomes for each unit do not depend on the potential outcomes of the other units. Equivalently, this assumption states that there is no interference between units and no hidden versions of treatment.

Second, conditioning strategies require the assumption of *conditional ignorability*: that after conditioning on the  $X$ , treatment assignment is independent of the potential outcomes. Alternatively, this assumption is called 'selection on observables' or 'no omitted variables'. If unobserved confounders affect both treatment and outcome, conditioning on  $X$  cannot provide unbiased causal estimates. This requires that the analyst know which variables are confounders. Variables that are part of the *assignment mechanism* of  $T$  should be included in  $X$ . The most compelling matching applications carefully theorize the assignment mechanism of  $T$  and then exhaustively measure each variable.

Third, we require *common support*: there must be treated and control units at all levels of  $X$  for which we wish to make inference.

## CONSTRUCTING AND ANALYZING MATCHED SAMPLES

The core activity of matching is constructing a matched sample that plausibly satisfies the assumptions of stable unit treatment values, conditional ignorability, and common support. A *matching algorithm* is the procedure

for constructing a matched sample. Matching algorithms typically identify similar treatment and control units and match them to each other, discarding (or ‘pruning’) unmatched units from the data set. This is equivalent to giving discarded units a weight of zero in subsequent statistical calculations, so a matched sample is merely a reweighted version of an unmatched sample.<sup>5</sup> After constructing a matched sample, the analyst can calculate the effect of treatment by calculating the difference of the (weighted) average outcome of the treated units and the weighted average outcome of the control units ( $\bar{Y}_{\text{treated}} - \bar{Y}_{\text{control}}$ ). Or, the analyst may wish to further adjust via regression (Ho et al., 2007).

Although matching algorithms are the focus of much of the matching literature, from a certain perspective, the algorithm is not very important. What matters is getting the best possible matched sample by whatever means. What is the best possible matched sample? Matched samples are evaluated on two main criteria: the similarity of treated and control units in the sample (called *balance*) and sample size. A larger sample provides more statistical power for estimating precise results. A balanced sample is more likely to satisfy the assumptions of common support and conditional ignorability. These two criteria are in tension because the way to improve balance is to discard observations. Matching is a constrained optimization problem of maximizing balance subject to a sample size constraint (or maximizing sample size subject to a balance constraint).

Assessing sample size is straightforward, but the notion of balance deserves elaboration. When  $X_{\text{treated}} = X_{\text{control}}$ , the data set is unambiguously balanced. However, when  $X_{\text{treated}} \neq X_{\text{control}}$ , there is no universal definition of balance. Various balance metrics have been proposed and each may give different answers. Improvement on one balance metric may not correspond to improvements in other balance metrics. Applied scholars often fixate on justifying their choice of matching algorithm, but they should be more concerned

with justifying their choice of balance metric. As I show below, these are integrally connected, but the prevalence of nonsensical practices in the literature reveals that analysts may not fully appreciate this connection.

Each balance metric offers different answers to two questions: How should differences between  $X_{\text{treated}}$  and  $X_{\text{control}}$  be summarized? And which differences are small enough for a data set to be declared sufficiently balanced? Analysts must inescapably answer these questions by choosing a balance metric, and their choice reflects an implicit set of assumptions about how confounding threatens their inferences.

Analysts who believe that confounding is omnipresent worry that minor differences between  $X_{\text{treated}}$  and  $X_{\text{control}}$  could induce substantial confounding. These analysts prefer balance metrics that are sensitive to small differences in many possible dimensions. Other analysts might be less concerned about confounding from small differences between  $X_{\text{treated}}$  and  $X_{\text{control}}$  and prefer balance metrics that do not capture these small differences.

One set of balance metrics focus on the discrepancies between pairs (or groups) of treated and control units. The *Average Mahalanobis Imbalance* metric uses the average of the pairwise Mahalanobis distance<sup>6</sup> from each unit to the nearest unit of the opposite treatment status (King et al., 2017: 479). The *L1 Imbalance* metric is a discretized version of the same idea. The analyst selects a multivariate histogram binning, calculates the difference in the frequency of treated and control units in each bin, and averages over the bins of the histogram (King et al., 2017: 479). For both metrics, larger values indicate more imbalance. These discrepancy-based balance metrics are sensitive to small differences in many possible dimensions. Both of them have the highly desirable ability to detect an exactly matched subset of observations if such a subset exists. However, they offer no definitive threshold for declaring a data set ‘balanced’ short of zero difference ( $X_{\text{treated}} = X_{\text{control}}$ ).

An alternative style of balance metrics uses statistical tests to detect differences between  $X_{\text{treated}}$  and  $X_{\text{control}}$ . Some analysts calculate the difference in means of the covariates in  $X$  and declare the data set balanced if variable-by-variable t-tests fail to reject the null hypotheses that the means are the same. Some analysts prefer Kolmogorov–Smirnov tests for the difference in distributions to t-tests. Hansen and Bowers (2008) propose a global p-value for a test that the combined dimensions of  $X$  are detectably different from each other. The metrics rely on widely recognized significance thresholds for declaring a data set sufficiently balanced, even if  $X_{\text{treated}} \neq X_{\text{control}}$ . Unfortunately, hypothesis testing depends on the sample size, so ‘balance’ may be achieved once enough observations are removed whether the remaining observations in the matched sample are actually similar or not (Stuart, 2010). Also, optimizing these metrics will not reliably uncover an exactly matched subsample, even if one exists.

In addition to these formal approaches, analysts also inspect balance by plotting the difference in means between treated and control groups for each covariate before and after matching (akin to the variable-by-variable t-tests, but without a test).

### **Matching Algorithms**

It does not matter how one constructs a matched sample with sufficient balance and sample size. In theory, analysts could directly optimize their preferred balance metric subject to a sample size constraint. In practice, this is usually not possible because the optimization problem is intractable, though it can be done for some balance metrics (King et al., 2017). One could attempt brute-force optimization by randomly sampling all possible subsets of the data and finding the sufficiently sized subset with the best balance. However, this is infeasible for most data sets because the number of possible combinations is too large. Instead, analysts turn to

matching algorithms that dramatically reduce the time necessary to construct candidate matched samples and select the best one.

Ideally, matches should be exact: each treated unit paired to a control unit (or units) with identical values of all confounding pre-treatment covariates. Exact matches can be identified by checking whether each treated unit has a matching control unit with identical values of  $X$  (using, for example, the MatchIt library in R).

In most data sets, exact matching leaves too few observations for precise estimation in the subsequent statistical analysis. With continuous covariates, exact matches are unlikely to exist at all. Approximate matching methods offer a way to identify matched subsets of the data that plausibly satisfy the three key assumptions above, even if the treated and control units are not exactly matched.

Each approximate matching algorithm deals with the problem of inexact matches by explicitly or implicitly constructing a distance metric from each treated unit to each control unit and then attempting to identify a matched sample of sufficient size that minimizes that distance metric.

There is no established ‘best’ matching algorithm. Because each algorithm has a different distance metric, each is optimizing a different implied balance metric, and each potentially finds different matched samples. Standard practice is to match with one algorithm and then check balance with one or more metrics *not* implied by that algorithm. If one is lucky, optimizing the balance metric implied by the algorithm will lead to improvements with respect to other balance metrics. In this happy circumstance, the choice of balance metric does not matter much because they all move together.

However, in many cases, a matching algorithm does not lead to notable improvements on all balance metrics. Standard practice in this situation is to change the tuning parameters of the algorithm, rematch, and check balance again, repeating until balance on various balance metrics is satisfactory. More generally,

analysts often match using a method that minimizes one metric, and then evaluate balance by checking another metric. This is a strange form of indirect optimization. If balance on one metric is really the target, then iteratively optimizing it using hit-or-miss trials from an unrelated matching algorithm seems nonsensical. One justification might be that some balance metrics cannot be directly optimized, but even so, the standard practice of checking balance and manually adjusting the algorithm after each iteration of is inefficient.

Instead, analysts who cannot find exactly matched observations within their data sets must make a philosophical commitment to a particular balance metric and, if possible, chose a matching algorithm that optimizes it. Checking whether matching improves alternative balance metrics may be revealing about the sensitivity of matches to the choice of balance metric, but it is not dispositive about whether the data set is balanced. When two balance metrics diverge meaningfully, there is no way to optimize both at the same time. The choice of one over the other is philosophical.

There is a long and growing list of matching techniques.<sup>7</sup> I'll review the key families of algorithms which are the foundation for many, though not all, of the options available.

### ***Mahalanobis Distance Matching (MDM)***

The first matching approach proposed by Rubin (1980) matches treated and control units based on continuous distances from each other in the  $k$ -dimensional space defined by the  $k$  covariates in  $X$ . The basic matching machinery works for any continuous distance metric: First, calculate the distance from each treated unit to each control unit. Then, form matches by selecting the closest control observation to each treated observation. Finally, discard the unmatched control observations. Methodologists prefer Mahalanobis distance (Euclidian distance normalized by the covariance matrix) because it accounts for

correlations between the variables and makes them unit-less and scale-invariant. In order to avoid matched pairs that are not adequately similar, analysts often use a *caliper*: a defined maximum distance for a match. If a treated unit does not have a matching control unit within the specified caliper, it is discarded and the quantity of interest shifts from the SATT to the FSATT.

### ***Propensity Score Matching (PSM)***

Rubin and Rosenbaum (1983) introduce the *propensity score*, defined as the probability that unit  $i$  receives treatment, conditional on  $X$ . They show that the propensity score is a sufficient statistic for  $X$ ; all of the information about  $X$  necessary to make the potential outcomes independent of treatment is contained in this score. Assuming conditional ignorability, matching units exactly on their true propensity scores results in unbiased estimates, even if  $X_{\text{treated}} \neq X_{\text{control}}$ . The true propensity score is usually unknown, so we must estimate it – typically with predicted values from a logistic regression predicting treatment as a function of  $X$ . We then match treated and control units using the similarity of their propensity scores. Again, analysts often define calipers. A common rule of thumb is that observations should be within 0.25 of a standard deviation of the propensity scores to match (Rosenbaum and Rubin, 1985: 114; but see King and Nielsen, 2019).

The theory underlying PSM is elegant and intuitive, but it requires exact matches on the true propensity score. In practice, we typically must make do with inexact matches on estimated propensity scores, and this can sometimes make causal inferences worse, rather than better (King and Nielsen, 2019). Several methods combine information from the propensity score and some other covariate information (Diamond and Sekhon, 2013; Imai and Ratkovic, 2014), and these methods side-step the problem.

There are a host of complications that can be vexing for both MDM and PSM.

Methodologists assume a large pool of control observations from which to match (Rubin and Thomas, 1996), but this is not the case in many IR settings. Also, the same control might be the best available match for multiple treated units. Which should get it? Optimal matching (Rosenbaum, 1989), optimal full matching (Hansen, 2004), and optimal cardinality matching (Zubizarreta et al., 2014) are promising solutions that improve global balance as well as local balance.

### ***Coarsened Exact Matching (CEM)***

Coarsened exact matching is an extension of exact matching methods to situations where exact matches are not possible (Iacus et al., 2012) and is the most popular example of a class of matching methods that are monotonically imbalance bounding (Iacus et al., 2011). Monotonically imbalance bounding methods set a fixed bound on the imbalance that the analyst will permit and achieve that balance, though often at the expense of matched sample size.

CEM ‘coarsens’ each variable into categories that are less restrictive than the measured values. For example, if the democracy level of country is measured on a 21 point scale, as in the Polity IV scores (Marshall et al., 2002), we might coarsen this variable into five categories based on the scores: strongly democratic, leaning democratic, inchoate, leaning autocratic, and strongly autocratic. Then, we match units that fall in the same category, rather than requiring that matches have exactly the same score. Observations that do not have matches, both treated and control, are discarded, and the remaining observations are reweighted to account for the fact that matches may not be one-to-one. Then, the analyst ‘uncoarsens’ the data and estimates causal effects using weighted least squares on the matched data set.

### ***The Matching Frontier***

A new class of approaches seeks to improve inferences with each of the prior methods by

introducing the ‘matching frontier’ – a set of matched samples that are optimal according to the balance metric they maximize (King et al., 2017). Users of MDM and PSM often use *calipers*, which define the maximum allowable distance between a treated and control match. A stricter caliper will result in a smaller, but more balanced, matched sample. Analysts also tend to create one-to-one matched samples out of convenience. However, if more than one control unit is a good match for a treated unit, arbitrarily limiting the number of matches decreases the sample size and decreases the precision of causal estimates without removing bias.

The matching frontier generalizes the idea of a caliper and allows many-to-many matching. The frontier is defined by matched samples of various sizes, from  $n = N$  (the full unmatched sample) down to  $n = 2$  (the closest matching pair). At each possible sample size  $n$  between

$N$  and 2, there are  $\binom{N}{n}$  possible matched

samples. The frontier finds the optimal one for each  $n$ , resulting in a series of optimal matched samples of various sizes.

This continuum of matched samples defines the trade-off between bias and variance. Matched samples that discard all but a few extremely close matches will reduce bias the most but may lack the statistical power to precisely estimate effects. On the other end, those that retain most of the original sample will usually result in more precise estimates, but there may be more bias because the observations are less similar to each other. Any matching solution that is not on this frontier is suboptimal: there is an achievable matched sample that has either more observations or better balance.

Calculating matching frontiers for arbitrary balance metrics is hard because the number of possible matched samples for real-life data sets is extremely large. King et al. (2017) develop fast algorithms that calculate the matching frontier for MDM, PSM, and CEM-style methods.

### ***Categorical and Continuous Treatments***

All of these algorithms struggle to accommodate categorical and continuous treatment variables. The problem is philosophical as well as technical. As the number of treatment levels increases, so does the number of counterfactual questions we must answer to describe what would have happened to a given unit if it received each alternative treatment level. One possibility is to construct a matched sample for each possible treatment contrast, but this is difficult if each level of treatment is only assigned to a few observations. With a truly continuous treatment variable, each unit receives a unique level of treatment, so constructing a matched sample for each treatment level is impossible. Instead, methodologists have proposed matching methods that effectively consider similar levels of treatment to be equivalent (Imai and Van Dyk, 2004; Hirano and Imbens, 2004).

Regression appears to side-step this challenge, effortlessly estimating effects of categorical and continuous treatments. In fact, regression faces the same challenge but limits the analyst's ability to diagnose it; regression automatically fits a hyperplane that extrapolates between treatment levels while obscuring the fact that this extrapolation may be based on very little data at any single treatment level. Matching calls attention to the true degree of difficulty in making credible causal inferences when there are many versions of treatment.

### ***Inference***

After matching, analysts typically calculate their desired treatment effect by fitting a parametric regression model predicting  $Y$  as a function of  $T$ , possibly controlling for variables in  $X$  if matching was inexact. If the assumptions hold, this results in an unbiased estimate. But analysts are almost always interested in testing whether this estimate is statistically different from some null hypothesis

(typically that treatment has no effect). This requires estimates of the uncertainty surrounding an effect estimated via matching and there is substantial debate in the literature about how to do so.

Because it is convenient, the standard practice among practitioners is to report the standard errors from regression analysis as measures of uncertainty for the treatment effect. Methodologists worry that these standard errors do not account for the uncertainty of the matching procedure. Abadie and Imbens (2008) show that naive bootstrap standard errors are not asymptotically valid and propose a correction, though recent work proposes an asymptotically valid bootstrap (Otsu and Rai, 2017). Others propose algorithm-specific approaches such as a Bayesian estimate incorporating the uncertainty inherent in estimating propensity scores (An, 2010). However, Iacus et al. (2019) argue that these concerns are misplaced; if analysts are willing to change their axioms about sampling, then unaltered regression standard errors are correct.

One especially clever approach to justifying measures of effect size uncertainty comes from Imai and Kim (2019). They demonstrate an equivalence between linear models with unit fixed effects and a class of within-unit matching schemes. From this equivalence, they are able to show that accepted model-based standard errors are valid, without resorting to the more computationally intensive proposed alternatives. While this equivalence approach currently only applies to an uncommon subclass of matching approaches, it might be extensible.

### ***Comparison to Regression***

Regression is an alternative approach to conditioning on  $X$  that has a much longer history of use in IR. Regression conditions on  $X$  by calculating smooth hyperplanes that summarize the central tendency of the data at all levels of the variables in the regression.

To make causal inference, a researcher simply compares the average distance between the conditional central tendency for the treated units and the conditional central tendency of the control units. Regression without matching works well for causal inference if the assumptions listed above hold, and if the regression hyperplanes accurately reflect the central tendency of  $X$  – in other words, if the model fits well. However, matching offers several advantages over traditional regression.

Benefit 1: matching automatically conditions on complex interactions between covariates. Linear regression can do this in principle, but in practice, most applied researchers specify linear, non-interactive regression models and only check model fit haphazardly. In a balanced matched sample, treatment effect estimates are not dependent on whether the analyst includes interactions and nonlinear terms in a subsequent regression (Ho et al., 2007).

Benefit 2: matching identifies and corrects issues of ‘common support’, where some subset of treated or control units is so dissimilar from any units with other treatments that any inference about the outcomes of these units is determined almost entirely by modeling assumptions rather than data. Matching also draws attention to the related challenge of estimating the effects of categorical and continuous treatments. Linear regression obscures these challenges.

Benefit 3: analysts have intuitions about cases, but regression makes it difficult for analysts to tell which cases are really being compared. Analysts can examine matched cases to directly assess the quality of their counterfactual comparisons.

## **MATCHING WITH TIME-SERIES CROSS-SECTIONAL DATA**

IR researchers frequently use time-series cross-sectional (TSCS) data to estimate causal effects. With the exception of a few

working papers (Nielsen and Sheffield, 2009, Imai et al., 2020), there has been very little attention to TSCS data in the matching literature, leaving applied researchers with few guidelines. My survey of the applied matching literature in IR shows that 65% of IR articles that use matching use it on TSCS data, but only half of these do anything to account for the structure of the data, and few analysts defend their choices.

In this section, I explain the challenges of matching with TSCS data and describe current best practices. This is an area of active research (Imai et al., 2020), so future scholarship may offer amendments and improvements to the approach I endorse here.

To match TSCS data, IR scholars must be attentive to their unit of analysis. Outside of IR, matching is typically applied to cross-sectional data, in which each row in a data matrix records information about a single unit (e.g., a patient in a medical study). Analysts generally assume that these units are independent, so each row of the matrix is exchangeable. If this assumption holds, then every treated unit may be matched to any control unit without fear of violating the stable unit treatment value assumption or the assumption of conditional ignorability. Virtually all matching algorithms have been designed to take the data matrix provided by the analyst and match each treated row to the nearest available control row without constraint. In this setting, each row of the data set matches the analyst’s conceptual unit of analysis.

TSCS data are typically formatted in a ‘long’ format matrix where each column is a variable and each row contains information about a given unit in a given time period, such as a country-year. These rows are no longer plausibly independent observations. Repeated observations of the same unit are probably correlated, as are observations of different units in the same time period. Applying standard matching algorithms to these data sets often results in matches that are likely to be dependent (e.g., if the country of Ghana received a treatment in the year

2005, the row of the data matrix with the most similar  $X$  values is likely to be Ghana in 2004). This is not necessarily bad. In fact, Imai and Kim (2019) show that estimates from a within-unit matched sample are equivalent to a popular fixed effects model. Generally, however, if countries (rather than country-years) are the analyst's unit of analysis, then I advise modifying the TSCS data matrix to correspond. For most TSCS matching applications, the fix is simple: transform the data matrix from 'long' to 'wide' format and then perform matching.

The cross-sectional data sets used in traditional matching applications can be thought of as TSCS data sets in 'wide' format with only two time-periods: pre-treatment (covariates) and post-treatment (outcome). Extending matching methods to data sets with a longer pre-treatment time-series component merely requires appending this information in the right way.<sup>8</sup> Each prior time period simply adds to the available set of covariates on which to match treated and control units. Rather than representing this additional time-period with a new row in the data matrix  $X$ , represent it with a new column. Similarly, lagged values of the outcome variable should be represented as additional columns on the data matrix.

Analysts may be interested in long-range outcomes. If so, long-range measurements of the outcome variable should be included in a single row as well. However, analysts should not adjust for post-treatment covariates that might affect the outcome, because they might also be a result of treatment, and would then bias the estimated treatment effect. The difficulty of adjusting for post-treatment confounding means that inferences about immediate outcomes will be much more precise than inferences about long-range outcomes.

Several complications remain. First, should TSCS matching include every observed lag of every covariate? Can units provide multiple observations? And if so, how should treatment and control units be defined? Finally, should

analysts remain concerned about interference between units inducing SUTVA violations? I discuss each of these issues in turn.

### ***Should Matching Include Every Lag of Every Covariate?***

With a TSCS set in 'wide' format, analysts will suddenly feel that they have 'lost' most of their data because they have exchanged a data matrix with many rows and few columns for one with few rows and many columns. This 'loss' is illusory; the ' $N$ ' of the data matrix in 'long' format gives a highly inflated sense of the effective number of observations. The 'wide' data matrix encodes the same information but reflects dependencies between repeated observations of the same unit.

With TSCS data in 'wide' format, the analyst may have more variables than observations. Suppose they observe 180 countries over, say, 74 years since the end of World War II and wish to 'control for' 20 variables (e.g., GDP, Democracy, Trade). Structured as country-years, the  $X$  matrix would have 13,320 observations and 20 variables, but when reformatted to be 'wide', the same data set now has 180 observations and 1,480 variables (e.g.,  $GDP_{1946}, \dots, GDP_{2019}; Democracy_{1946}, \dots, Democracy_{2019}; Trade_{1946}, \dots, Trade_{2019}$ ). Most matching methods fail when the number of covariates exceeds the number of observations (Roberts et al., forthcoming).

If a country receives treatment in 2015, does the analyst really need to consider each observation of each control variable back to the end of World War II? Probably not. Determining how many lags of a given covariate to use depends on how treatment is assigned. Analysts should choose the smallest set of lags that ensures conditional ignorability. Ideally, theory and prior evidence should guide this choice, but they are rarely precise enough to dictate whether  $democracy_{t-4}$  is a confounder after including  $democracy_{t-1}$ ,  $democracy_{t-2}$ , and  $democracy_{t-3}$ . Without strong theory to guide the choice about lags,



analysts can reasonably turn to heuristics. For example, if the analyst assumes that treatment assignment is largely a function of recent events, she might match placing greater weight on the recent past while still placing a small weight on the distant past (see Nielsen, 2016: 588 for an example).

One practical reason to omit unnecessary lags is that some are likely to be missing. Analysts can impute these missing values, but it complicates matching (D'Agostino Jr and Rubin, 2000). If treatment assignment is not a function of these missing lags, then avoid the complication.

### ***Can Units Provide Multiple Observations?***

Considering each row of a 'long' TSCS data set as an independent observation dramatically overstates the amount of information the data contain. But transforming it to 'wide' format implicitly assumes that repeated observations are completely dependent, which is probably too conservative. There is often independent information in repeated observations of each unit that could be exploited.

This motivates some IR scholars to use a different unit of analysis – the country-block (see, for example, Simmons and Hopkins, 2005; Hollyer and Rosendorff, 2012; Nielsen and Simmons, 2015). The analyst selects a number of  $l$  lags which they consider to be sufficient for conditional ignorability on all variables, and a number of  $m$  post-treatment periods for measuring the outcome. Each treatment block is defined by the  $l$  time-period observations before treatment and the  $m$  time periods after. Control units are also divided into blocks of  $l+m$  consecutive observation periods. If the analyst wishes to match exactly on time, then time subscripts of treated and control blocks must match. If exactly matching time is unnecessary, then treated blocks can match control blocks with different time subscripts, greatly increasing the number of available controls. Each unit

that never receives treatment can potentially offer multiple control blocks, rather than one. Additional control blocks can come from the pre-treatment life histories of units that eventually get treatment. In some IR data sets, almost every unit eventually receives treatment, so drawing from the pre-treatment histories is the only source of control blocks.

Approximately 15% of the TSCS matching papers in IR have landed on this strategy, including the pioneering paper of Simmons and Hopkins (2005). Simmons and Hopkins use the following procedure: first, for each treated unit, they drop all lags except for the four years prior to treatment, the year of treatment, and the year after. They divide the control units into blocks of six consecutive observations, using both countries that never receive treatment and countries that receive treatment later. They then average the lagged covariates over the first four years of both treated and control blocks, and do a propensity score matching with these averages, reporting improvements in balance and, presumably, a reduction in bias.

### ***The Potential for Violations of SUTVA***

Many applications in international relations and comparative politics are likely to violate the stable unit treatment value assumption (SUTVA). This is not a problem induced by matching; matching merely illuminates a problem that standard TSCS regression techniques often overlook.

One obvious violation of SUTVA arises from the dependence between repeated observations of the same unit that I have just discussed; there may be 'interference' between the observations. Causal inference with regression requires the same SUTVA assumptions as causal inference with matching, so it is puzzling that IR scholars sometimes wring their hands about matching repeated observations of the same unit to each other but then blithely throw them all

into a pooled regression. If pooled matching makes the analyst uncomfortable, then pooled regression is inappropriate. Matching TSCS data in the ‘wide’ format avoids the problem of dependence within repeated observation of a single unit, but interactions between units also threaten to violate SUTVA. Unlike a medical study in which subjects can be isolated, it is not generally reasonable to assume that there is no interference between units in the international system. There is currently no widely accepted solution to this problem, though causal inference under networked interference is an active area of research (Aronow and Samii, 2017).

Hidden versions of treatment are another major source of SUTVA violations. Common TSCS practices for estimating the effects of ‘sticky’ treatments can inadvertently create hidden versions of treatment. For example, when estimating the effect of a binary democracy variable on an outcome like trade flows, researchers typically estimate a model that compares each observed year of democracy to each year of non-democracy. However, the first year of democracy after democratization may not be equivalent to a year of mature democracy many years after democratization. Conceptualizing the treatment variable as a *transition to treatment* and using countries, rather than country-years, as the unit of analysis avoids this problem.

SUTVA is a strong assumption that may not be plausible in many TSCS applications. If so, TSCS matching will still mitigate model specification dependence but the results should probably not be interpreted causally. There has been very little research exploring when and how violations of SUTVA are problematic and methods for proceeding without SUTVA are in their infancy (Tchetgen Tchetgen and Vanderweele, 2012). The standard practice for applied researchers facing possible SUTVA violations is to simply proceed as if they did not exist. Rather than ignoring SUTVA violations entirely, first order violations of SUTVA may be avoided by following the advice above.

## ***Estimation after Matching***

What should analysts do after creating a matched TSCS sample? Following Ho et al. (2007), analysts may be able to proceed with the model they would have used on unmatched data. The data may need to be transformed back into ‘long’ format for use with popular TSCS model software. However, analysts may realize in the course of matching that the pooled regression they initially wanted to estimate probably violates the assumptions above.

If so, a reasonable approach is to estimate treatment effects using the difference-in-differences estimator (Keele, Chapter 43, this *Handbook*). This estimator assumes that the units observed have parallel over-time trajectories prior to treatment. If so, differences in the differences of their outcome trajectories after treatment of some panels are the effect of treatment (hence the name). However, it is difficult to find settings where the parallel trends assumption is plausible. Heckman et al. (1997) have suggested that researchers use matching to select units with similar trends prior to difference-in-differences estimation. The TSCS matching methods outlined above and in Imai et al. (2020) offer a way to perform this matching.

## **MATCH OR MISMATCH?**

Will matching help me get better, more reliable answers to my research questions? This is a crucial question for any International Relations scholar considering using statistical matching in their work. With over a decade of matching under our belts, we can step back and evaluate its usefulness for IR. Potential criticism comes in two varieties: matching is generally problematic, and matching is especially problematic when applied to the TSCS data sets common in IR.

To review, matching is a method for selecting cases to use in a subsequent statistical analysis. A matched sample is fully characterized by a data matrix of cases and a set of

weights (generally derived from a matching algorithm) for each case. Observations with weights of zero are considered ‘pruned’ from the data set, resulting in a data set that is more ‘balanced’ than the original (meaning treated and control observations are more similar on a set of possible confounders represented by  $X$ ). Once a matched sample is in hand, a researcher can typically use any appropriate regression technique to analyze the data without much modification.

As Geddes (1990) reminds us, the cases we choose affect the results we get. Because matching is case selection based on  $X$ , calculations based on the matched sample are conditional on  $X$  whether the subsequent estimation is explicitly conditional or not. This means that matching ‘controls’ for  $X$  in ways similar to regression, with added flexibility for interactions and ensuring common support. It also means that statistical results from matched data will depend less on which specific regression model an analyst chooses than results from unmatched data (Ho et al., 2007).

Critics of matching warn that it can produce misleading results (Arceneaux et al., 2006) and facilitate false discovery through p-hacking (Miller, 2019). At seminars and conferences, I have seen critics scold matching advocates for overstating the benefits (and benefiting professionally from those overstatements!). It is fair to say that some proponents of matching (including myself) have at times been overly sanguine about the method, though the same could be said for virtually every statistical technique. Perceptions that researchers benefit professionally from applying matching in their research are also probably correct. Using the data set of IR articles I described above, I estimate that articles using matching have gotten ten more citations on average than comparable articles that did not use matching.<sup>9</sup> It is worth asking whether this additional attention is warranted.

The critics have a point. Matching is not a statistical truth serum to extract causal information from even the most recalcitrant data sets. There is no secret ‘causality’ sauce in

the inner workings of matching methods. Matching is for researchers who want to make causal inferences, but it won’t necessarily allow them to do so. Matching cannot transform your unruly data into a neat and tidy experimental study. Moreover, existing matching methods have been used in an off-the-shelf manner not appropriate for the TSCS structure of many IR data sets. And it is unlikely that matching alleviates the risk of fishing for desired results despite early optimism on this score. Now that a decade has passed, my sense is that the critics of matching have successfully tempered these overly-optimistic claims about what matching can do.

Matching is not magic, but it is not all hype either. The upside of matching is that it is likely to help you understand your data better and make clear to you and your readers what assumptions underlie your conclusions. Fitting a good regression model for causal inference can be very hard, and bad regression models abound. My experience is that matching makes the assumptions necessary for reliable causal inference more transparent to researchers. The intuition of matching is easier for many people than the intuition of regression, and understanding the intuition can help analysts avoid pitfalls. Matching has proven to be a reliably useful method for causal inference with messy, non-experimental data, precisely the sort that most of us in IR are used to dealing with. Although conditioning with regression can get the same result as conditioning with matching, I generally find that someone using matching is going to have a better sense for whether the counterfactuals implied by their model make any sense.

Yet, even if matching is useful for cross-sectional data analysis, it may be that the dependencies of time-series cross-sectional data are too challenging for existing matching methods to tackle. I acknowledge that the challenges are formidable. It is difficult to confidently assert that the stable unit treatment value assumption and conditional ignorability have been satisfied in any

cross-national comparison study. However, these assumptions are necessary for making credible causal inferences from regressions analysis without matching. Experiments are currently infeasible for many pressing IR questions and genuine natural experiments are rare. This leaves us with the necessity of extracting the most credible inferences we can from the observational data provided by history. Matching offers a transparent, useful framework for doing so while foregrounding the challenges.

So, where does this leave us? Matching is here to stay and rightfully belongs in the tool kit of IR scholars. It is no silver bullet. Like any other method, it can be misused, oversold, and misunderstood. It should not be, if we care about learning the truth about international relations.

## Notes

- 1 Department of Political Science, MIT. Eliza Riley provided research assistance. Luigi Curini, Elizabeth Dekeyser, Rob Franzese, Kosuke Imai, In Song Kim, Kacie Miura, and Eliza Riley generously gave advice.
- 2 Matching is also useful as a case selection strategy for some qualitative studies. See Nielsen (2016) for explanation and Weeks (2018) for an example. However, my focus in this chapter is on matching for statistical inference.
- 3 I consider IR data to include either an international predictor variable, an international outcome variable, or both. I use the TRIP article database (Teaching, Research, and International Policy Project, 2017; Maliniak et al, 2018) to identify IR articles in 12 journals: APSR, AJPS, BJPS, JOP, IO, IS, ISQ, WP, SS, JCR, JPR, and EJIR. Using several methods for searching, I find 124 articles that use matching. A full list is in the online supplemental information.
- 4 These papers are collected into one volume in Rubin (2006).
- 5 Some matching algorithms also give units non-integer weights, which can be thought of as ‘partial inclusion’ in the matched sample.
- 6 Mahalanobis distance is a generalization of Euclidean distance that is unit-less, scale-invariant, and accounts for correlated variables.
- 7 See the online supplemental information for a list with citations.

- 8 There are many complicated issues involved with making causal inferences about time-varying (dynamic) treatments and treatment regimes (Murphy et al., 2001). But most IR scholars using matching have worked with static treatments, so I focus on how to extend the standard matching apparatus for static treatments to TSCS data.
- 9 See the online Supplemental Information (Nielsen, 2019) at <https://doi.org/10.7910/DVN/HEFNHA>.

## REFERENCES

- Abadie, Alberto, and Guido W. Imbens. On the Failure of the Bootstrap for Matching Estimators. *Econometrica* 76(6) (2008): 1537–1557.
- An, Weihua. Bayesian Propensity Score Estimators: Incorporating Uncertainties in Propensity Scores into Causal Inference. *Sociological Methodology* 40(1) (2010): 151–189.
- Arceneaux, Kevin, Gerber, Alan S., and Green, Donald P. Comparing Experimental and Matching Methods Using a Large-Scale Field Experiment on Voter Mobilization. *Political Analysis* 14(1) (2006): 37–62.
- Arceneaux, Kevin, Gerber, Alan S., and Green, Donald P. A Cautionary Note on the Use of Matching to Estimate Causal Effects: An Empirical Example Comparing Matching Estimates to an Experimental Benchmark. *Sociological Methods & Research* 39(2) (2010): 256–282.
- Aronow, Peter M., and Cyrus Samii. Estimating Average Causal Effects Under General Interference, with Application to a Social Network Experiment. *The Annals of Applied Statistics* 11(4) (2017): 1912–1947.
- Bowers, Jake, and Thomas Leavitt. ‘Causality and Design-Based Inference’ in *The SAGE Handbook of Research Methods in Political Science and International Relations*, Luigi Curini and Robert Franzese, editors, Sage (2020): 771–806.
- Carter, Christopher L., and Thad Dunning. ‘Instrumental Variables: From Structural Equation Models to Design-Based Causal Inference’ in *The SAGE Handbook of Research Methods in Political Science and International Relations*, Luigi Curini and Robert Franzese, editors, Sage (2020): 750–770.
- Cattaneo, Matias D., Rocío Titiunik, and Gonzalo Vazquez-Bare. ‘The Regression Discontinuity

- Design' in *The SAGE Handbook of Research Methods in Political Science and International Relations*, Luigi Curini and Robert Franzese, editors, Sage (2020): 837–859.
- D'Agostino Jr, Ralph B., and Donald B. Rubin. Estimating and Using Propensity Scores with Partially Missing Data. *Journal of the American Statistical Association* 95(451) (2000): 749–759.
- Diamond, Alexis, and Jasjeet S. Sekhon. Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies. *Review of Economics and Statistics* 95(3) (2013): 932–945.
- Findley, Michael G., Daniel L. Nielson, and Jason C. Sharman. Using Field Experiments in International Relations: A Randomized Study of Anonymous Incorporation. *International Organization* 67(4) (2013): 657–693.
- Geddes, Barbara. How the Cases You Choose Affect the Answers You Get: Selection Bias in Comparative Politics. *Political Analysis* 2 (1990): 131–150.
- Hansen, Ben B. Full Matching in an Observational Study of Coaching for the SAT. *Journal of the American Statistical Association* 99(467) (2004): 609–618.
- Hansen, Ben B., and Jake Bowers. Covariate Balance in Simple, Stratified and Clustered Comparative Studies. *Statistical Science* 23(2) (2008): 219–236.
- Heckman, James, Hidehiko Ichimura and Petra Todd. 1997. Matching as an Econometric Evaluation Estimator. *Review of Economic Studies* 65(2) (1997): 261–294.
- Hirano, Keisuke, and Guido W. Imbens. The Propensity Score with Continuous Treatments. *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives* 226164 (2004): 73–84.
- Ho, Daniel E., Kosuke Imai, Gary King, and Elizabeth A. Stuart. Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference. *Political Analysis* 15(3) (2007): 199–236.
- Holland, Paul W. Statistics and Causal Inference. *Journal of the American Statistical Association* 81(396) (1986): 945–960.
- Hollyer, James R., and B. Peter Rosendorff. Leadership Survival, Regime Type, Policy Uncertainty and PTA Accession. *International Studies Quarterly* 56(4) (2012): 748–764.
- Iacus, Stefano M., Gary King, and Giuseppe Porro. Multivariate Matching Methods that are Monotonic Imbalance Bounding. *Journal of the American Statistical Association* 106(493) (2011): 345–361.
- Iacus, Stefano M., Gary King, and Giuseppe Porro. Causal Inference without Balance Checking: Coarsened Exact Matching. *Political Analysis* 20(1) (2012): 1–24.
- Iacus, Stefano M., Gary King, and Giuseppe Porro. A Theory of Statistical Inference for Matching Methods in Causal Research. *Political Analysis* 27(1) (2019): 46–68.
- Imai, Kosuke, and David A. Van Dyk. Causal Inference with General Treatment Regimes: Generalizing the Propensity Score. *Journal of the American Statistical Association* 99(467) (2004): 854–866.
- Imai, Kosuke, and In Song Kim. When Should We Use Unit Fixed Effects Regression Models for Causal Inference with Longitudinal Data? *American Journal of Political Science* 63(2) (2019): 467–490.
- Imai, Kosuke, In Song Kim, and Erik Wang. 2020. Matching Methods for Causal Inference with Time-Series Cross-Section Data, Available at <http://web.mit.edu/insong/www/pdf/tscs.pdf> (Accessed on 15 January, 2020).
- Imai, Kosuke, and Marc Ratkovic. Covariate Balancing Propensity Score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76(1) (2014): 243–263.
- Imbens, Guido W. Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009). *Journal of Economic Literature* 48(2) (2009): 399–423.
- King, Gary, Christopher Lucas, and Richard A. Nielsen. The Balance-Sample Size Frontier in Matching Methods for Causal Inference. *American Journal of Political Science* 61(2) (2017): 473–489.
- King, Gary, and Richard Nielsen. Why Propensity Scores Should Not Be Used for Matching. *Political Analysis* 27(4) (2019): 435–454.
- Keele, Luke. 'Difference-in-Differences: Neither Natural nor an Experiment' in *The SAGE Handbook of Research Methods in Political Science and International Relations*, Luigi Curini and Robert Franzese, editors, Sage (2020): 824–836.

- Maliniak, Daniel, Susan Peterson, Ryan Powers, and Michael J. Tierney. Is International Relations a Global Discipline? Hegemony, Insularity, and Diversity in the Field. *Security Studies*, 27(3) (2018): 448–484.
- Marshall, Monty G., Keith Jagers, and Ted Robert Gurr. Polity IV Project: Dataset Users' Manual. College Park: University of Maryland (2002). Available at <http://www.systemicpeace.org/inscr/p4manualv2018.pdf> (Accessed on 15 January, 2020).
- Miller, Michael K. The Uses and Abuses of Matching in Political Science. Unpublished. <https://sites.google.com/site/mkmtwo/Miller-Matching.pdf> (accessed 2 July, 2019).
- Murphy, Susan A., Mark J. van der Laan, James M. Robins, and Conduct Problems Prevention Research Group. Marginal Mean Models for Dynamic Regimes. *Journal of the American Statistical Association* 96(456) (2001): 1410–1423.
- Nielsen, Richard A. Case Selection via Matching. *Sociological Methods & Research* 45(3) (2016): 569–597.
- Nielsen, Richard. Supplemental Information for: 'Statistical Matching with Time-Series Cross-Sectional Data: Magic, Malfeasance, or Something in Between?' Harvard Dataverse (2019) <https://doi.org/10.7910/DVN/HEFNHA>.
- Nielsen, Richard A., and Beth A. Simmons. Rewards for Ratification: Payoffs for Participating in the International Human Rights Regime? *International Studies Quarterly* 59(2) (2015): 197–208.
- Nielsen, Richard, and John Sheffield. *Matching with Time-Series Cross-Sectional Data*. Polmeth XXVI, Yale University (2009).
- Otsu, Taisuke, and Yoshiyasu Rai. Bootstrap Inference of Matching Estimators for Average Treatment Effects. *Journal of the American Statistical Association* 112(520) (2017): 1720–1732.
- Roberts, Margaret E., Brandon M. Stewart, and Richard Nielsen. Adjusting for Confounding with Text Matching. *American Journal of Political Science* Forthcoming (n.d.).
- Rosenbaum, Paul R. Optimal Matching for Observational Studies. *Journal of the American Statistical Association* 84(408) (1989): 1024–1032.
- Rosenbaum, Paul R., and Donald B. Rubin. The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika* 70(1) (1983): 41–55.
- Rosenbaum, Paul R., and Donald B. Rubin. The Bias Due to Incomplete Matching. *Biometrics* 41(1) (1985): 103–116.
- Rubin, Donald B. Matching to Remove Bias in Observational Studies. *Biometrics* 29(1) (1973): 159–183.
- Rubin, Donald B. Bias Reduction Using Mahalanobis-Metric Matching. *Biometrics* 36(2) (1980): 293–298.
- Rubin, Donald B. *Matched Sampling for Causal Effects*. New York: Cambridge University Press (2006).
- Rubin, Donald B., and Neal Thomas. Matching Using Estimated Propensity Scores: Relating Theory to Practice. *Biometrics* 52(1) (1996): 249–264.
- Sekhon, Jasjeet S. Opiates for the Matches: Matching Methods for Causal Inference. *Annual Review of Political Science* 12 (2009): 487–508.
- Simmons, Beth A., and Daniel J. Hopkins. The Constraining Power of International Treaties: Theory and Methods. *American Political Science Review* 99(4) (2005): 623–631.
- Stuart, Elizabeth A. Matching Methods for Causal Inference: A Review and a Look Forward. *Statistical Science: A Review Journal of the Institute of Mathematical Statistics* 25(1) (2010): 1.
- Tarrow, Sidney. The Strategy of Paired Comparison: Toward a Theory of Practice. *Comparative Political Studies* 43(2) (2010): 230–259.
- Tchetgen Tchetgen, Eric J., and Tyler J. VanderWeele. On Causal Inference in the Presence of Interference. *Statistical Methods in Medical Research* 21(1) (2012): 55–75.
- Teaching, Research, and International Policy Project. TRIP Journal Article Database Release (Version 3.1). (2017). Available at <https://trip.wm.edu/>.
- Weeks, Ana Catalano. Why Are Gender Quota Laws Adopted by Men? The Role of Inter- and Intraparty Competition. *Comparative Political Studies* 51(14) (2018): 1935–1973.
- Zubizarreta, José R., Ricardo D. Paredes, and Paul R. Rosenbaum. Matching for Balance, Pairing for Heterogeneity in an Observational Study of the Effectiveness of For-Profit and Not-For-Profit High Schools in Chile. *The Annals of Applied Statistics* 8(1) (2014): 204–231.

# Differences-in-Differences: Neither Natural nor an Experiment

Luke Keele

## INTRODUCTION

When analysts seek to estimate causal effects, one of the most commonly used tools is differences-in-differences (DID). Widespread use of DID is most likely a result of two factors. One is that there are many cases where the method can be used. That is, it can be used in any context where before and after data exist for treated and control groups. Moreover, estimation can be done with linear regression models and, as such, no special software or programming is necessary. Another reason DID is widely employed is that many investigators assume that if one can apply DID, then the analysis tends to be classified as a type of natural experiment. Evidence from natural experiments is now widely viewed as the best alternative to evidence from randomized trials. As such, social scientists are often urged to identify a natural experiment when randomization is not possible (Angrist and Pischke, 2010). Research that can be classified as a natural experiment

is often imbued with a sense of greater credibility. Given that DID is often classified as a type of a natural experiment, I would argue this has given the method wider popularity.

In this essay, I highlight several issues in the use of DID. I specifically focus on whether an application that uses DID should be classified as a natural experiment. I argue that use of DID does not imply that one is analyzing a natural experiment. DID can be applied to any application with the right data configuration, and that configuration of the data does not imply that the treatment assignment mechanism is as-if randomized, which is the hallmark of a natural experiment. While DID can be readily applied to settings that are natural experiments, it is often applied to contexts where nothing about the treatment assignment mechanism implies a natural experiment. Next, I demonstrate that while DID protects against a specific form of bias, there are also instances where the use of DID may increase the bias in treatment effect estimates compared with a more basic

treatment and control comparison. Turning to estimation, I describe a specific form of matching that can be used as an alternative to linear regression models. Finally, I present an empirical example to demonstrate concepts. In the example, I use data on the introduction of election day registration in the state of Wisconsin. This essay is not, however, a tutorial on DID for readers hoping to learn the method for the first time. Readers interested in basic introduction to DID should see Angrist and Pischke (2009: ch. 5). Readers interested in a complete review of estimation methods for DID should see Lechner (2011).

## NOTATION AND METHOD

Here, I outline notation for the basic DID framework following the notation in Abadie (2005). In this framework, we observe data at the individual level for at least two time periods. For each unit  $i$  at time  $t$ , we denote the outcome as  $Y(i, t)$ , where  $t = 0$  for the pre-treatment time period and  $t = 1$  for the post-treatment time period. Some fraction of the units are exposed to a treatment between  $t = 0$  and  $t = 1$ . We denote a unit as treated if  $D(i, 1) = 1$ , and units are controls if  $D(i, 0) = 0$ . Given that we stipulate treatment is only given after  $t = 0$ , we can simplify the notation for treatment to  $D(i)$ . Next, for each unit, we define the effect of the treatment in terms of potential outcomes. The potential outcome if unit  $i$  at time  $t$  is untreated is  $Y^0(i, t)$ . Similarly, the potential outcome under treatment for unit  $i$  and time  $t$  is  $Y^1(i, t)$ . The unit level causal effect of the treatment is  $Y^1(i, t) - Y^0(i, t)$ . The fundamental problem of causal inference is that we cannot observe both potential outcomes for each unit. Instead, we can only observe the realized outcome which is a function of treatment status and potential outcomes:  $Y(i, t) = Y^0(i, t)(1 - D(i)) + Y^1(i, t)D(i)$ . Hereafter, we drop  $i$  to simplify the notation.

While individual level treatment effects cannot be computed, researchers generally focus on some population average effect. In the DID context, that effect is usually defined as the average effect of the treatment on the treated:  $E[Y^1(1) - Y^0(1) | D = 1]$ . However, further assumptions are needed to identify the causal effect. Specifically, in a DID analysis, we assume:

$$\begin{aligned} E[Y^0(1) - Y^0(0) | D = 1] \\ = E[Y^0(1) - Y^0(0) | D = 0] \end{aligned}$$

In words, this assumption requires that, in the absence of treatment, the potential outcomes for treated and control units are following the same common time trend. That is, we assume that the treated and control units are evolving over time, and absent treatment the two groups would evolve the same way. The only factor that causes the treated units to deviate from this common time trend is exposure to the treatment. Typically, researchers make this assumption conditional on a set of covariates:

$$\begin{aligned} E[Y^0(1) - Y^0(0) | X, D = 1] \\ = E[Y^0(1) - Y^0(0) | X, D = 0] \end{aligned}$$

Now, we stipulate that the outcomes of treated and control units that look similar in terms of observed characteristics change at the same rate across time. Under either of these identifying assumptions, one can straightforwardly estimate the DID treatment effect.

Next, I change the notation to compactly refer to outcomes in each for the four treatment time periods. That is, we now denote the observed outcome as  $Y(t, d)$  where  $Y(1, 1)$  refers to treated outcomes in the post-treatment time period. The DID estimator for the treatment effect of  $D$  is

$$\begin{aligned} \tau = & (\mathbb{E}[Y(1, 1)] - \mathbb{E}[Y(0, 1)]) \\ & - (\mathbb{E}[Y(1, 0)] - \mathbb{E}[Y(0, 0)]), \end{aligned}$$



Close examination of the basic DID estimator reveals the logic behind the term ‘differences-in-differences’. The DID estimator is the difference between the treated outcome trend and the control outcome trend. That is, one takes the difference in the outcomes after removing the time trends for the treated and control groups. Moreover, investigators can use a linear regression model to estimate the DID effect as well, which also simplifies inference when the data are IID (Donald and Lang, 2007; Bertrand et al., 2004).

The simple form of the estimator also makes DID very flexible in terms of data requirements. That is, DID can be applied to data settings where the same units are observed over time, but it can also be applied to settings where different units are observed over time. For example, survey data may be collected before and after a treatment goes into effect. The survey data need not be collected from the same individuals before and after treatment. This has made DID highly applicable to studying the effect of policy changes. That is, so long as survey data can be collected before and after a new law or policy is implemented, then DID can be easily applied.

DID is considered useful since it removes two types of bias. The first bias is a uniform time trend, which we denote  $\lambda_t$ , and the second bias is a constant difference between treated and control groups, which we denote  $\lambda_d$ . If only these two biases are present, then  $Y(1,1) - \tau = Y(0,0) + \lambda_t + \lambda_d$  where  $\tau$  is a constant-additive treatment effect. The differences-in-differences contrast:  $(Y(1,1) - Y(0,1)) - (Y(1,0) - Y(0,0))$  removes the bias from  $\lambda_t$  and  $\lambda_d$ . As such, DID removes these two forms of additive bias even if they arise from observed and unobserved sources. Thus, the particular advantage of DID is that it provides some protection against unobserved confounding. However, as I outline below, this leverage over unobservables is fairly limited.

## HISTORY

The DID method is one of the oldest in social science data analysis. The first recorded

example of DID is in Jon Snow’s ground-breaking study showing that cholera was a waterborne disease (Snow, 1854). What is perhaps more striking is how another early application of DID is remarkably similar to how DID is currently employed. Obenauer et al. (1915) sought to estimate the effect of a minimum wage law in Oregon that led to higher wages in Portland but not the rest of the state. In their study, they used Salem – another city in Oregon that the authors thought was similar to Portland – as the control group. They then compared the overtime changes in employment in Portland to the overtime changes in employment in Salem. Other early studies are also typical of current DID usage: a change in state or local laws is the treatment of interest. Data are then collected in the pre and post-treatment time periods for a treated group, the place with the new policy, and a control group, some place without a change in policy (Lester, 1946; Rose, 1952). In political science, changes in state regulation of the voting registration process are routinely analyzed using DID (Leighley and Nagler, 2013; Hanmer, 2009). Again, one reason for widespread use of DID is that the data configuration that makes DID possible is common. Moreover, as I outlined above, DID does protect against a specific form of unobserved bias. In sum, the attraction of DID as a strategy for estimating treatment effects is fairly obvious. Next, I review several issues related to DID.

## ISSUES IN DID

### *Scale Dependence*

As I highlighted above, the strength of DID is that it can remove bias due to time-invariant confounders. That is, if there are unobserved differences in the treated and control groups that do not change over time, those differences are removed by applying DID to the data. That would appear to be a significant advantage, given that bias from unobserved confounders is the specific weakness of most attempts to

estimate causal effects with observational data. However, closer examination reveals that the bias reducing properties of DID are highly dependent on functional form assumptions. That is, for DID to be successful, the bias can only be additive. There is, however, no reason to assume the bias is additive. One way to understand how restrictive this assumption is is to note that the bias reduction properties of DID do not hold under logarithmic transformations.

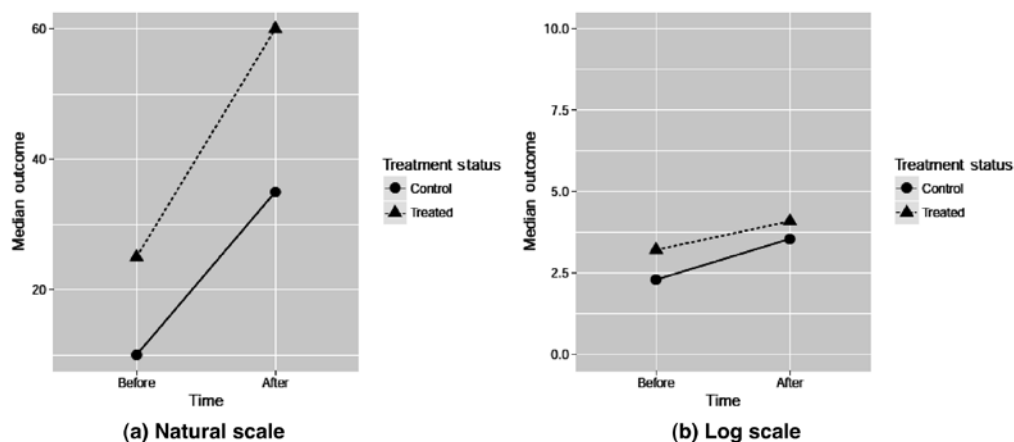
The left panel of Figure 43.1 contains the specific pattern in data for which DID is useful. By construction, both the treated and control groups increase over time by 25 units due to fixed over time changes in outcomes. The treated group outcome increases an additional 10 points due to exposure to the treatment. The pattern in the left panel of Figure 43.1 would appear to be the prototypical example of why one should use DID. However, let's say that an analyst decides to log transform this data, which is often done to decrease the variability in the outcome. The right panel of Figure 43.1 contains the same data after a log transformation of the outcome. Notice, however, that under the log transform, the treatment effect is mostly eliminated. This is due to the fact that the additive bias removed by DID may not hold for a strictly increasing transformation of the outcome. That is, when the bias takes any form that is more complex

than additive, DID offers little protection. There is, however, no reason to think that bias from unobservables takes such a simple form.

### Assignment Mechanism

One trend in empirical analysis over the last 15 years has been an increased use of natural experiments. Given that randomized experiments are often infeasible in many empirical settings, investigators attempt to find natural circumstances that assign treatments in a manner something like a randomized experiment. The hope is to reduce bias from confounding by exploiting circumstances where treatments are not purposely assigned. What is a natural experiment? Keele (2015) defines a natural experiment as a real-world situation that produces haphazard assignment to a treatment. See Dunning (2012) for a useful review of how to judge the quality of a natural experiment relative to actual randomization.

The DID method is closely identified with natural experiments. In fact, many authors often refer to any use of DID as a type of natural experiment. For example, Mayne et al. (2015) conducted a review of natural experiments in the study of obesity. Any study that applied differences-in-differences was classified as a natural



**Figure 43.1** Schematic representation of response in treated and control groups, before and after treatment, with and without transformation to log scale

experiment in their review. The close identification between DID and natural experiments is probably due to the fact that DID has been used to analyze several famous natural experiments including Jon Snow's cholera study. Another well-known example where DID was applied to natural experiment includes the study by Card (1990) which used the Mariel Boatlift from Cuba to estimate the effect of immigration on employment rates. Dynarski (1999) studied the treatment effect of the additional aid on the decision to attend college, using the Social Security Student Benefit Program, which awarded aid to high school seniors with deceased fathers of Social Security recipients. Referring to DID as a type of natural experiment tends to imbue the method with a level of credibility that is not deserved, since there is nothing about DID that implies the study is a natural experiment. One way to understand this limitation is by reviewing the concept of treatment assignment mechanisms.

In the literature on casual inference, treatment assignment mechanisms are often an area of emphasis (Rubin, 2008). For example, the power of randomized experiments largely stems from the fact that the treatment assignment mechanism follows a known probabilistic rule. An application can be categorized as a convincing natural experiment when the treatment assignment mechanism appears as-if random. For example, Lyall (2009) used artillery fire by Russian troops in Chechnya to understand whether indiscriminate violence increases insurgent attacks. More specifically, he exploited the fact that the shelling patterns of the Russian army appeared to be at worst indiscriminate and at best as-if random. That is, instead of firing on known military targets, the troops fired in what may have been random patterns. Research studies that utilize natural experiments are advised to use considerable qualitative knowledge to describe the process by which haphazard treatment occurs and to compare the results of haphazard assignment to what would have occurred under true random assignment (Dunning, 2012). That is,

when investigators claim to be exploiting a natural experiment, it requires considerable evidence to make a convincing case. Lyall (2009), for example, does find that the treatment, being shelled, appears to be uncorrelated with pre-treatment covariates as would be the case in a randomized experiment. Regression discontinuity designs are built on a series of falsification tests that compare haphazard assignment to random assignment (Cattaneo et al., 2018: ch. 5).

The critical point is that use of DID tells you nothing about the assignment mechanism. That is, DID can be applied to applications where treatments are assigned in a highly plausible, as-if random fashion. However, DID can also be applied to contexts where treatments are applied in a highly selective fashion. As an example, Mayne et al. (2015) classify studies that use DID to estimate the effect of built environment interventions on obesity as natural experiments. In a built environment intervention study, the investigator would be seek to estimate the effect of building a new playground on health outcomes in a community (Gustat et al., 2012). Like many policy changes, there is little that is haphazard about an intervention of this type. That is, without clear evidence to the contrary, it seems highly unlikely that a city would choose neighborhoods for playgrounds in an as-if random fashion. In all probability, the city government selects areas for new playgrounds in a highly purposive fashion. For example, the city may select neighborhoods where the local citizens are highly engaged and have strongly lobbied for a new playground. Such places are a likely to be different from other places in both observed and unobserved ways. The key point is that being able to apply DID to a specific application does not make it a natural experiment.

As I noted above, DID is widely used to evaluate the effect of changes in state or local laws. For example, DID might be used to evaluate a new voting regulation such as election day registration (EDR). In general, such changes in state policy are not natural experiments in that changes in such laws are typically highly purposeful. It is not an accident

that two upper Midwestern states – Minnesota and Wisconsin – with high levels of social capital enacted EDR years before any other state. The state legislatures in Minnesota and Wisconsin adopted these laws not by accident, but to maximize turnout in those states (Smolka, 1977). Moreover, the reason those two states adopted EDR is a function of most likely both observed and unobserved differences in the civic cultures of those states relative to other states.

In general, the DID framework should not be conflated with the concept of natural experiments. DID is essentially an estimation strategy that is applied to a particular data configuration in hopes that it removes a specific form of bias. Conflating DID with natural experiments gives it a status it does not deserve. Applications should be classified as natural experiments based on whether the assignment mechanism appears to approximate as-if random assignment. Widespread use of DID results from the fact that the necessary data configuration is common. That is, frequent use of DID is not unlike assuming that ‘one’s cancer is curable by a particular drug simply because one has an abundant supply of this specific drug’.<sup>1</sup>

### ***Bias Due to an Unobserved Time-Varying Confounder***

While the benefits of DID tend to be well understood, the general weakness of the method tends to be less well understood. Next, I use a simple numerical example to demonstrate how time-varying confounders can undermine estimates based on DID in a specific way. In this hypothetical example, the true outcome values for the four groups are as follows:

- $\mathbb{E}[Y(1,1)] = 75$ ,
- $\mathbb{E}[Y(0,1)] = 15$ ,
- $\mathbb{E}[Y(1,0)] = 25$ ,
- $\mathbb{E}[Y(0,0)] = 15$ .

The DID estimate of  $\tau$  is  $E[75 - 15] - E[25 - 15] = 50$ . Next, we define  $\gamma(t, d)$  as bias from a confounder that alters the average

outcomes in at least one or possibly more of the four groups. For example,  $\gamma(1, 0)$  is a bias that increases the average outcome in the control group when  $t = 1$ . First, we consider a bias of the form  $\gamma(1, 1) = \gamma(1, 0) = 10$ , where both treated and control outcomes increase when  $t = 1$  for reasons unrelated to treatment. If we apply the usual DID estimator, it is still the case that  $\hat{\tau} = 50$ . That is, when the bias takes this specific time-invariant form, DID works as advertised and removes that bias. Next, we consider a bias of the form  $\gamma(0, 0) = \gamma(1, 1) = 20$ . This type of bias increases outcomes in the control when  $t = 0$  and increases outcomes in the treated when  $t = 1$ . Here, the bias is no longer time invariant. Under a bias of this type, the DID estimate is now  $E[95 - 15] - E[25 - 35] = 90$ . A bias of this form distorts in a particular way such that DID is unable to offer any protection. Moreover, for this type of bias, if we did not apply DID and simply used the treated and control difference in the post-treatment time period, the estimate would be  $95 - 25 = 70$ . The key point is that under the right form of time-varying bias, one would be better off simply using a standard treated control difference instead of applying DID. This is due to the fact that under DID, one must assume that the bias only takes a specific time-invariant form.

One might reasonably ask whether a bias of this form is plausible? Perhaps a bias that changes outcomes in this specific way over time is hard to imagine and, thus, not something most investigators should worry about. Or perhaps not – let us return to the EDR example and assume we have two states: one adopts EDR and the other does not. It is well understood that during presidential elections, some states are considered battleground states due to the competitive nature of the election in that state. Voters in battleground states tend to be targeted for mobilization by the campaigns, while voters in non-battleground states may be subject to few, if any, mobilization efforts. Some states are perennial battleground states, and other

states become competitive in certain elections. Imagine that when  $t = 0$ , the presidential race is competitive in the control state, and that state is subject to national mobilization efforts, as it is designated a battleground state. However, after EDR goes into effect, the treated state becomes a battleground state while the control state is no longer competitive. As a result, the treated state experiences nationalized get-out-the-vote efforts in the post-treatment time period. A mechanism of this type is completely consistent with a bias of the form  $\gamma(0, 0) = \gamma(1, 1) = \delta$ . That is, it is not difficult to conceptualize a time varying bias of the type that biases differences-in-differences in a serious way.

### ***Bracketing Methods for DID***

DID, at the most basic level, requires longitudinal data. While one need not observe the exact same units over time, the investigator must at the very least observe grouped data over time. As such, in many cases, one alternative to DID is to condition on past outcomes directly. In the simplest form, this approach simply means including lagged outcomes on the right-hand side of a regression model designed to estimate the treatment effect of interest. Under this approach, the investigator is assuming that potential outcomes are independent of treatment once he or she conditions on observed covariates including lagged outcomes. Using the notation from above, the key identification assumption is

$$Y^1(i, t), Y^0(i, t) \perp D_i \mid \mathbf{X}, Y_{it-h}$$

where  $h$  is some unspecified lag length, such that  $h = 2$  would indicate conditioning on two past periods of the outcome. This assumption can be viewed as a variant of the common selection on observable assumption. Here, one assumes that potential outcomes are independent of treatment once one has conditioned on observables including measures of the outcome from time periods before the

treatment is in effect. What is gained by this identification strategy? First, unlike under DID, we need not assume treated and control observations had similar longitudinal trajectories. Here, we can directly control for the histories of each unit. Moreover, past outcomes are a function of both observed and unobserved factors; therefore, conditioning on past outcomes is an indirect way to condition on unobservables. Of course, given that DID can be implemented using regression models, in many cases, it would appear natural to combine the lagged outcome identification strategy with DID. However, OLS estimates are inconsistent in this case (Nickell, 1981). Solutions are possible, but they typically require exclusion restrictions that may be hard to justify. See Xu (2017) for an approach that combines past outcomes and fixed effects but relies on an exclusion restriction.

An alternative approach is to use a solution outlined in Angrist and Pischke (2009) and fully generalized in Ding and Li (2019). Ding and Li (2019) show that if the past outcomes identification strategy holds, but the analyst estimates the treatment effect using DID, the estimated treatment effect will be too large since the unestimated lag parameter will be additive with the treatment effect through the error term. Conversely, if the DID assumptions hold, and the analyst estimates a model that conditions on past outcomes, this will generate a correlation between the treatment and the lagged outcome which will bias the treatment effect downward. One can therefore view the estimates from these two methods as bounds for the causal effect of interest.

How might this work in practice? One key advantage of this approach is that it can be implemented quite simply. First, the investigator would estimate the treatment effect using DID methods. Next, the investigator would estimate the treatment effect controlling for lags of the outcome. In the simplest scenario, one would use linear models in both cases. The estimates from each approach then form bounds for the treatment effect of interest. Ideally, the sign of the two estimates

agree, which provides some evidence that the researcher has at least managed to identify the sign of the treatment effect of interest. See Keele et al. (2013) for an example of using this combined approach. However, in one recent comparative study, conditioning on past outcomes proved to be generally more robust than DID (O'Neill et al., 2016).

## DID ESTIMATION

Estimation in the DID framework typically relies on linear models. See Angrist and Pischke (2009) for a basic introduction and Lechner (2011) for a more detailed treatment of DID estimation. Instead of reviewing linear models for DID estimation, I present an alternative way to estimate treatment effects in the DID framework. Specifically, I present a matching framework outlined in Keele et al. (2018). Next, I then present an actual empirical example using this framework.

### *DID and Matching*

The key advantage to using matching over linear regression is that matching is more robust to a variety of data configurations (Imbens, 2015). Matching was first used for DID estimation in Heckman et al. (1998), but still remains under utilized. Here, I outline one matching plan that can be used to estimate the DID treatment effect in the basic setting with treated and control groups and one pre- and post-treatment time period. This form of DID matching is actually comprised of three different matches. These three matches balance units both with respect to treatment and control but also with respect to time periods. I motivate these matches using a familiar empirical application. Here, I apply this matching plan to estimate the DID treatment effect for EDR. Specifically, I estimate the EDR effect for Wisconsin. Wisconsin was one of the first states to adopt EDR, and

it is a state where the effect of EDR is widely understood to have contributed to an increase in turnout (Hanmer, 2009). The data I use are extracts from the 1972 and 1980 Current Population Survey (CPS) and are a subset of the data from Keele and Minozzi (2012). Wisconsin first used EDR in 1976, and I use turnout levels in the 1980 presidential election as the post-treatment period in case of any delay in the effect of EDR. I use voters from Illinois as controls, since it is adjacent to Wisconsin, and both states have large metropolitan areas with minority communities but have large rural populations as well.

The first match in the plan is to match treated and control units in the pre-treatment period. That is, in this application that requires matching Wisconsin voters to Illinois voters in 1972. The second match matches treated and control units in the post-treatment time period. In the EDR application, I match Wisconsin voters to Illinois voters in 1980. The final match pairs pre-treatment pairs to post-treatment pairs. Specifically, I would match the 1972 Wisconsin and Illinois pairs to the 1980 Wisconsin and Illinois pairs. This final match balances observed covariates with respect to time.

The first two matches are straightforward to implement and almost any form of matching can be used to form these two sets of matched pairs. In the results reported below, I used a matching method based on integer programming in the R package *designmatch* (Zubizarreta, 2012; Zubizarreta and Kilcioglu, 2016; Kilcioglu and Zubizarreta, 2017). The third match, which requires matching matched pairs to matched pairs, is not a standard matching problem. I can reduce this to a standard matching problem by creating a new dataset from the pre-treatment matched pairs and post-treatment matched pairs based on summary statistics. For example, in the EDR data, I calculate the within pair mean for covariates like age. This collapses the data for each set of pairs in each time period. With the collapsed data, standard pair matching techniques can be used to match pre-treatment pairs to post-treatment

pairs. For nominal covariates, exact matching can be applied to avoid having to apply summary statistics. For example, in the CPS data, income is recorded as categories. I use exact matching when forming the first two sets of pairs. Thus, when I collapse the matched pairs, no summary statistics are needed. The end result is matched quadruple comprised of two pre-treatment units and two post-treatment units. Estimation of the treatment effect can be accomplished by applying the usual DID formula. Testing can be done by applying paired tests to the over-time differences in outcomes. Moreover, the method of Rosenbaum bounds can also be applied so that sensitivity to hidden bias is also assessed. See Keele et al. (2018) for details.

When implementing a match of this type, the investigator must carefully select the covariates for matching. In general, one should only match on pre-treatment covariates to avoid bias from adjusting for a post-treatment covariate (Rosenbaum, 1984). The DID match, however, requires matching in the post-treatment period. Therefore, for all stages of the matching, analysts should be careful to match on covariates that are thought to be unaffected by the intervention but may affect the outcome. In the EDR example, education would appear to be a safe covariate to match on, since it is unlikely that a change in voter registration laws would affect levels of education. The reason we match across time is to remove bias due to group status that is thought to vary as a function of observed covariates (Keele et al., 2018).

## Empirical Results

Next, I present the results from the DID matching plan using the CPS data. I begin by matching Wisconsin residents to Illinois residents in 1972. I matched on age, an indicator if he or she is African-American, female, a categorical scale of education, a categorical scale of income and an interaction between education and income categories. In the match, I matched exactly on whether a resident was African-American, and I applied near-fine balance to education, income and the interaction between education and income categories. I allowed for a deviation of two categories on the near-fine balance in the match. After matching in 1972, we have 1,427 matched pairs. Table 43.1 contains the balance statistics for this first match.

I next implemented a match of the same form using the data from 1980. After matching in 1980, we have 1,718 matched pairs. Table 43.2 contains the balance statistics. In both of these matches, the largest discrepancy is for race as Illinois has a larger African-American population. However, after matching, all the standard differences are 0.05 or less.

Finally, I matched the pairs from 1972 to the pairs from 1980. For the pair-to-pair match, I used cardinality matching. After this final match, there are 938 matched pairs from 1972 matched to 938 matched pairs from 1980. Table 43.3 contains the balance statistics. It is worth noting that the imbalances are much larger across the two time periods than within each year. In particular, the standardized difference on income and the income education

**Table 43.1 Standardized differences and p-values for treated to control match in the pre-treatment period for the election day registration application**

	<i>Before matching</i>		<i>After matching</i>	
	<i>Std dif</i>	<i>P-val</i>	<i>Std dif</i>	<i>P-val</i>
Age	0.00	0.98	-0.05	0.19
African-American	-0.31	0.00	0.04	0.13
Female	-0.01	0.72	-0.04	0.28
Education	0.07	0.02	-0.05	0.20
Income	0.02	0.52	-0.05	0.18
Education X income	0.07	0.02	0.05	0.19

**Table 43.2 Standardized differences and p-values for treated to control match in the post-treatment period for the election day registration application**

	<i>Before matching</i>		<i>After matching</i>	
	<i>Std dif</i>	<i>P-val</i>	<i>Std dif</i>	<i>P-val</i>
Age	-0.06	0.04	0.05	0.15
African-American	-0.24	0.00	-0.04	0.13
Female	-0.01	0.77	0.04	0.25
Education	0.17	0.00	-0.05	0.23
Income	0.11	0.00	0.05	0.17
Education X income	0.16	0.00	-0.05	0.19

**Table 43.3 Standardized differences and p-values for treated to control match in the pair-to-pair match for the election day registration application**

	<i>Before matching</i>		<i>After matching</i>	
	<i>Std dif</i>	<i>P-val</i>	<i>Std dif</i>	<i>P-val</i>
Age	0.18	0.00	-0.05	0.29
African-American	-0.05	0.15	-0.07	0.12
Female	0.10	0.00	-0.02	0.67
Education	-0.27	0.00	0.05	0.28
Income	-1.27	0.00	-0.05	0.15
Education X income	-1.10	0.00	-0.05	0.18

interaction both exceed one. This implies fairly large differences in the composition of the samples across years. As is often the case, balance checking reveals what is mostly likely an important lack of overlap in the covariate distributions across the two time periods.

Next, I estimate the DID treatment effect of EDR. Here, I simply apply the usual DID formula to the quadruple outcomes. According to this estimate, the turnout rate increased 12.6 percentage points in Wisconsin as compared with Illinois. The DID estimate of the EDR treatment effect is also statistically significant ( $p < 0.001$ ). However, as noted in Keele et al. (2018), sensitivity analysis reveals that this result could easily be explained by an unobserved confounder.

In general, the EDR application is a case where DID is easily applied but the evidence is unconvincing. First, there is nothing as-if random about treatment assignment in this case. Presumably, lawmakers in Wisconsin selected this voter registration regime for a specific

reason. Thus, we have highly intentional self-selection into treatment. Moreover, we must assume that no other events beside the change in voter registration altered the temporal path of either the treated or control groups. Here, for example, no other events can occur in Wisconsin that might also boost turnout. In general, the key DID assumption is suspect in this case given that there are lots of ways that turnout might have changed overtime for reasons unrelated to EDR. Moreover, the covariates in the CPS data do not measure the temporal dynamics of turnout during this time period, since those tend to be driven by state specific campaign factors. As such, the use of DID does little to enhance the credibility of a study of this type.

## CONCLUSION

All in all, I have presented a fairly pessimistic take on differences-in-differences. My primary



objection is that investigators assume that since they can apply DID, then the design is necessarily a type of natural experiment. While it is true that DID can be applied to natural experiments, DID itself tells one little about whether a set of circumstances are indeed a natural experiment. Once this point is understood, justification of a DID identification strategy should proceed along lines that are similar to if one were assuming selection on observables. That is, a DID identification strategy requires careful explication of the key assumption and a detailed defense of why that assumption is plausible in a given context. One effective way to defend the DID identification strategy is through the use of falsification tests.

I conclude by reviewing one persuasive use of DID based on a falsification or placebo test. Gruber (1994) studied the labor market effects of mandates passed by 23 states between 1975 and 1979 that outlawed treating pregnancy differently from ‘comparable illnesses’, and mandated comprehensive coverage for childbirth in health insurance policies. Mandates of this type increase the costs of employing women of childbearing age and their husbands, under whose insurance these women might have been covered. Gruber employed DID to analyze whether the costs of these mandates were shifted to the wages of married women of childbearing age. Table 43.4 presents the results from a DID analysis. The results based on DID

indicate a decline in wages for married women of child-bearing age during this period  $(1.513 - 1.547) - (1.397 - 1.369) = 0.062$ . The evidence in Table 43.4 is far from conclusive given that a wide variety of other changes to the labor market might explain the decline in wages.

While key assumptions in observational studies are untestable, in many cases these assumptions can be tested indirectly via falsification. Angrist and Krueger (1999) refer to such tests as instances of ‘refutability’. Falsification tests arise from the fact that causal theories may do more than predict the presence of a causal effect; they may also predict an absence of causal effects (Rosenbaum, 2002; Lipsitch et al., 2010). One form of falsification test exploits the fact that treatment is known to be zero in some populations. Gruber (1994) noted that if larger trends in wages could explain the decline in wages observed in Table 43.4, that change in wages should also be present in single men aged between 20 and 40 and women over 40. However, if the decline in wages is due to childbirth mandates, no such treatment effect should be present in this population. Table 43.5 contains the results from the DID applied to women over 40 and single men aged 20–40. Indeed, we observe no significant change in wages for this group  $(1.748 - 1.759) - (1.627 - 1.630) = 0.008$ . That is, in the treated group, wages declined 6.2% while in the placebo test, the decline was an insignificant 0.8%.

**Table 43.4 Average log hourly wages, married women 20–40 years**

	<i>Before law change</i>	<i>After law change</i>
Treated states	1.547	1.513
Control states	1.369	1.397
DID estimate	6.2%	

**Table 43.5 Average log hourly wages, women over 40 and single men 20–40**

	<i>Before law change</i>	<i>After law change</i>
Treated states	1.759	1.748
Control states	1.630	1.627
DID estimate	0.8%	

In general, research design elements of this type tend to be far more effective than when DID is used in isolation. In short, research designs are generally not persuasive just because DID was used. Instead, analysts should make the case for why they are analyzing a natural experiment and use of DID should be largely incidental.

## Note

- 1 This last line is a quote by Paul Rosenbaum from personal communication.

## REFERENCES

- Abadie, Alberto. 2005. Semiparametric Difference-in-Difference Estimators. *Review of Economic Studies* 75(1): 1–19.
- Angrist, Joshua D and Alan B Krueger. 1999. Empirical Strategies in Labor Economics. In *Handbook of Labor Economics*, ed. O. Ashenfelter and D. Card. Vol. 3A, San Diego, CA: Elsevier Science Publishers, pp. 1277–1366.
- Angrist, Joshua D and J'orn-Steffen Pischke. 2009. *Mostly Harmless Econometrics*. Princeton, NJ: Princeton University Press.
- Angrist, Joshua D and J'orn-Steffen Pischke. 2010. The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con Out of Econometrics. *Journal of Economic Perspectives* 24(2): 3–30.
- Bertrand, Marianne, Esther Duflo and Sendhil Mullainathan. 2004. How Much Should We Trust Differences-in-Differences Estimates? *The Quarterly Journal of Economics* 119(1): 249–275.
- Card, David. 1990. The Impact of the Mariel Boatlift on the Miami Labor Market. *Industrial & Labor Relations Review* 43(2): 245–257.
- Cattaneo, Matias D, Nicolás Idrobo and Roc io Titiunik. 2018. *A practical Introduction to Regression Discontinuity Designs: Volume I*. Cambridge; New York: Cambridge University Press.
- Ding, Peng and Fan Li. 2019. A Bracketing Relationship between Difference-in-Differences and Lagged-Dependent-Variable Adjustment. *Political Analysis* 27(4): 605–615.
- Donald, Stephen G and Kevin Lang. 2007. Inference With Differences-In-Differences and Other Panel Data. *The Review of Economics and Statistics* 89(2): 221–233.
- Dunning, Thad. 2012. *Natural Experiments in the Social Sciences: A Design-Based Approach*. Cambridge, UK: Cambridge University Press.
- Dynarski, Susan M. 1999. Does Aid Matter? Measuring the Effect of Student Aid on College Attendance and Completion. *American Economic Review* 93(1): 279–288.
- Gruber, Jonathan. 1994. The Incidence of Mandated Maternity Benefits. *The American Economic Review* 84(3): 622–641.
- Gustat, Jeanette, Janet Rice, Kathryn M Parker, Adam B Becker and Thomas A Farley. 2012. Effect of Changes to the Neighborhood Built Environment on Physical Activity in a Low-income African American Neighborhood. *Preventing Chronic Disease* 9(2): E57.
- Hanmer, Michael J. 2009. *Discount Voting*. New York: Cambridge University Press.
- Heckman, James J, Hidehiko Ichimura and Petra Todd. 1998. Matching as an Econometric Evaluation Estimator. *Review of Economic Studies* 65(2): 261–294.
- Imbens, Guido W. 2015. Matching Methods in Practice: Three Examples. *Journal of Human Resources* 50(2): 373–419.
- Keele, Luke J. 2015. The Statistics of Causal Inference: A View From Political Methodology. *Political Analysis* 23(3): 313–335.
- Keele, Luke J, Jesse Hsu, Dylan S Small and Colin B Fogarty. 2018. *Patterns of Effects and Sensitivity Analysis for Differences-in-Differences*. Available at <https://arxiv.org/pdf/1901.01869.pdf> (Accessed on 15 January, 2020).
- Keele, Luke J, Neil Malhotra and Colin H McCubbins. 2013. Do Term Limits Restrain State Fiscal Policy? Approaches for Causal Inference in Assessing the Effects of Legislative Institutions. *Legislative Studies Quarterly* 38(3): 291–326.
- Keele, Luke J and William Minozzi. 2012. How Much is Minnesota Like Wisconsin? Assumptions and Counterfactuals in Causal Inference with Observational Data. *Political Analysis* 21(2): 193–216.
- Kilcioglu, Cinar and José R Zubizarreta. 2017. Maximizing The Information Content Of a

- Balanced Matched Sample In a Study Of The Economic Performance Of Green Buildings. *The Annals of Applied Statistics* 10(4):1997–2020.
- Lechner, Michael. 2011. The Estimation of Causal Effects by Difference-in-Difference Methods. *Foundations and Trends in Econometrics* 4(3): 165–224.
- Leighley, Jan E and Jonathan Nagler. 2013. *Who Votes now?: Demographics, Issues, Inequality, and Turnout in the United States*. Princeton, NJ: Princeton University Press.
- Lester, Richard A. 1946. Shortcomings of Marginal Analysis for Wage-Employment Problems. *The American Economic Review* 36(1): 63–82.
- Lipsitch, Marc, Eric Tchetgen Tchetgen and Ted Cohen. 2010. Negative Controls: A Tool for Detecting Confounding and Bias in Observational Studies. *Epidemiology (Cambridge, Mass.)* 21(3): 383–388.
- Lyall, Jason. 2009. Does Indiscriminate Violence Incite Insurgent Attacks? Evidence from Chechnya. *Journal of Conflict Resolution* 53(3): 331–362.
- Mayne, Stephanie L, Amy H Auchincloss and Yvonne L Michael. 2015. Impact of Policy and Built Environment Changes on Obesity-Related Outcomes: A Systematic Review of Naturally Occurring Experiments. *Obesity Reviews* 16(5): 362–375.
- Nickell, Stephen. 1981. Biases In Dynamic Models with Fixed Effects. *Econometrica* 49(6): 1417–1426.
- Obenauer, Marie Louise, Burtha von der Nienburg and Royal Meeker. 1915. *Effect of Minimum-Wage Determinations in Oregon*. Washington, DC: US Government Printing Office.
- O'Neill, Stephen, Noémi Kreif, Richard Grieve, Matthew Sutton and Jasjeet S Sekhon. 2016. Estimating Causal Effects: Considering Three Alternatives to Difference-in-Differences Estimation. *Health Services and Outcomes Research Methodology* 16(1-2): 1–21.
- Rose, Arnold M. 1952. Needed Research on the Mediation of Labor Disputes. *Personnel Psychology* 5(3): 187–200.
- Rosenbaum, Paul R. 1984. The Consequences of Adjusting For a Concomitant Variable That Has Been Affected By The Treatment. *Journal of The Royal Statistical Society Series A* 147(5): 656–666.
- Rosenbaum, Paul R. 2002. *Observational Studies*. 2nd ed. New York: Springer.
- Rubin, Donald B. 2008. For Objective Causal Inference, Design Trumps Analysis. *The Annals of Applied Statistics* 2(3): 808–840.
- Smolka, Richard G. 1977. *Election Day Registration: The Minnesota and Wisconsin Experience in 1976*. Washington, DC: American Enterprise Institute for Public Policy Research.
- Snow, John. 1854. The Cholera Near Golden-square, and at Deptford. *Medical Times and Gazette* 9: 321–322.
- Xu, Yiqing. 2017. Generalized Synthetic Control Method: Causal Inference with Interactive Fixed Effects Models. *Political Analysis* 25(1): 57–76.
- Zubizarreta, José R. 2012. Using Mixed Integer Programming for Matching in an Observational Study of Kidney Failure after Surgery. *Journal of the American Statistical Association* 107(500): 1360–1371.
- Zubizarreta, JR and C Kilcioglu. 2016. Designmatch: Construction of Optimally Matched Samples for Randomized Experiments and Observational Studies that are Balanced by Design. R Package Version 0.1.1.

# The Regression Discontinuity Design<sup>1</sup>

Matias D. Cattaneo, Rocío Titiunik, and  
Gonzalo Vazquez-Bare

## INTRODUCTION

The Regression Discontinuity (RD) design has emerged in recent decades as one of the most credible non-experimental research strategies to study causal treatment effects. The distinctive feature behind the RD design is that all units receive a score, and a treatment is offered to all units whose score exceeds a known cutoff, and it is withheld from all the units whose score is below the cutoff. Under the assumption that the units' characteristics do not change abruptly at the cutoff, the change in treatment status induced by the discontinuous treatment-assignment rule can be used to study different causal treatment effects on outcomes of interest.

The RD design was originally proposed by Thistlethwaite and Campbell (1960) in the context of an education policy, where an honorary certificate was given to students with test scores above a threshold. Over time, the design has become common in areas

beyond education and is now routinely used by scholars and policy-makers across the social, behavioral, and biomedical sciences. In particular, the RD design is now part of the standard quantitative toolkit of political science research, and it has been used to study the effect of many different interventions including party incumbency, foreign aid, and campaign persuasion.

In this chapter, we provide an overview of the basic RD framework, discussing the main assumptions required for identification, estimation, and inference. We first discuss the most common approach for RD analysis, the *continuity-based* framework, which relies on assumptions of continuity of the conditional expectations of potential outcomes given the score and defines the basic parameter of interest as an average treatment effect at the cutoff. We discuss how to estimate this effect using local polynomials, devoting special attention to the role of the bandwidth, which determines the

neighborhood around the cutoff where the analysis is implemented. We consider the bias-variance trade-off that is inherent in the most common bandwidth-selection method (which is based on mean-squared-error minimization) and how to make valid inferences within this bandwidth choice. We also discuss the local nature of the RD parameter, including recent developments in extrapolation methods that may enhance the external validity of RD-based results.

In the second part of the chapter, we overview an alternative framework for RD analysis that, instead of relying on continuity of the potential outcome-regression functions, makes the assumption that the treatment is as-if randomly assigned in a neighborhood around the cutoff. This interpretation was the intuition provided by Thistlethwaite and Campbell (1960) in their original contribution, though it now has become less common, due to the stronger nature of the assumptions it requires. We discuss situations in which this *local randomization* framework for RD analysis may be relevant, focusing on cases where the running variable has mass points, which occurs very frequently in applications.

To conclude, we discuss a battery of data-driven falsification tests that can provide empirical evidence about the validity of the design and the plausibility of its key identifying assumptions. These falsification tests are intuitive and easy to implement and thus should be included as part of any RD analysis in order to enhance its credibility and replicability.

Due to space limitations, we do not discuss variations and extensions of the canonical (sharp) RD designs such as fuzzy, kink, geographic, multi-cutoff, or multi-score RD designs. A practical introduction to those topics can be found in Cattaneo et al. (2019a, 2020a) and in the recent edited volume Cattaneo and Escanciano (2017) and the references therein. For a recent review on program evaluation methods see Abadie and Cattaneo (2018).

## GENERAL SETUP

We start by introducing the basic notation and framework. We consider a study where there are multiple units from a population of interest (such as politicians, parties, students, households, or firms), and each unit  $i$  has a *score* or *running variable*, denoted by  $X_i$ . This running variable could be, for example, a party's vote share in a congressional district, a student's score from a standardized test, a household's poverty index, or a firm's total revenue over a certain period of time. This running variable may be continuous, in which case no two units will have the same value of  $X_i$ , or not continuous, in which case the same value of  $X_i$  might be shared by multiple units. The latter case is usually called 'discrete', but in many empirical applications the score variable is actually both.

In the simplest RD design, each unit receives a binary treatment  $D_i$  when their score exceeds some fixed threshold  $c$  and does not receive the treatment otherwise. This type of RD design is commonly known as the *sharp RD design*, where the word *sharp* refers to the fact that the assignment of treatment coincides with the actual treatment taken – that is, compliance with treatment assignment is perfect. When treatment compliance is imperfect, the RD design becomes a *fuzzy RD design* and its analysis requires additional methods beyond the scope of this chapter. The methods described here for analyzing sharp RD designs can be applied directly in the context of fuzzy RD designs when the parameter of interest is the intention-to-treat effect.

The sharp RD treatment assignment rule can be formally written as

$$D_i = \mathbb{I}(X_i \geq c) = \begin{cases} 1 & \text{if } X_i \geq c \\ 0 & \text{if } X_i < c \end{cases} \quad (1)$$

where  $\mathbb{I}(\cdot)$  is the indicator function. For example,  $D_i$  could be a scholarship for college students that is assigned to those with a

score of seven or higher in an entry exam on a scale from 0 to 10. In this example,  $X_i$  is the exam score,  $c = 7$  is the cutoff used for treatment assignment, and  $D_i = \mathbb{I}(X_i \geq 7)$  is the binary variable that indicates receipt of the scholarship.

Our goal is to assess the effect of the binary treatment  $D_i$  on a certain outcome variable. For instance, in the previous scholarship example, we may be interested in analyzing whether the scholarship increases the academic performance during college or the probability of graduating. This problem can be formalized within the potential outcomes framework (Imbens and Rubin, 2015). In this framework, each unit  $i$  from the population of interest has two potential outcomes, denoted  $Y_i(1)$  and  $Y_i(0)$ , which measure the outcome that would be observed for unit  $i$  with and without treatment, respectively. For example, for a certain college student  $i$ ,  $Y_i(1)$  could be the student's GPA at a certain stage had she received the scholarship and  $Y_i(0)$  the student's GPA had she not received the scholarship. The individual-level treatment 'effect' for unit  $i$  is defined as the difference between the potential outcomes under treatment and control status,  $\tau_i = Y_i(1) - Y_i(0)$ .

Because the same unit can never be observed under both treated and control status (a student can either receive or not receive the scholarship, not both), one of the potential outcomes is always unobservable. The observed outcome, denoted  $Y_i$  equals  $Y_i(1)$  when  $i$  is treated and  $Y_i(0)$  if  $i$  is untreated, that is,

$$Y_i = Y_i(1)D_i + Y_i(0)(1 - D_i) \\ = \begin{cases} Y_i(1) & \text{if } D_i = 1 \\ Y_i(0) & \text{if } D_i = 0 \end{cases} \quad (2)$$

The observed outcome can never provide information on both potential outcomes. Hence, for each unit in the population, one of the potential outcomes is observed and the other one is a *counterfactual*. This problem is known as the *fundamental problem of causal inference* (Holland, 1986).

The RD design provides a way to address this problem by comparing treated units that are 'slightly above' the cutoff to control units that are 'slightly below' it. The rationale behind this comparison is that under appropriate assumptions, which will be made more precise in the upcoming sections, treated and control units in a small neighborhood or *window* around the cutoff are comparable in the sense of having similar observed and unobserved characteristics (with the only exception being treatment status). Thus, observing the outcomes of units just below the cutoff provides a valid measure of the average outcome that treated units just above the cutoff would have had if they had not received the treatment.

In the remainder of this chapter, we describe two alternative approaches for analyzing RD designs. The first one, which we call the *continuity-based framework*, assumes that the observed sample is a random draw from an infinite population of interest and invokes assumptions of continuity. In this framework, identification of the parameter of interest (defined precisely in the next section) relies on assuming that the average potential outcomes given the score are continuous as a function of the score. This assumption implies that the researcher can compare units marginally above the cutoff to units marginally below in order to identify (and estimate) the average treatment effect at the cutoff.

The second approach for RD analysis, which we call the *local randomization framework*, assumes that the treatment of interest is as-if randomly assigned in a small region around the cutoff. This approach formalizes the interpretation of RD designs as local experiments and allows researchers to use the standard tools from the classical analysis of experiments. In addition, if the researcher is willing to assume that potential outcomes are fixed (non-random) and that the  $n$  units observed in the sample form the finite population of interest, this approach also allows the researcher to use finite-sample exact randomization inference tools, which are especially

appealing in applications where the number of observations near the cutoff is small.

For both frameworks, we discuss the parameters of interest, estimation, inference, and bandwidth or window selection methods. We then compare the two approaches and provide a series of falsification methods that are commonly employed to assess the validity of the RD design. See also Cattaneo et al. (2017) for an overview and practical comparisons between these RD approaches.

## THE CONTINUITY-BASED FRAMEWORK

Under the continuity-based framework, the observed data  $\{Y_i(1), Y_i(0), X_i, D_i\}$ , for  $i = 1, 2, \dots, n$ , is a random sample from an infinite population of interest (or data-generating process). The main objects of interest under this framework are the conditional-expectation functions of the potential outcomes,

$$\begin{aligned} \mu_1(x) &= \mathbb{E}[Y_i(1)|X_i = x] \\ \text{and } \mu_0(x) &= \mathbb{E}[Y_i(0)|X_i = x] \end{aligned} \quad (3)$$

which capture the population average of the potential outcomes for each value of the score. In the sharp RD design, for each value of  $x$ , only one of these functions is observed:  $\mu_1(x)$  is observed for  $x$  at or above the cutoff and  $\mu_0(x)$  is observed for values of  $x$  below the cutoff.

The observed conditional expectation function is

$$\mu(x) = \mathbb{E}[Y_i|X_i = x] = \begin{cases} \mu_1(x) & \text{if } x \geq c \\ \mu_0(x) & \text{if } x < c \end{cases} \quad (4)$$

We start by defining the function  $\tau(x)$ , which gives the average treatment effect conditional on  $X_i = x$ :

$$\begin{aligned} \tau(x) &= \mathbb{E}[Y_i(1) - Y_i(0)|X_i = x] \\ &= \mu_1(x) - \mu_0(x) \end{aligned} \quad (5)$$

The first step is to establish conditions for identification, that is, conditions under which we can write the parameter of interest, which depends on unobservable quantities due to the fundamental problem of causal inference, in terms of observable (i.e., identifiable) and thus estimable quantities. In the continuity-based framework, the key assumption for identification is that  $\mu_1(x)$  and  $\mu_0(x)$  are continuous functions of the score at the cutoff point  $x = c$ . Intuitively and informally, this condition states that the observable and unobservable characteristics that determine the average potential outcomes do not jump abruptly at the cutoff. When this assumption holds, the only difference between units on opposite sides of the cutoff whose scores are ‘very close’ to the cutoff is their treatment status.

Intuitively, we may think that treated and control units with very different score values will generally be very different in terms of important observable and unobservable characteristics affecting the outcome of interest, but as their scores approach the cutoff and become similar in that dimension, the only remaining difference between them will be their treatment status, thus ensuring comparability between units just above and just below the cutoff, at least in terms of their potential outcome mean regression functions.

More formally, Hahn et al. (2001) showed that when conditional expectation functions are continuous in  $x$  at the cutoff level  $x = c$ ,

$$\tau(c) = \lim_{x \downarrow c} \mathbb{E}[Y_i|X_i = x] - \lim_{x \uparrow c} \mathbb{E}[Y_i|X_i = x] \quad (6)$$

that is, the difference between average observed outcomes for units just above and just below the cutoff is equal to the average treatment effect at the cutoff,  $\tau(c) = \mathbb{E}[Y_i(1) - Y_i(0)|X_i = c]$ . Note that this identification result expresses the estimand  $\tau(c)$ , which is unobservable, as a function of two limits that depend only on observable (i.e., identifiable) quantities that are estimable from the data.

As a consequence, in a sharp RD design, a natural parameter of interest is  $\tau(c)$ , the

average treatment effect at the cutoff. This parameter captures the average effect of the treatment on the outcome of interest, given that the value of the score is equal to the cutoff. It is useful to compare this parameter with the average treatment effect,  $ATE = \mathbb{E}[Y_i(1) - Y_i(0)]$ , which is the difference that we would see in average outcomes if all units were switched from control to treatment. In contrast to ATE, which is the average of  $\tau(x)$  over  $x$ ,  $ATE = \mathbb{E}[\tau(X_i)]$ ,  $\tau(c)$  is only the average effect of the treatment at a particular value of the score,  $x = c$ . For this reason, the RD parameter of interest  $\tau(c)$  is often referred to as a *local* average treatment effect, because it is only informative of the effect of the treatment for units whose value of the score is at (or, loosely speaking, in a local neighborhood of) the cutoff. This limits the external validity of the RD parameter  $\tau(c)$ . A recent and growing literature studies how to extrapolate treatment effects in RD designs (Angrist and Rokkanen, 2015; Dong and Lewbel, 2015; Cattaneo et al., 2016a; Bertanha and Imbens, 2019; Cattaneo et al., 2020c).

The main advantage of the identification result in equation (6) is that it relies on continuity conditions of  $\mu_1(x)$  and  $\mu_0(x)$  at  $x = c$ , which are nonparametric in nature and reasonable in a wide array of empirical applications. The fifth section describes several falsification strategies to provide indirect empirical evidence in order to assess the plausibility of this assumption. Assuming continuity holds, the estimation of the RD parameter  $\tau(c)$  can proceed without making parametric assumptions about the particular form of  $\mathbb{E}[Y_i|X_i = c]$ . Instead, estimation can proceed by using nonparametric methods to approximate the regression function  $\mathbb{E}[Y_i|X_i = x]$  separately for values of  $x$  above and below the cutoff.

However, estimation and inference via nonparametric local approximations near the cutoff is not without challenges. When the score is continuous, there are, in general, no units that have a value of the score exactly equal to the cutoff. Thus, estimation

of the limits of  $\mathbb{E}[Y_i|X_i = x]$  will necessarily require extrapolation as  $x$  tends to the cutoff from above or below. To this end, estimation in RD designs requires specifying a neighborhood or bandwidth around the cutoff in which to approximate the regression function  $\mathbb{E}[Y_i|X_i = x]$ , and then, based on that approximation, calculate the value that the function has exactly at  $x = c$ . In what follows, we describe different methods for estimation and bandwidth selection under the continuity-based framework.

### **Bandwidth Selection**

Selecting the bandwidth around the cutoff in which to estimate the effect is a crucial step in RD analysis, as the results and conclusions are typically sensitive to this choice. We now briefly outline some common methods for bandwidth selection in RD designs. See also Cattaneo and Vazquez-Bare (2016) for an overview of neighborhood selection methods in RD designs.

The approach for bandwidth selection used in early RD studies is what we call *ad hoc* bandwidth selection, in which the researcher chooses a bandwidth without a systematic data-driven criterion, perhaps relying on intuition or prior knowledge about the particular context. This approach is not recommended, since it lacks objectivity, does not have a rigorous justification, and, by leaving bandwidth selection to the discretion of the researcher, opens the door for specification searches. For these reasons, the ad hoc approach to bandwidth selection has been replaced by systematic, data-driven criteria.

In the RD continuity-based framework, the most widely used bandwidth selection criterion in empirical practice is the *mean squared error* (MSE) criterion (Imbens and Kalyanaraman, 2012; Calonico et al., 2014b; Arai and Ichimura, 2018; Calonico et al., 2019b), which relies on a tradeoff between the bias and variance of the RD point estimator.



The bandwidth determines the neighborhood of observations around the cutoff that will be used to approximate the unknown function  $\mathbb{E}[Y_i|X_i = x]$  above and below the cutoff. Intuitively, choosing a very small bandwidth around the cutoff will tend to reduce the misspecification error in the approximation, thus reducing bias. A very small bandwidth, however, requires discarding a large fraction of the observations and hence reduces the sample, leading to estimators with larger variance. Conversely, choosing a very large bandwidth allows the researcher to gain precision using more observations for estimation and inference, but it is at the expense of a larger misspecification error, since the function  $\mathbb{E}[Y_i|X_i = x]$  now has to be approximated over a larger range. The goal of bandwidth selection methods based on this tradeoff is therefore to find the bandwidth that optimally balances bias and variance.

We let  $\hat{\tau}$  denote a local polynomial estimator of the RD treatment effect  $\tau(c)$  – we explain how to construct this estimator in the next section. For a given bandwidth  $h$  and a total sample size  $n$ , the MSE of  $\hat{\tau}$  is

$$MSE(\hat{\tau}) = Bias^2(\hat{\tau}) + Variance(\hat{\tau}) \tag{7}$$

$$= B^2 + V,$$

which is the sum of the squared bias and the variance of the estimator. The MSE-optimal bandwidth,  $h_{MSE}$ , is the value of  $h$  that balances bias and variance by minimizing the MSE of  $\hat{\tau}$ ,

$$h_{MSE} \approx \arg \min_{h>0} MSE(\hat{\tau}) \tag{8}$$

The shape of the MSE depends on the specific estimator chosen. For example, when  $\hat{\tau}$  is obtained using local linear regression (LLR), which will be discussed in the next section, the MSE can be approximated by

$$MSE(\hat{\tau}) \approx h^4 B^2 + \frac{1}{nh} V$$

where  $B$  and  $V$  are constants that depend on the data-generating process and specific features of the estimator used. This expression clearly highlights how a smaller bandwidth reduces the bias term while increasing the variance and vice versa. In this case, the optimal bandwidth, simply obtained by setting the derivative of the above expression with respect to  $h$  equal to zero, is

$$h_{MSE}^{LLR} = C_{MSE} n^{-1/5}, \tag{9}$$

where the constant  $C_{MSE} = \left(\frac{V}{4B^2}\right)^{1/5}$  is unknown but estimable. This shows that the MSE-optimal bandwidth for a local linear estimator is proportional to  $n^{-1/5}$ .

While  $h_{MSE}$  is optimal for point estimation, it is generally not optimal for conducting inference. Calonico et al. (2018, 2019a, 2019b, 2020) show how to choose the bandwidth to obtain confidence intervals minimizing the coverage error probability (CER). More precisely, let  $CI(\hat{\tau})$  be an  $(1-\alpha)$ -level confidence interval for the RD parameter  $\tau(c)$ , based on the estimator  $\hat{\tau}$ . A CER-optimal bandwidth makes the coverage probability as close as possible to the desired level  $1 - \alpha$ :

$$h_{CER} \approx \arg \min_{h>0} \left| \mathbb{P}([\tau(c) \in CI(\hat{\tau})] - (1 - \alpha)) \right| \tag{10}$$

For the case of local linear regression, the CER-optimal  $h$  is

$$h_{CER}^{LLR} = C_{CER} n^{-1/4}$$

where, again, the constant  $C_{CER}$  is unknown, because it depends in part on the data-generating process, but it is estimable. Hence, the CER-optimal bandwidth is smaller than the MSE-optimal bandwidth, at least in large samples.

Based on the ideas above, several variations of optimal bandwidth selectors exist, including one-sided CER-optimal and MSE-optimal bandwidths with and without

accounting for covariate adjustment, clustering, or other specific features. In all cases, these bandwidth selectors are implemented in two steps: first, the constant (e.g.,  $C_{MSE}$  or  $C_{CER}$ ) is estimated, and then the bandwidth is chosen using that preliminary estimate and the appropriate rate formula (e.g.,  $n^{-1/5}$  or  $n^{-1/4}$ ).

### Estimation and Inference

Given a bandwidth  $h$ , continuity-based estimation in RD designs consists on estimating the outcome regression functions, given the score, separately for treated and control units whose scores are within the bandwidth. Recall from equation (6) that we need to estimate the limits of the conditional expectation function of the observed outcome from the right and from the left.

One possible approach would be to simply estimate the difference in average outcomes between treated and controls within  $h$ . This strategy is equivalent to fitting a regression that includes only an intercept at each side of the cutoff. However, since the goal is to estimate two boundary points, this local constant approach will have a bias that can be reduced by including a slope term in the regression. More generally, the most common approach for point estimation in the continuity-based RD framework is to employ *local polynomial* methods (Fan and Gijbels, 1996), which involve fitting a polynomial of order  $p$  separately on each side of the cutoff, only for observations inside the bandwidth. Local polynomial approximations usually include a weighting scheme that places more weight on observations that are closer to the cutoff; this weighting scheme is based on a *kernel function*, which we denote by  $K(\cdot)$ .

More formally, the treatment effect is estimated as:

$$\hat{\tau} = \hat{\alpha}_+ - \hat{\alpha}_-$$

where  $\hat{\alpha}_+$  is obtained as the intercept from the (possibly misspecified) regression model:

$$Y_i = \alpha_+ + \beta_{1+}(X_i - c) + \dots + \beta_{p+}(X_i - c)^p + u_i$$

on the treated observations using weights  $K((X_i - c)/h)$ , and similarly  $\hat{\alpha}_-$  is obtained as the intercept from an analogous regression fit employing only the control observations. Although theoretically a large value of  $p$  can capture more features of the unobserved regression functions,  $\mu_1(x)$  and  $\mu_0(x)$ , in practice, high-order polynomials can have erratic behavior, especially when estimating boundary points, a fact usually known as *Runge's phenomenon* (Calonico et al., 2015a: 1756–7). In addition, global polynomials can lead to counter-intuitive weighting schemes, as discussed by Gelman and Imbens (2019). Common choices for  $p$  are  $p = 1$  or  $p = 2$ .

As we can see, once the bandwidth has been appropriately chosen, the implementation of local polynomial regression reduces to simply fitting two linear or quadratic regressions via weighted least-squares – see Cattaneo et al. (2019a) for an extended discussion and practical introduction. Despite the implementation and algebraic similarities between ordinary least squares (OLS) methods and local polynomial methods, there is a crucial difference: OLS methods assume that the polynomial used for estimation is the true form of the function, while local polynomial methods see it as just an approximation to an unknown regression function. Thus, inherent in the use of local polynomial methods is the idea that the resulting estimate will contain a certain error of approximation or misspecification bias.

This difference between OLS and local polynomial methods turns out to be very consequential for inference purposes – that is, for testing statistical hypotheses and constructing confidence intervals. The conventional OLS inference procedure to test the null hypothesis of no treatment effect at the cutoff,  $H_0: \tau(c) = 0$ , relies on the assumption that the

distribution of the t-statistic is approximately standard normal in large samples:

$$\frac{\hat{\tau}}{\sqrt{\mathcal{V}}} \stackrel{a}{\sim} N(0,1) \quad (11)$$

where  $\mathcal{V}$  is the (conditional) variance of  $\hat{\tau}$ , that is, the square of the standard error.

However, this will only occur in cases where the misspecification bias or approximation error of the estimator  $\hat{\tau}$ , for  $\tau(c)$  becomes sufficiently small in large samples, so that the distribution of the t-statistic is correctly centered at zero. In general, this will not occur in RD analysis, where the local polynomials are used as a non-parametric approximation device and do not make any specific functional form assumptions about the regression functions  $\mu_1(x)$  and  $\mu_0(x)$ , which will generally be misspecified. The general approximation to the t-statistic in the presence of misspecification error is

$$\frac{\hat{\tau} - \mathcal{B}}{\sqrt{\mathcal{V}}} \stackrel{a}{\sim} N(0,1) \quad (12)$$

where  $\mathcal{B}$  is the (conditional) bias of  $\hat{\tau}$  for  $\tau(c)$ . This approximation will be equivalent to the one in equation (11) only when  $\mathcal{B} / \sqrt{\mathcal{V}}$  is small, at least in large samples.

More generally, it is crucial to account for the bias  $\mathcal{B}$  when conducting inference. The magnitude of the bias depends on the shape of the true regression functions and on the length of the bandwidth. As discussed before, the smaller the bandwidth, the smaller the bias. Although the conventional asymptotic approximation in equation (11) will be valid in some special cases, such as when the bandwidth is small enough, it is not valid in general. In particular, if the bandwidth chosen for implementation is the MSE-optimal bandwidth discussed in the prior section, the bias will remain even in large samples, making inferences based on equation (11) invalid. In other words, the

MSE-optimal bandwidth, which is optimal for point estimation, is too large when conducting inference, according to the usual OLS approximations.

Generally valid inferences thus require researchers to use the asymptotic approximation in equation (12), which contains the bias. In particular, Calonico et al. (2014b) propose a way to construct a t-statistic that corrects the bias of the estimator (thus making the approximation valid for more bandwidth choices, including the MSE-optimal choice) and simultaneously adjusts the standard errors to account for the variability that is introduced in the bias-correction step – this additional variability is introduced because the bias is unknown and thus must be estimated. This approach is known as *robust bias-corrected* inference.

Based on the approximation (12), Calonico et al. (2014b) propose robust bias-corrected confidence intervals

$$CI_{rbc} = \left[ \hat{\tau} - \hat{\mathcal{B}} \pm 1.96 \sqrt{\mathcal{V}_{bc}} \right], \quad (13)$$

where, in general,  $\mathcal{V}_{bc} > \mathcal{V}$  because  $\mathcal{V}_{bc}$  includes the variability of estimating  $\mathcal{B}$  with  $\hat{\mathcal{B}}$ . In terms of implementation, the infeasible variance  $\mathcal{V}_{bc}$  can be replaced by a consistent estimator  $\hat{\mathcal{V}}_{bc}$ , which can account for heteroskedasticity and clustering as appropriate.

Robust bias-correction methods for RD designs have been further developed in recent years. For example, see Calonico et al. (2019b) for robust bias-correction inference in the context of RD designs with covariate adjustments, clustered data, and other empirically relevant features. In addition, see Calonico et al. (2018, 2019a, 2019b, 2020) for theoretical results justifying some of features of robust bias-correction inference. Finally, see Ganong and Jäger (2018) and Hyytinen et al. (2018) for two recent applications and empirical comparisons of robust bias-correction methods.

### Continuity-based framework: summary

#### 1 Key assumptions:

- a) Random potential outcomes drawn from an infinite population
- b) The regression functions are continuous at the cutoff

#### 2 Bandwidth selection:

- a) Systematic, data-driven selection based on nonparametric methods
- b) Optimality criteria: MSE, coverage error

#### 3 Estimation:

- a) Nonparametric local polynomial regression within bandwidth
- b) Choice parameters: order of the polynomial, weighting method (kernel)

#### 4 Inference:

- a) Large-sample normal approximation
- b) Robust, bias corrected

## THE LOCAL RANDOMIZATION FRAMEWORK

The local randomization approach to RD analysis provides an alternative to the continuity-based framework. Instead of relying on assumptions about the continuity of regression functions and their approximation and extrapolation, this approach is based on the idea that close enough to the cutoff, the treatment can be interpreted to be ‘as good as randomly assigned’. The intuition is that if units either have no knowledge of the cutoff or have no ability to precisely manipulate their own score, units whose scores are close enough to the cutoff will have the same chance of being barely above the cutoff as barely below it. If this is true, then close enough to the cutoff, the RD design may create experimental-like variation in treatment assignment. The idea that RD designs create conditions that resemble an experiment near the cutoff has been present since the origins of the method (see Thistlethwaite and Campbell, 1960) and sometimes has been proposed as a heuristic interpretation of continuity-based RD results.

Cattaneo et al. (2015) used this local randomization idea to develop a formal framework and to derive alternative assumptions for the analysis of RD designs, which are stronger than the typical continuity conditions. The formal local randomization framework was further developed by Cattaneo et al. (2017). The central idea behind the local randomization approach is to assume the existence of a neighborhood or window around the cutoff where the assignment to above or below the cutoff behaves as it would have behaved in an experiment that assigned all units to treatment with equal (and known) probability.

The formalization of these assumptions requires a more general notation. In prior sections, we used  $Y_i(D_i)$  to denote the potential outcome under treatment  $D_i$ , which could be equal to one (treatment) or zero (control). Since  $D_i = \mathbb{I}(X_i \geq c)$ , this also allowed the score  $X_i$  to indirectly affect the potential outcomes; moreover, this notation did not prevent  $Y_i(\cdot)$  from being a function of  $X_i$ ; but this was not explicitly noted. We now generalize the notation to explicitly note that the potential outcomes may be a direct function of  $X_i$ , so we

write  $Y_i(D_i, X_i)$ . In addition, note that here and in all prior sections, we are implicitly assuming that potential outcomes only depend on unit  $i$ 's own treatment assignment and running variable, an assumption known as SUTVA (stable unit treatment value assumption). While some of the methods described in this section are robust to some violations of the SUTVA, we impose this assumption to ease exposition. See Cattaneo et al. (2017) for more discussion.

To formalize the local randomization RD approach, we assume that there exists a window  $W_0$  around the cutoff where the following two conditions hold:

- **Unconfounded assignment.** The distribution function of the score inside the window,  $F_{X_i | X_i \in W_0}(r)$ , does not depend on the potential outcomes, is the same for all units, and is known:

$$F_{X_i | X_i \in W_0}(x) = F_0(x), \quad (14)$$

where  $F_0(x)$  is a known distribution function.

- **Exclusion restriction.** The potential outcomes do not depend on the value of the running variable inside the window, except via the treatment assignment indicator

$$Y_i(d, x) = Y_i(d), \quad \forall i \quad (15)$$

*such that*  $X_i \in W_0$ .

This condition requires the potential outcomes to be unrelated to the score inside the window.

Importantly, these two assumptions would *not* be satisfied by randomly assigning the value of the score (and hence the treatment) inside  $W_0$ , because the random assignment of the score inside  $W_0$  does not by itself guarantee that the score and the potential outcomes are unrelated (the exclusion restriction). For example, imagine a RD design based on elections, where the treatment is the electoral victory of a political party, the score is the vote share, and the party wins the election if the vote share is above 50%. Even in

very close races, donors might still believe that districts where the party obtained a bare majority are more likely to support the party again, and thus they may donate more money to the races where the party's vote share was just above 50% than to races where the party was just below 50%. If donations are effective in boosting the party, this would induce a positive relationship near the cutoff between the running variable (vote share) and the outcome of interest (victory in the future election), even when the running variable is initially randomly assigned.

This illustrates why the unconfounded assignment assumption in equation (14) is not enough for a local randomization approach to RD analysis. In addition, we must explicitly assume that the score and the potential outcomes are unrelated inside  $W_0$ , which is not implied by equation (14). This issue is discussed in detail by Sekhon and Titiunik (2017), who use several examples to show that the exclusion restriction in equation 15 is implied neither by assuming statistical independence between the potential outcomes and the treatment in  $W_0$  nor by assuming that the running variable is randomly assigned in  $W_0$ . In addition, see Sekhon and Titiunik (2016) for a discussion of the status of RD designs among observational studies and Titiunik (2020) for a discussion of the connection between RD designs and natural experiments.

### ***Estimation and Inference within a Known Window***

The local randomization conditions (14) and (15) open new possibilities for RD estimation and inference. Of course, these conditions are strong and, just like the continuity conditions in the third section, they are not implied by the RD treatment assignment rule but rather must be assumed in addition to it (Sekhon and Titiunik, 2016). Because these assumptions are strong and are inherently untestable, it is

crucial for researchers to provide as much information as possible regarding their plausibility. We discuss this issue in the fifth section, where we present several strategies for empirical falsification of the RD assumptions.

The key assumption of the local randomization approach is that there exists a neighborhood around the cutoff in which equations (14) and (15) hold – implying that we can treat the RD design as a randomized experiment near the cutoff. We denote this neighborhood by  $W_0 = [c - w, c + w]$ , where  $c$  continues to be the RD cutoff, but we now use the notation  $w$  as opposed to  $h$  to emphasize that  $w$  will be chosen and interpreted differently from the previous section. Furthermore, to ease the exposition, we start by assuming that  $W_0$  is known and then discuss how to select  $W_0$  based on observable information. This data-driven window selection step will be crucial in applications, as in most empirical examples,  $W_0$  is fundamentally unknown, if it exists at all – but see Hyytinen et al. (2018) for an exception.

Given a window  $W_0$ , the local randomization framework summarized by assumptions (14) and (15) allows us to analyze the RD design employing the standard tools of the classical analysis of experiments. Depending on the available number of observations inside the window, the experimental analysis can follow two different approaches. In the Fisherian approach, also known as a randomization inference approach, potential outcomes are considered non-random, the assignment mechanism is assumed to be known, and this assignment is used to calculate the exact finite-sample distribution of a test statistic of interest under the null hypothesis that the treatment effect is zero for every unit. On the other hand, in the large-sample approach, the potential outcomes may be fixed or random, the assignment mechanism need not be known, and the finite-sample distribution of the test statistic is approximated under the assumption that the number of observations is large. Thus,

in contrast to the Fisherian approach, in the large-sample approach inferences are based on test statistics whose finite-sample properties are unknown but whose null distribution can be approximated by a normal distribution under the assumption that the sample size is large enough.

Next, we briefly review both Fisherian and large-sample methods for analysis of RD designs under a local randomization framework. Fisherian methods will be most useful when the number of observations near the cutoff is small, which may render large-sample methods invalid. In contrast, in applications with many observations, large-sample methods will be the most natural approach, and Fisherian methods can be used as a robustness check.

### *Fisherian Approach*

In the Fisherian framework, the potential outcomes are seen as fixed, non-random magnitudes from a finite population of  $n$  units. The information on the observed sample of units  $i = 1, \dots, n$  is not seen as a random draw from an infinite population but as the population of interest. This feature allows for the derivation of the finite-sample-exact distribution of test statistics without relying on approximations.

We follow the notation in Cattaneo et al. (2017), slightly adapting our previous notation. Let  $\mathbf{X} = (X_1, \dots, X_n)'$  denote the  $n \times 1$  column vector collecting the observed running variable of all units in the sample, and let  $\mathbf{D} = (D_1, \dots, D_n)'$  be the vector collecting treatment assignments. The non-random potential outcomes for each unit  $i$  are denoted by  $y_i(d, x)$  where  $d$  and  $x$  are possible values for  $D_i$  and  $X_i$ . All the potential outcomes are collected in the vector  $\mathbf{y} = (\mathbf{d}, \mathbf{x})$ . The vector of observed outcomes is simply the vector of potential outcomes, evaluated at the observed values of the treatment and running variable,  $\mathbf{Y} = \mathbf{y}(\mathbf{D}, \mathbf{X})$ .

Because potential outcomes are assumed non-random, all the randomness in the model

enters through the running variable vector  $\mathbf{X}$  and the treatment assignment  $\mathbf{D}$ , which is a function of it. In what follows, we let the subscript ‘0’ indicate the sub-vector inside the neighborhood  $W_0$ , so that  $\mathbf{X}_0$ ,  $\mathbf{D}_0$ , and  $\mathbf{Y}_0$  denote the vectors of running variables, treatment assignments, and observed outcomes inside  $W_0$ . Finally,  $N_0^+$  denotes the number of observations inside the neighborhood and above the cutoff (treated units inside  $W_0$ ) and  $N_0^-$  the number of units in the neighborhood below the cutoff (control units in  $W_0$ ), with  $N_0 = N_0^+ + N_0^-$ . Note that using the fixed-potential outcomes notation, the exclusion restriction becomes  $y_i(d, x) = y_i(d)$ ,  $\forall i \in W_0$  (see assumption 1(b) in Cattaneo et al., 2015).

In this Fisherian framework, a natural null hypothesis is the *sharp null of no effect*:

$$H_0^s: y_i(1) = y_i(0), \quad \forall i \in W_0.$$

This sharp null hypothesis states that switching treatment status does not affect potential outcomes, implying that the treatment does not have an effect on *any* unit inside the window. In this context, a hypothesis is *sharp* when it allows the researcher to impute all the missing potential outcomes. Thus,  $H_0^s$  is sharp because when there is no effect, all the missing potential outcomes are equal to the observed ones.

Under  $H_0^s$ , the researcher can impute all the missing potential outcomes and, since the assignment mechanism is assumed to be known, it is possible to calculate the distribution of any test statistic  $T(\mathbf{D}_0, \mathbf{Y}_0)$  to assess how far in the tails the observed statistic falls. This reasoning provides a way to calculate a p-value for  $H_0^s$  that is finite-sample exact and does not require any distributional approximation. This randomization inference p-value is obtained by calculating the value of  $T(\mathbf{D}_0, \mathbf{Y}_0)$  for all possible values of the treatment vector inside the window  $\mathbf{D}_0$  and calculating the probability of  $T(\mathbf{D}_0, \mathbf{Y}_0)$  being larger than

the observed value  $T_{obs}$ . See Cattaneo et al. (2015), Cattaneo et al. (2017) and Cattaneo et al. (2016b) for further details and implementation issues. See also Cattaneo et al. (2020a) for a practical introduction to local randomization methods.

In addition to testing the null hypothesis of no treatment effect, the researcher may be interested in obtaining a point estimate for the effect. When condition (15) holds, a difference in means between treated and controls inside the window,

$$\delta = \frac{1}{N_0^+} \sum_{i=1}^n Y_i D_i - \frac{1}{N_0^-} \sum_{i=1}^n Y_i (1 - D_i),$$

where the sum runs over all observations inside  $W_0$ , is unbiased for the sample average treatment effect in  $W_0$ ,

$$\tau_0 = \frac{1}{N_0} \sum_{i=1}^n (y_i(1) - y_i(0)).$$

However, it is important to emphasize that the randomization inference method described above cannot test hypotheses on  $\tau_0$ , because the null hypothesis that  $\tau_0 = 0$  is not sharp – that is, does not allow the researcher to unequivocally impute all the missing potential outcomes without further restrictive assumptions, which is a necessary condition to use Fisherian methods. Hence, under the assumptions imposed so far, hypothesis testing on  $\tau_0$  has to be based on asymptotic approximations, as described in the next section, on large-sample approaches.

The assumption that the potential outcomes do not depend on the running variable, stated in equation (15), can be relaxed by assuming a local parametric model for the relationship between  $\mathbf{Y}_0$  and  $\mathbf{X}_0$ . Specifically, Cattaneo et al. (2017) assume there exists a transformation  $\varphi(\cdot)$  such that the transformed outcomes do not depend on  $\mathbf{X}_0$ . This transformation could be, for instance, a linear adjustment that

removes the slope whenever the relationship between outcomes and the running variable is assumed to be linear. The case where potential outcomes do not depend on the running variable is a particular case in which  $\varphi(\cdot)$  is the identity function. Both inference and estimation can therefore be conducted using the transformed outcomes when the assumption that potential outcomes are unrelated is not reasonable or as a robustness check.

### *Large-Sample Approach*

In the most common large-sample approach, we treat potential outcomes as random variables and often see the units in the study as a random sample from a larger population (though in the Neyman large-sample approach, potential outcomes are fixed; see Imbens and Rubin (2015) for more discussion). In addition to the randomness of the potential outcomes, this approach differs from the Fisherian approach in its null hypothesis of interest. Given the randomness of the potential outcomes, the focus is no longer on the sharp null but rather typically on the hypothesis that the average treatment effect is zero. In our RD context, this null hypothesis can be written as

$$H_0^s : \mathbb{E}[Y_i(1)] = \mathbb{E}[Y_i(0)], \quad \forall_i \in W_0$$

Inference in this case is based on the usual large-sample methods for the analysis of experiments, relying on usual difference-in-means tests and normal-based confidence intervals. See Imbens and Rubin (2015) and Cattaneo et al. (2020a) for details.

### *Window Selection*

In practice, the window  $W_0$  in which the RD design can be seen as a randomized experiment is not known and needs to be estimated. Cattaneo et al. (2015) propose a window

selection mechanism based on the idea that in a randomized experiment, the distribution of observed covariates has to be equal between treated and controls. Thus, if the local assumption is plausible in any window, it should be in a window where we cannot reject the premise that the predetermined characteristics of treated and control units are, in distribution identical.

The idea of this procedure is to select a test statistic that summarizes differences in a vector of covariates between groups, such as the difference-in-means or the Kolmogorov–Smirnov statistic, and start with an initial ‘small’ window. Inside this initial window, the researcher conducts a test of the null hypothesis that covariates are balanced between treated and control groups. This can be done, for example, by assessing whether the minimum p-value from the tests of differences-in-means for each covariate is larger than some specified level or by conducting a joint test using, for instance, a Hotelling statistic. If the null hypothesis is not rejected, enlarge the window and repeat the process. The selected window will be the widest window in which the null hypothesis is not rejected. Common choices for the test statistic  $T(\mathbf{D}_0, \mathbf{Y}_0)$  are the difference-in-means between treated and controls, the two-sample Kolmogorov–Smirnov statistic, and the rank sum statistic. The minimum window to start the procedure should contain enough observations to ensure enough statistical power to reject the null hypothesis of covariate balance. The appropriate minimum number of observations will naturally depend on unknown, application-specific parameters, but based on standard power calculations, we suggest using no fewer than approximately 10 observations in each group.

See Cattaneo et al. (2015) and Cattaneo et al. (2017) for methodological details, Cattaneo, et al. (2020a) for a practical introduction, and Cattaneo et al. (2016b) for software implementation.



### Local randomization framework: summary

#### 1 Key assumptions:

- a) There exists a window  $W_0$  in which the treatment assignment mechanism satisfies two conditions:
- Probability of receiving a particular score value in  $W_0$  does not depend on the potential outcomes and is the same for all units
  - Exclusion restriction or parametric relationship between  $\mathbf{Y}$  and  $\mathbf{X}$  in  $W_0$

#### 2 Window selection:

- a) Goal: find a window where the key assumptions are plausible  
 b) Iterative procedure to balance observed covariates between groups  
 c) Choice parameters: test statistic, stopping rule

#### 3 Estimation:

- a) Difference in means between treated and controls within neighborhood, OR  
 b) Flexible parametric modeling to account for the effect of  $X_i$

#### 4 Inference:

- a) Fisherian randomization-based inference or large-sample inference  
 b) Conditional on sample and chosen window  
 c) Choice parameter: test statistic, randomization mechanism in Fisherian

## FALSIFICATION METHODS

Every time researchers use an RD design, they must rely on identification assumptions that are fundamentally untestable and that do not hold by construction. On one hand, if we employ a continuity-based approach, we must assume that the regression functions are smooth functions of the score at the cutoff. If, on the other hand, we employ a local randomization approach, we must assume that there exists a window where the treatment behaves as if it had been randomly assigned. These assumptions may be violated for many reasons. Thus, it is crucial for researchers to provide as much empirical evidence as possible about its validity.

Although testing the assumptions directly is not possible, there are several empirical regularities that we expect to hold in most cases where the assumptions are met. We discuss some of these tests below. Our discussion is brief, but we refer the reader to

Cattaneo et al. (2019a) for an extensive practical discussion of RD falsification methods and additional references.

- 1 **Covariate balance.** If either the continuity or local randomization assumptions hold, the treatment should not have an effect on any predetermined covariates, that is, on covariates whose values are realized before the treatment is assigned. Since the treatment effect on predetermined covariates is zero by construction, consistent evidence of non-zero effects on covariates that are likely to be confounders would raise questions about the validity of the RD assumptions. For implementation, researchers should analyze each covariate as if it were an outcome. In the continuity-based approach, this requires choosing a bandwidth and performing local polynomial estimation and inference within that bandwidth. Note that the optimal bandwidth is naturally different for each covariate. In the local randomization approach, the null hypothesis of no effect should be tested for each covariate using the same choices as used for the outcome. If the window is chosen

using the covariate-balance procedure discussed above, the selected window will automatically be a region where no treatment effects on covariates are found.

- 2 **Density of running variable.** Another common falsification test is to study the number of observations near the cutoff. If units cannot manipulate precisely the value of the score that they receive, we should expect as many observations just above the cutoff as just below it. In contrast, if units had the power to affect their score and they knew that the treatment were very beneficial, for example, we should expect more people just above the cutoff (where the treatment is received) than below it. In the continuity-based framework, the procedure is to test the null hypothesis that the density of the running variable is continuous at the cutoff (McCrary, 2008), which can be implemented in a more robust way via the novel density estimator proposed in Cattaneo et al. (2020b). In the local randomization framework, Cattaneo et al. (2017) propose a novel implementation via a finite-sample exact binomial test of the null hypothesis that the number of treated and control observations in the chosen window is compatible with a 50% probability of treatment assignment.
- 3 **Alternative cutoff values.** Another falsification test estimates the treatment effect on the outcome at a cutoff value different from the actual cutoff used for the RD treatment assignment, using the same procedures that were used to estimate the effect in the actual cutoff but only using observations that share the same treatment status (all treatment observations if the artificial cutoff is above the real one or all control observations if the artificial cutoff is below the real cutoff). The idea is that no treatment effect should be found at the artificial cutoff, since the treatment status is not changing.
- 4 **Alternative bandwidth and window choices.** Another approach is to study the robustness of the results to small changes in the size of the bandwidth or window. For implementation, the main analysis is typically repeated for values of the bandwidth or window that are slightly smaller and/or larger than the values used in the main analysis. If the effects completely change or disappear for small changes in the chosen neighborhood, researchers should be cautious in interpreting their results.

## EMPIRICAL ILLUSTRATION

To illustrate all the RD methods discussed so far, we partially reanalyze the study by Klašnja and Titiunik (2017). These authors study municipal mayor elections in Brazil between 1996 and 2012, examining the effect of a party's victory in the current election on the probability that the party wins a future election for mayor in the same municipality. The unit of analysis is the municipality, the score is the party's margin of victory at election  $t$  – defined as the party's vote share minus the vote share of the party's strongest opponent, and the treatment is the party's victory at  $t$ . Their original analysis focuses on the unconditional victory of the party at  $t + 1$  as the outcome of interest. In this illustration, our outcome of interest is instead the party's margin of victory at  $t + 1$ , which is only defined for those municipalities where the incumbent party runs for reelection at  $t + 1$ . We analyze this effect for the incumbent party (defined as the party that won election  $t - 1$ , whatever this party is) in the full sample. Klašnja and Titiunik (2017) discuss the interpretation and validity issues that arise when conditioning on the party's decision to rerun, but we ignore such issues here for the purposes of illustration.

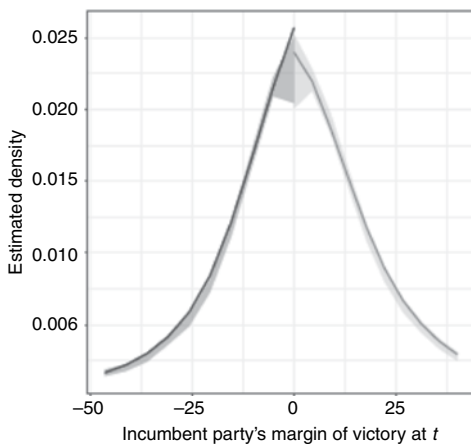
In addition to the outcome and score variables used for the main empirical analysis, our covariate-adjusted local polynomial methods, window selection procedure, and falsification approaches employ seven covariates at the municipality level: per capita GDP, population, number of effective parties, and indicators for whether each of the four parties (the Democratas, PSDB, PT, and PMDB) had won the prior ( $t - 1$ ) election.

We implement the continuity-based analysis with the `rdrobust` software (Calonico et al., 2014a, 2015b; Calonico et al., 2017), the local randomization analysis using the `rdlocand` software (Cattaneo et al., 2016b), and the density test falsification using the `rddensity` software (Cattaneo et al., 2018).

The packages can be obtained for R and Stata from <https://sites.google.com/site/rdpackages/>. We do not present the code to conserve space, but the full code employed is available in the packages' website. Cattaneo et al. (2019a, 2020b) offer a detailed tutorial on how to use these packages, employing a different empirical illustration.

### Falsification Analysis

We start by presenting a falsification analysis. In order to falsify the continuity-based analysis, we analyze the density of the running variable and also the effect of the RD treatment on several predetermined covariates. We start by reporting the result of a continuity-based density test, using the local polynomial density estimator developed by Cattaneo et al. (2019b). The estimated difference in the density of the running variable at the cutoff is  $-0.0753$ , and the p-value associated with the test of the null hypothesis that this difference is zero is 0.94. This test is illustrated in Figure 44.1, which shows the local polynomial estimated density of the incumbent party's margin of victory at  $t$  at the cutoff, separately estimated from above and below the cutoff. These results indicate



**Figure 44.1.** Estimated density of running variable

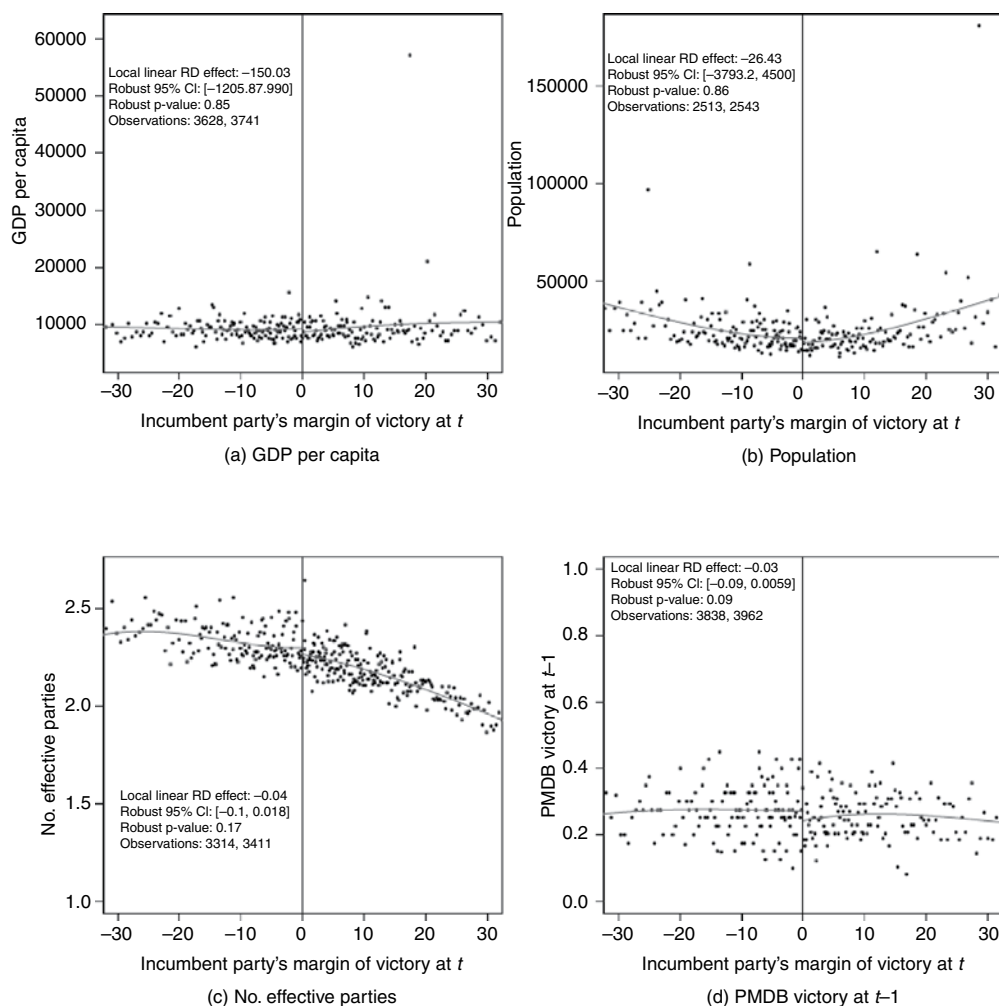
that the density of the running variable does not change abruptly at the cutoff and are thus consistent with the assumption that parties do not precisely manipulate their margin of victory to ensure a win in close races.

In addition, we also implemented the finite-sample exact binomial tests proposed in Cattaneo et al. (2017), which confirmed the empirical results obtained via local polynomial density methods. We do not report these numerical results, in order to conserve space, but they can be consulted using the accompanying replication files.

We also present local polynomial point estimates of the effect of the incumbent party's victory on each of the seven predetermined covariates mentioned above, and we perform robust local-polynomial inference to obtain confidence intervals and p-values for these effects. Since these covariates are all determined before the outcome of the election at  $t$  is known, the treatment effect on each of them is zero by construction. Our estimated effects and statistical inferences should therefore be consistent with these known null effects.

We present the results graphically in Figures 44.2 and 44.3 using typical RD plots (Calonico et al., 2015a), where binned means of the outcome within intervals of the score are plotted against the mid point of the score in each interval. A fourth-order polynomial, separately estimated above and below the cutoff, is superimposed to show the global shape of the regression functions. In these plots, we also report the formal local polynomial point estimate, 95% robust confidence interval, robust p-value, and number of observations within the bandwidth. The bandwidth (not reported) is chosen in each case to be MSE-optimal.

As we can see, the incumbent party's bare victory at  $t$  does not have an effect on any of the covariates. All 95% confidence intervals contain zero, most of these intervals are approximately symmetric around zero, and most point estimates are small. These results show that there are no obvious or notable



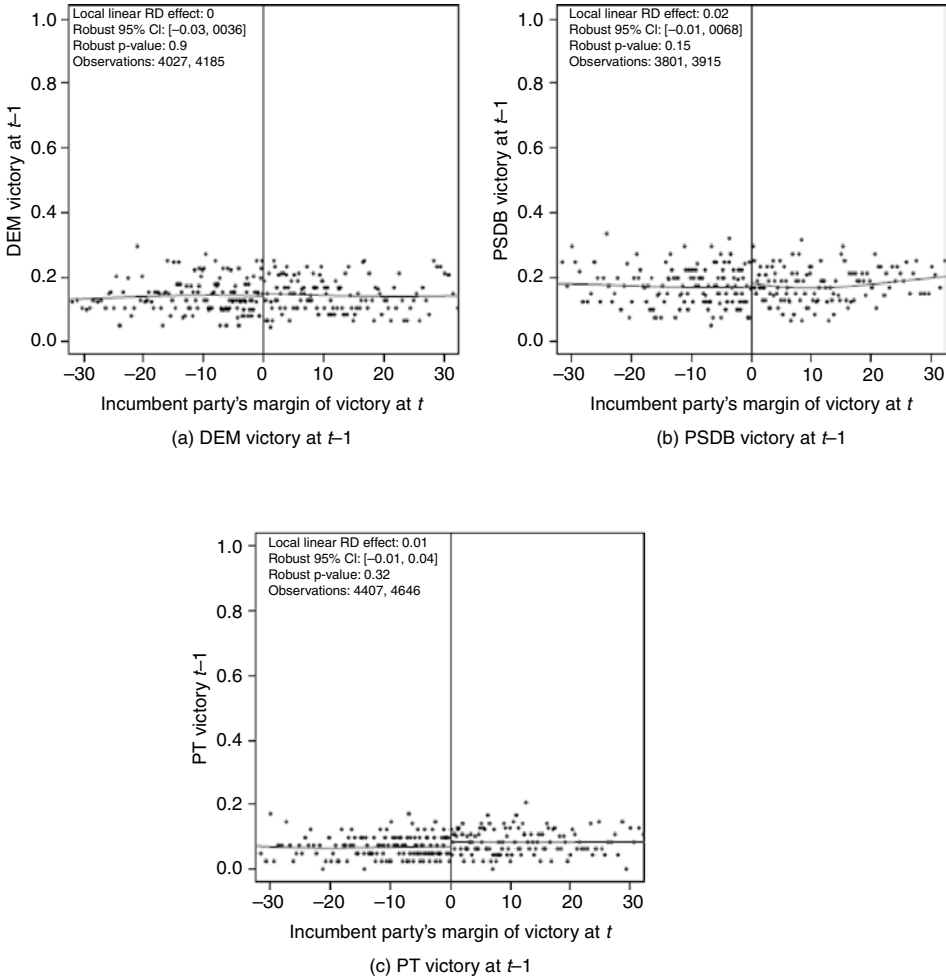
**Figure 44.2.** RD effects on predetermined covariates

covariate differences at the cutoff between municipalities where the incumbent party barely won at  $t$  and municipalities where the incumbent party barely lost at  $t$ .

### Outcome Analysis

Since the evidence from our falsification analysis is consistent with the validity of our RD design, we now proceed to analyze the treatment effect on the main outcome of

interest – the incumbent party’s margin of victory at  $t + 1$ . This effect is illustrated in Figure 44.4. A stark jump can be seen at the cutoff, where the margin of victory of the incumbent party at  $t + 1$  abruptly decreases as the score crosses the cutoff. This indicates that municipalities where the incumbent party barely wins at  $t$  obtain a lower margin of victory at election  $t + 1$  compared with municipalities where the incumbent party barely loses at  $t$ , one of the main substantive findings in Klačnjana and Titiunik (2017).

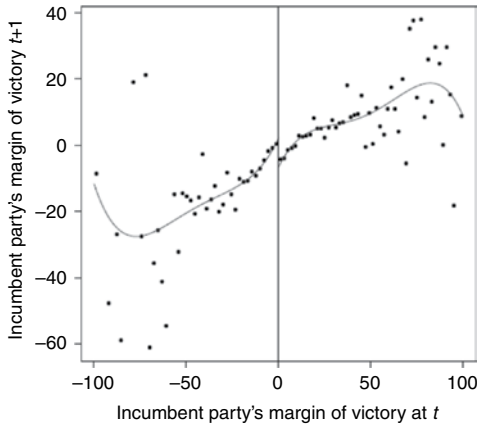


**Figure 44.3. RD effects on predetermined covariates**

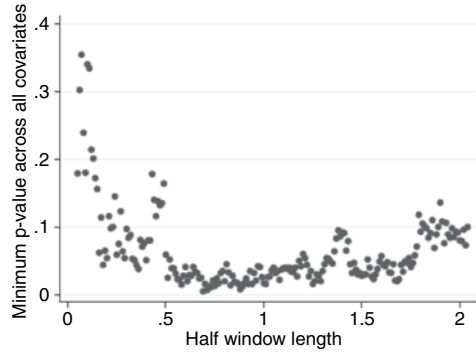
We now analyze this effect formally. We first analyze RD effects using the continuity-based framework, employing local polynomial methods with  $p = 1$  and a MSE-optimal bandwidth. For inference, we use robust bias-corrected 95% confidence intervals. As we can see in Table 44.1, the MSE-optimal bandwidth is estimated to be around 15.3 percentage points, and within this bandwidth, the RD local-polynomial point estimate is about  $-6.3$ . This shows that at the cutoff, a victory at  $t$  reduces the incumbent party's vote margin at  $t + 1$  by about six percentage points

in those municipalities where the party seeks reelection. The 95% robust bias-corrected confidence interval ranges from  $-10.224$  to  $-2.945$ , rejecting the null hypothesis of no effect with a robust p-value of about 0.0004. Including covariates leads to very similar results: the MSE-optimal bandwidth changes to 14.45 and the point estimate moves from  $-6.28$  to  $-6.10$  – a very small change, as expected when the covariates are truly predetermined.

Second, we analyze the main outcome using a local randomization approach. For



**Figure 44.4.** Effect of victory at  $t$  on vote margin at  $t + 1$ . Incumbent party, Brazilian mayoral elections, 1996–2012



**Figure 44.5** Window selector based on covariates. Incumbent party, Brazilian mayoral elections, 1996–2012. The running variable is incumbent party’s margin of victory at  $t$

**Table 44.1** Continuity-based RD analysis: effect of victory at  $t$  on vote margin at  $t + 1$ . Incumbent party, Brazilian mayoral elections, 1996–2012

	RD effect ( $\hat{\tau}$ )	95% robust conf. int.	Robust p-value	$h$	$N_l$	$N_r$
Standard	-6.281	[-10.224, -2.945]	0.0004	15.294	1533	1740
Using covariates	-6.106	[-9.881, -2.656]	0.0007	14.453	1482	1672

this, we must choose the window around the cutoff where the assumption of local randomization appears plausible (if such a window exists). We implement our window selection procedure using the list of covariates mentioned above, an increment of 0.01 percentage points, and a cutoff p-value of 0.15. We use Fisherian randomization-based inference with the difference-in-means as the test statistic and assuming a fixed-margins randomization procedure using the actual number of treated and controls in each window. As shown in Table 44.2, starting at the [0.05, -0.05] window and considering all symmetric windows in 0.01 increments, we see that all windows between [0.05, -0.05] and [0.15, -0.15] have a minimum p-value above 0.15. The window [-0.16, 0.16] is the first window where the minimum p-value drops below 0.15; indeed, it drops all the way

down to 0.061. Thus, our selected window is [-0.15, 0.15], which has exactly 38 observations on each side of the cutoff.

In order to further illustrate the results in Table 44.2, Figure 44.5 shows the associated p-values for all symmetric windows in 0.01 increments between [0.05, -0.05] and [-2.00, 2.00].

In Table 44.3, we present our inference results in the chosen window [-0.15, 0.15], reporting both Fisherian inference (using the same choices as those used in the window selection procedure) and large-sample p-values. The treated-control difference-in-means is -9.992, with a Fisherian p-value of approximately 0.083 and a large-sample p-value of about 0.070, rejecting both the sharp null hypothesis and the hypothesis of no average effect at 10% level. The fact that the point estimate continues to be negative and that

**Table 44.2 Minimum p-value in first 20 symmetric windows around cutoff running variable is vote margin at  $t$  of incumbent party, Brazilian mayoral elections, 1996–2012**

<i>Window</i>	<i>Minimum balance p-value</i>	<i>Covariate of minimum p-value</i>	$N_0^-$	$N_0^+$
[0.05,0.05]	0.179	PSDB previous victory	10	14
[0.06,0.06]	0.302	PSDB previous victory	13	16
[0.07,0.07]	0.357	No. effective parties	16	16
[0.08,0.08]	0.231	No. effective parties	18	20
[0.09,0.09]	0.176	No. effective parties	18	22
[0.10,0.10]	0.34	PT previous victory	23	28
[0.11,0.11]	0.335	Population	24	30
[0.12,0.12]	0.208	No. effective parties	26	31
[0.13,0.13]	0.201	PT previous victory	28	33
[0.14,0.14]	0.167	No. effective parties	34	36
[0.15,0.15]	0.157	No. effective parties	38	38
[0.16,0.16]	0.062	PT previous victory	42	41
[0.17,0.17]	0.114	PT previous victory	43	43
[0.18,0.18]	0.044	PT previous victory	49	45
[0.19,0.19]	0.065	PT previous victory	51	50
[0.20,0.20]	0.054	PT previous victory	53	50

**Table 44.3 Local randomization RD analysis: effect of victory at  $t$  on vote margin at  $t + 1$ . Incumbent party, Brazilian mayoral elections, 1996–2012**

<i>RD effect <math>\hat{\tau}_0</math></i>	<i>Fisher p-value</i>	<i>Large-sample p-value</i>	<i>Window</i>	$N_0^-$	$N_0^+$
-9.992	0.083	0.0697	[-0.15, 0.15]	19	20

the p-values are 8% and below suggests that the continuity-based results are broadly robust to a local randomization assumption, as both approaches lead to similar conclusions. The local randomization p-value is much larger than the p-value from the continuity-based local polynomial analysis, but this is likely due – at least in part – to the loss of observations, as the sample size goes from a total of 3,412 (1,740 + 1,672) observations to just 39 (19 + 20) (the discrepancy in the number of observations in [-0.15, 0.15] between the outcome analysis and the window-selector analysis stems from missing values in the outcome, as the margin of victory is undefined for races where the party does not run).

## CONCLUSION

We reviewed two alternative frameworks for analyzing sharp RD designs. First, the continuity-based approach, which is more common in empirical work, assumes that the unknown regression functions are continuous at the cutoff. Estimation is conducted non-parametrically using local polynomial methods, and bandwidth selection relies on minimizing a criterion such as the MSE or the coverage error probability. Inference under this framework relies on large sample distributional approximations and requires robust bias correction to account for misspecification errors local to the cutoff. Second, the local randomization approach

formalizes the intuition that RD designs can be interpreted as local experiments in a window around the cutoff. In this case, the window is chosen to ensure that treated and controls are comparable in terms of observed predetermined characteristics, as in a randomized experiment. Within this window, inference is conducted using randomization inference methods assuming that potential outcomes are non-random or other canonical analysis of experiments methods based on large-sample approximations.

These two approaches rely on different assumptions, each with its own advantages and disadvantages, and thus we see them as complementary. On the one hand, the continuity-based approach is agnostic about the data-generating process and does not require any modeling or distributional assumptions on the regression functions. This generality comes at the expense of basing inference on large-sample approximations, which may not be reliable when the sample size is small (a case that is common in RD designs, given their local nature). On the other hand, the Fisherian local randomization approach provides tools to conduct inference that is exact in finite samples and does not rely on distributional approximations. This type of inference is more reliable than large-sample-based inference when the sample size is small. And if the sample size near the cutoff is large, the analysis can also be conducted using standard large-sample methods for the analysis of experiments. However, the conclusions drawn under the local randomization approach (either Fisherian or large-sample) require stronger assumptions (unconfounded assignment, exclusion restriction) than the continuity-based approach, are conditional on a specific sample and window, and do not generalize to other samples or populations.

In sum, as in Cattaneo et al. (2017), we recommend the continuity-based approach as the default approach for analysis, since it does not require parametric modeling assumptions and automatically accounts for misspecification bias in the regression

functions when conducting estimation and inference. The local randomization approach can be used as a robustness check, especially when the sample size is small and the large-sample approximations may not be reliable.

There is one particular case, however, in which the continuity-based approach is not applicable: when the running variable exhibits only a few distinct values or mass points (even if the sample size is large because of repeated values). In this case, the nonparametric methods for estimation, inference, and bandwidth selection described above do not apply, since they are developed under the assumption of local approximations and continuity of the score variable, which are violated by construction when the running variable is discrete with a small number of mass points. Thus, in settings where the running variable has few mass points, local randomization methods – possibly employing only the closest observations to the cutoff – are a more natural approach for analysis. We refer the reader to Cattaneo et al. (2019a) for a more detailed discussion and practical illustration of this point.

## Note

- 1 We thank Rich Nielsen for his comments and suggestions on a previous version of this chapter.

## REFERENCES

- Abadie, A. and Cattaneo, M. D. (2018), 'Econometric Methods for Program Evaluation', *Annual Review of Economics*, 10, 465–503.
- Angrist, J. D. and Rokkanen, M. (2015), 'Wanna get away? Regression Discontinuity Estimation of Exam School Effects Away from the Cutoff', *Journal of the American Statistical Association*, 110, 1331–1344.
- Arai, Y. and Ichimura, H. (2018), 'Simultaneous Selection of Optimal Bandwidths for the Sharp Regression Discontinuity Estimator', *Quantitative Economics*, 9, 441–482.



- Bertanha, M. and Imbens, G. W. (2019), 'External Validity in Fuzzy Regression Discontinuity Designs', *Journal of Business & Economic Statistics*, forthcoming.
- Calonico, S., Cattaneo, M. D. and Farrell, M. H. (2018), 'On the Effect of Bias Estimation on Coverage Accuracy in Nonparametric Inference', *Journal of the American Statistical Association*, 113, 767–779.
- Calonico, S., Cattaneo, M. D. and Farrell, M. H. (2019a), 'Coverage Error Optimal Confidence Intervals for Local Polynomial Regression', *arXiv:1808.01398*.
- Calonico, S., Cattaneo, M. D. and Farrell, M. H. (2020), 'Optimal Bandwidth Choice for Robust Bias Corrected Inference in Regression Discontinuity Designs', *Econometrics Journal*, forthcoming.
- Calonico, S., Cattaneo, M. D., Farrell, M. H. and Titiunik, R. (2017), 'rdrobust: Software for Regression Discontinuity Designs', *Stata Journal*, 17, 372–404.
- Calonico, S., Cattaneo, M. D., Farrell, M. H. and Titiunik, R. (2019b), 'Regression Discontinuity Designs Using Covariates', *Review of Economics and Statistics*, 101, 442–451.
- Calonico, S., Cattaneo, M. D. and Titiunik, R. (2014a), 'Robust Data-Driven Inference in the Regression-Discontinuity Design', *Stata Journal*, 14, 909–946.
- Calonico, S., Cattaneo, M. D. and Titiunik, R. (2014b), 'Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs', *Econometrica*, 82, 2295–2326.
- Calonico, S., Cattaneo, M. D. and Titiunik, R. (2015a), 'Optimal Data-Driven Regression Discontinuity Plots', *Journal of the American Statistical Association*, 110, 1753–1769.
- Calonico, S., Cattaneo, M. D. and Titiunik, R. (2015b), 'rdrobust: An R Package for Robust Nonparametric Inference in Regression-Discontinuity Designs', *R Journal*, 7, 38–51.
- Cattaneo, M. D. and Escanciano, J. C. (2017), *Regression Discontinuity Designs: Theory and Applications (Advances in Econometrics)*, vol. 38, Emerald Group Publishing, Bingley.
- Cattaneo, M. D., Frandsen, B. and Titiunik, R. (2015), 'Randomization Inference in the Regression Discontinuity Design: An Application to Party Advantages in the U.S. Senate', *Journal of Causal Inference*, 3, 1–24.
- Cattaneo, M. D., Idrobo, N. and Titiunik, R. (2019a), *A Practical Introduction to Regression Discontinuity Designs: Foundations*, In preparation for Cambridge Elements: Quantitative and Computational Methods for Social Science, Cambridge University Press, Cambridge.
- Cattaneo, M. D., Idrobo, N. and Titiunik, R. (2020a), *A Practical Introduction to Regression Discontinuity Designs: Extensions*, Cambridge Elements: Quantitative and Computational Methods for Social Science, Cambridge University Press, Cambridge, forthcoming.
- Cattaneo, M. D., Jansson, M. and Ma, X. (2018), 'Manipulation Testing based on Density Discontinuity', *Stata Journal*, 18, 234–261.
- Cattaneo, M. D., Jansson, M. and Ma, X. (2020b), 'Simple Local Polynomial Density Estimators', *Journal of the American Statistical Association*, forthcoming.
- Cattaneo, M. D., Keele, L., Titiunik, R. and Vazquez-Bare, G. (2016a), 'Interpreting Regression Discontinuity Designs with Multiple Cutoffs', *Journal of Politics*, 78, 1229–1248.
- Cattaneo, M. D., Keele, L., Titiunik, R. and Vazquez-Bare, G. (2020c), 'Extrapolating Treatment Effects in Multi-Cutoff Regression Discontinuity Designs', *arXiv:1808.04416*.
- Cattaneo, M. D., Titiunik, R. and Vazquez-Bare, G. (2016b), 'Inference in Regression Discontinuity Designs under Local Randomization', *Stata Journal*, 16, 331–367.
- Cattaneo, M. D., Titiunik, R. and Vazquez-Bare, G. (2017), 'Comparing Inference Approaches for RD Designs: A Reexamination of the Effect of Head Start on Child Mortality', *Journal of Policy Analysis and Management*, 36, 643–681.
- Cattaneo, M. D. and Vazquez-Bare, G. (2016), 'The Choice of Neighborhood in Regression Discontinuity Designs', *Observational Studies*, 2, 134–146.
- Dong, Y. and Lewbel, A. (2015), 'Identifying the Effect of Changing the Policy Threshold in Regression Discontinuity Models', *Review of Economics and Statistics*, 97, 1081–1092.
- Fan, J. and Gijbels, I. (1996), *Local polynomial Modelling and its Applications*, vol. 66, CRC Press, Boca Raton, FL.

- Ganong, P. and Jäger, S. (2018), 'A Permutation Test for the Regression Kink Design', *Journal of the American Statistical Association*, 113, 494–504.
- Gelman, A. and Imbens, G. W. (2019), 'Why High-Order Polynomials Should Not be Used in Regression Discontinuity Designs', *Journal of Business & Economic Statistics*, 37, 447–456.
- Hahn, J., Todd, P. and van der Klaauw, W. (2001), 'Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design', *Econometrica*, 69, 201–209.
- Holland, P. W. (1986), 'Statistics and Causal Inference', *Journal of the American Statistical Association*, 81, 945–960.
- Hyytinen, A., Meriläinen, J., Saarimaa, T., Toivanen, O. and Tukiainen, J. (2018), 'When does Regression Discontinuity Design Work? Evidence from Random Election Outcomes', *Quantitative Economics*, 9, 1019–1051.
- Imbens, G. and Rubin, D. B. (2015), *Causal Inference in Statistics, Social, and Biomedical Sciences*, Cambridge University Press, Cambridge.
- Imbens, G. W. and Kalyanaraman, K. (2012), 'Optimal Bandwidth Choice for the Regression Discontinuity Estimator', *Review of Economic Studies*, 79, 933–959.
- Klašnja, M. and Titiunik, R. (2017), 'The Incumbency Curse: Weak Parties, Term Limits, and Unfulfilled Accountability', *American Political Science Review*, 111, 129–148.
- McCrary, J. (2008), 'Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test', *Journal of Econometrics*, 142, 698–714.
- Sekhon, J. S. and Titiunik, R. (2016), 'Understanding Regression Discontinuity Designs as Observational Studies', *Observational Studies*, 2, 174–182.
- Sekhon, J. S. and Titiunik, R. (2017), 'On Interpreting the Regression Discontinuity Design as a Local Experiment', in *Regression Discontinuity Designs: Theory and Applications (Advances in Econometrics)*, vol. 38, eds. M. D. Cattaneo and J. C. Escanciano, Emerald Group Publishing, Bingley, pp. 1–28.
- Thistlethwaite, D. L. and Campbell, D. T. (1960), 'Regression-discontinuity Analysis: An Alternative to the Ex-Post Facto Experiment', *Journal of Educational Psychology*, 51, 309–317.
- Titiunik, R. (2020), 'Natural Experiments', in *Advances in Experimental Political Science*, eds. J. Druckman and D. Green, Cambridge University Press, forthcoming.

# Network Analysis: Theory and Testing

Jennifer N. Victor and Elsa T. Khwaja

## INTRODUCTION

Although it is intuitive to conceptualize politics as a process that involves interactions, only within the last 20 years have political scientists had the theoretical and computational assets to fully engage an array of political questions using network-based theory and methods. Technological advances make computationally expensive network analysis more accessible to more scholars. This chapter proceeds by describing the major contributions that network analysis has made to the discipline of political science. The chapter continues with a discussion of the fundamental components of a network, which is tied to formal network theory. Then, a discussion follows of the basics of mathematical graph theory, as applicable to network analysis in politics. Next, the chapter gives practical advice about working with network data, including a section on how to visualize and represent relational data. Finally, the chapter closes with a discussion of properties of whole

networks and draws on political science examples, particularly the concepts of cohesion, reciprocity, transitivity, and centrality.

## THE ADDED VALUE OF NETWORKS IN POLITICAL SCIENCE

A diverse array of academic fields – anthropology, sociology, political science, and economics – have made fruitful applications of network theory and methods to important questions in public policy and international affairs. Social scientists from these intersecting fields have made significant contributions to the study of networks. As networks are omnipresent in social and political institutions, social network analysis is a multi-disciplinary methodology and paradigm, blending multiple methods and empirical applications that are evident in data collection and data analysis in both quantitative and qualitative techniques.

Network applications have made profound contributions to the study of US politics, from political discourse and voting, to legislative politics and political parties. Groundbreaking research from the so-called mid-20th-century Columbia school of political behavior focused attention on the relationship between mass communication and political behavior (Berelson, 1954; Lazarsfeld et al., 1948), which created the foundation on which the subfield of political behavior was built. The Michigan school developed the theoretical and empirical demonstrations of individual political behavior as being the psychological product of social context (Campbell et al., 1980; Putnam, 1966). Later, this thread of research became more explicitly relational and scholars expanded the understanding of individual political behavior by treating it as explicitly interdependent (Huckfeldt and Sprague, 1987a; Klofstad et al., 2013; McClurg, 2006; Mutz 2002; Sokhey and McClurg, 2012).

Researchers who specifically studied voting behavior then transformed a decades-long understanding of the individual motivation to vote, by applying network theory (Nickerson, 2008; Rolfe, 2013; Sinclair, 2012). In the legislative arena, scholars have long recognized the importance of social connections between lawmakers (Eulau, 1962; Patterson, 1959; Routt, 1938) but only recently had the computational power to study directly the patterns of relationships and their effect on legislative action by looking at co-sponsorship (Bratton and Rouse, 2011; Burkett and Skvoretz, 2001; Tam Cho and Fowler, 2010), committee assignments (Porter et al., 2005), campaign contributions (Kirkland, 2011; Koger and Victor, 2009; Victor and Koger, 2016), legislative staff (Montgomery and Nyhan, 2017; Ringe and Victor, 2013), and other subjects.

Beyond US politics, network methods and theories have been applied in many studies in the subfields of international relations, international development, human rights advocacy, and environmental studies, as well as to

the study of democratic peace, cooperation, and conflict. Hafner-Burton et al. (2009) write about network structure in international relations, specifically the power structures among international players. Analysis shows how state membership in international organizations creates positions of power that may sustain conflicts among states (Hafner-Burton and Montgomery, 2006). Further, Kinsella and Montgomery (2017) demonstrate the significance of networks of regional and global arms flows, lending critical observations on the study of arms supply and proliferation networks within the context of international security.

Network effects are also relevant in understanding major-party interventions on militarized interstate disputes, revealing that major parties influence other states' choices to join in conflicts (Corbetta and Dixon, 2005). Studies on terrorist or insurgency networks also contribute to critical observations about the significance of cohesion in networks. Staniland (2012) provides a qualitative account about the cohesion of insurgent groups in conflict-affected states, with a look into the dynamics of terrorist networks in Afghanistan, Kashmir, and Sri Lanka. In another study on terrorist networks, Horowitz and Potter (2013) write about how different terrorist organizations form intergroup alliances. They show how the organizations form preferential attachments with core-periphery network structures, which suggests the importance of a broader policy lens of addressing counterinsurgency tactics.

One common technique social scientists have used for identifying political networks between common citizens is the so-called 'name generator' approach, where survey respondents are asked to name their network contacts. While it is nearly impossible to identify the complete network of connections of a particular population through this method, it is highly useful for scholars who seek to engage in an 'egocentric' analysis. In such an approach, the scholar gathers information about the network of a particular

actor or set of actors and may draw inference about some outcome of interest based on the set of connections from these individualistic perspectives.

The approach was pioneered by sociologists in the 1980s (Burt, 1985) and adapted by political scientists (Huckfeldt and Sprague, 1987b). Its use has been critical to the study of social, political, and communication networks and has been adapted for application in the developing world – in particular to the spread of microfinance programs in India (Banerjee, 2012) and household level interactions in southern India (Shakya et al., 2017).

Recent policy networks studies have made significant contributions to an understanding of environmental sustainability and other topics critically related to environmental policy (Berardo and Lubell, 2016; Broadbent, 2017; Eilstrup-Sangiovanni, 2017). Broadbent's (2017) study examines a climate-change initiative that shows how differences among cases in the national policy-formation process are related to differences in their mitigation policies and carbon emissions. Additionally, with the use of policy networks, one can understand how bridging and bonding social capital are generated when governmental and non-governmental actors participate in forums on environment and climate change (Berardo and Lubell, 2016).

In part due to availability of data, scholars are benefiting from using human-generated data on social network platforms to study questions of political importance. An analysis of online blogs, social media including Facebook, Twitter, LinkedIn, and other online networks – as well as cyber-criminal networks – provide important insights on the online behavior of social media and how people engage with information. For example, a study of fake news and online engagement with misinformation is a relatively new area of research that attracts the interests of political networks analysts (Grinberg, 2018). Online networks intersect with a variety of topics from political discussion networks and engagement in political conversations on

social platforms, to social-media behaviors and cybercrimes (Klofstad et al., 2013; Lazer et al., 2015; Sokhey et al., 2015).

But even in a high-tech world with social media replacing traditional dialogue, many political scientists have increasingly incorporated fieldwork into their network-oriented research (Kapiszewski et al., 2015). One example is the application of Netmap, a tool used for understanding influence in a rural-development-program network, it uses participatory field-research methods that provide quantitative and qualitative data through low-cost, low-tech usage. (Schiffer and Hauck, 2010). It allows a low-tech, convenient way to conduct a social network analysis for a field study about all the stakeholders that can give a more sustainable solution for watershed management in the White Volta River Basin in northern Ghana. As a result, it generates a fuller understanding among stakeholders of their relationships and their level of influence (Schiffer and Hauck, 2010).

As the interdisciplinary nature of network analysis can be applied to diverse areas of study, network analysis can have great comparative value by demonstrating what types of networks matter over others (Razo, 2017; Vera and Schupp, 2006). Network analyses remain underutilized in comparative politics, where great insights are possible by blending network approaches with qualitative approaches. Some comparative politics scholars have noted the lack of network analysis and qualitative content analysis in the body of political-science literature (Fischer, 2011). Using both techniques, Fischer (2011) conducts an empirical study of Swiss political networks and compares 11 policy networks; he finds that the structure of the networks is the result of policy context, including the degree of Europeanization, federalism, intensity of the pre-parliamentary phase, and public attention.

Further, networks methods have been an increasingly relied upon tool to analyze global policy networks (Knoke and

Kostiuchenko, 2017). In recent decades, one emerging area in global public policy is the notion of ‘network governance’ (Eilstrup-Sangiovanni, 2017; Faul 2016). As the world becomes increasingly complex and interconnected, the toolbox of social scientific theories and methods to understand these complexities also expands.

One aspect of the trend is the growing literature about networks methods to analyze global and local human rights advocacy networks. The networks include local governance networks, civil-society organizations, also called CSOs, and transnational advocacy networks, or TANs (Keck and Sikkink, 1997; Murdie and Polizzi, 2017; Reinicke et al., 2000). For example, Murdie and Polizzi’s (2017) study examines how the human rights agenda, issues, and outcomes that receive international attention can be influenced by the nature of advocacy networks and their agendas.

While the studies mentioned herein are just a few examples of a much broader body of literature, they emphasize how network-based analyses give robust added value to the state of knowledge on important questions in politics, government, and policy.

## FUNDAMENTAL NETWORK COMPONENTS

The most consequential research-design choice in any network study is to identify the nodes and edges of the network. The *nodes* are the actors or units that take actions and have specific characteristics (Borgatti et al., 2018). A node typically has attributes – that is, features, characteristics, or traits. In a political network study of voting, the nodes might be individual voters, candidates, or parties. Their attributes may be age, political votes, or policy issues. A study of conflict between countries might use countries or treaty organizations as its nodes. When nodes are individual people, attributes may include demographic characteristics such as gender, age, marital status, or

profession. When nodes are institutions or coalitions, the attributes will describe characteristics of the institution or group.

*Edges* describe how they are related to one another. Edges might include a shared trait between two nodes (for example, both female, both members of the same group), or a shared action that two nodes take (for example, both nodes are parties to the same contract or both contributed to the same candidate). Defining the nodes and edges of a network study is akin to settling on the unit of analysis in a traditional statistical study. As with settling on a unit of analysis, defining the nodes and edges of the network give the study its boundaries as well as the objects through which the researcher can potentially draw inferences.

A network study usually has three potential units of analysis: node, edge, and network. When the researcher seeks to answer questions about individuals, it makes the most sense to perform analyses at the node level. If the researcher poses questions about the relationships between nodes, such as the conditions under which edges exist, the strength of relationships, or the properties of relationships, then the researcher will focus on edges as the key unit of analysis. It is relatively common in international relations to treat dyads, or pairs, as a unit of analysis. For example, in questions where a pair of countries is the defined critical unit of study, the focus becomes studying the edges of a network.

When conceptualizing relationships between nodes – factors to compare and contrast in a political analysis – there are generally five types of connections that can be described: similarities, relationships, cognition, events, and transactions. A similarity is any shared trait between two nodes, such as the same party, same state, same gender, or other. Relationship connections might describe those with familial connections, friendships, sexual partners, acquaintances, or something more hierarchical like teacher and student, boss and employee, conductor and musician. Cognitive relationships refer to connections between nodes that relate to

how they think about things, such as a shared attitude, common knowledge, a shared perception, trust or distrust. Edges can also be defined by events such as common participation in conflict, resolution, conference, convention, or other events. Edges can be described as transactions where one node transmits, sells, or communicates something to another node. In this way, networks can be used to describe the flow of information, money, or ideas across a group.

Notice that among the five types of relationships described above, they are not strictly social. For example, one may consider two members of Congress in a legislative caucus together to be connected by a common membership in a group, but this does not mean that the two legislators have a social connection to one another. Often, political scientists look at ties that are not social but either could be social or can be treated as a shared attribute for an analytical advantage, as demonstrated in the studies on campaign contributions (Koger and Victor, 2009; Victor and Koger, 2016). The downside to treating a non-social shared attribute as a relational edge in a network is that it may limit the inference one can draw on the edges. Challenges associated with causal inference and network studies are directly addressed in the next section.

Finally, scholars are frequently interested in properties found in the overall network or network component's relative positions to other components. Questions about the centrality of a political actor or set of political actors and the tendency to broker relationships between units can only be studied if one has information about the entire network. It has not been a standard approach for political scientists, but it is proving to be a form of study that brings thoughtful insights to questions of politics. Leveraging this level of analysis requires the researcher to think about their data and questions of study from a broader perspective than the one to which they are accustomed; however, as described below, many relevant politics can be studied at the level of whole networks.

## CAUSAL INFERENCE AND NETWORKS

It is important to define clearly the nodes and edges in a network analysis for the same reason it is important to define with clarity the unit of analysis in a traditional statistical analysis. Clear-cut definitions of units are the means to draw robust inferences. In other words, a claim of causal inference is constrained to the unit of study. Like traditional studies, network studies pose challenges in their pursuit of causal inference. Two particular challenges for causal inference in network research are 'homophily' and 'interference between units'.

*Homophily* refers to the extent to which characteristics in one unit are shared by connected units. A researcher who seeks to know if the presence of a characteristic is caused by a particular effect will have to wrestle with the fact that the presence of the characteristic is partially determined by the network effect. *Interference between units* refers to a spillover effect, where in an experimental setting, treatment effects spill over to adjacent, connected units because they are connected in a network. In a causal-inference framework, the researcher seeks to estimate the expected value of an observed outcome for units that have been randomly assigned a treatment.

In network studies, the assumption of random assignment might be violated due to *interference between units*, which is likely to lead to underestimating treatment effects (Rogowski and Sinclair, 2012, 2017). In general, the set of tools available to a scholar engaged in causal inference in a network study is no different than those available for a non-network study. Common approaches for identification include experiment, lagged variables, instrumental variables, and panel designs.

In observational studies of relational data, a researcher who seeks to draw inference about the effect of  $X$  on  $Y$  must be concerned about the possibility that variance in  $Y$  is determined by the dependent connection between units, rather than  $X$ . When similarities between actors cause connections

between actors, the connections themselves confound the observation of the independent and dependent variables. For example, in a study of voters, a researcher who seeks to know the effect of a particular message or campaign approach on a voter's probability of voting, or voting in a particular way, can randomize voters, administer treatment, and observe outcomes. However, if a voter changes conditions from non-voting to voting, the researcher cannot be certain whether the observed change is due to the treatment, whether it is due to the voter interacting with someone else who was treated, or because of some latent commonality between subjects. If it was due to the treatment, one would say the treatment caused the observed change. If it was due to the subject's interactions or other connections, one can say the observed difference is due to a 'peer effect', which might also be called 'contagion'. If it were due to the latent commonality, one would say that homophily explains the observed outcome. Discerning whether observations are due to peer effects, homophily, or some other treatment is a great challenge in network studies. The best practice under these conditions is to use sensitivity analysis.

A technique developed by VanderWeele (2011) allows a researcher to estimate the probability that an observed outcome is due to the presence of an unobserved factor, by estimating how sensitive outcomes are to varying levels of a latent factor. Demonstrating how sensitive a result is to an estimated bias factor reveals the extent to which observed differences are due to homophily, peer effects, or treatment. In an application of this technique to a well known study of peer effects on levels of obesity (Christakis and Fowler, 2007) and smoking cessation (Christakis and Fowler, 2008), VanderWeele (2011) finds evidence of contagion effects in obesity and smoking, which were 'reasonably' tolerant to latent factors for homophily and environmental confounders. The study also finds that the spread of happiness is not robust against such latent effects. Therefore, the conclusion

that happiness is contagious finds less support than claims of contagion effects for obesity and smoking when one uses sensitivity analysis (VanderWeele, 2011).

In experimental research designs, a network researcher's primary concern is the possibility of interference among units. That is, when units have received treatment and interacted with one another, the space of potential outcomes expands exponentially because each person in the study has had the potential to be exposed to treatment through every other person in the study. To address this problem, researchers need an understanding of the network structure among participants. When known, the pattern of connections among subjects can be used to constrain the outcome space. Individuals can be randomly assigned to treatment given a known network structure, as was done in the 61-million-person Facebook voting experiment (Bond et al., 2012). In the absence of a known network structure, a researcher can sometimes reasonably theorize and model such structure, using empirical estimates to evaluate the fit of the theorized structure (Bowers et al., 2013).

Now that the reader has a basic understanding of the levels of analysis and limits of inference in network studies, the chapter moves on to discuss the first principles behind network analysis, which come from mathematical graph theory and sociology.

## NETWORKS AND GRAPH THEORY

Graph theory is a subfield of mathematics that was developed in the 1940s. Mathematicians, such as Paul Erdos and Alfréd Rényi (1960), conceptualized nodes, edges, and graph properties by studying the distribution of edges in random graphs. Separately, sociologists began working on the 'small world problem', a phenomenon in which individual members of a large population are seemingly connected to everyone else by only a few steps. These two streams of work matured concurrently



without intersection for several decades. Mathematicians and sociologists appeared unaware of the parallel, related work advancing in other disciplines.

By the 1990s, physicists began applying graph theory to questions of particle physics. In 1999, Albert et al. published a paper on the size of the World Wide Web, drawing on the progress made in mathematics, sociology, and physics. Thus, it was just 20 years ago that scholars began to realize that networks are a natural phenomenon, present in nearly all aspects of life (Barabási, 2009; Barabási, 2014; Watts, 1999).

Today, one can use the known mathematical properties of networks to understand the relationships in everything from the human genome to the spread of disease, the distribution of oil fields across the globe, the distribution of insurance claims across beneficiaries, and participation in voting. One can use a common language to understand these phenomena because the same set of mathematical properties underlie the relationships between nodes, regardless of the identity of the nodes. In other words, when events, people, or units are relational, an observer or political scientist can analyze and understand how the relationships affect some outcomes of interest by studying the properties of the network such as its density, the distribution of dyadic and triadic connections, the number of components, its tendency to cluster, and so on.

Of course, the nature and context of connections matter in any specific instance – betweenness centrality in an arms-trade network does not imply the same phenomenon as betweenness centrality in a network that describes the spread of an infection. But some properties have consistent interpretations; for example, triadic closure (three nodes connected) is consistently associated with trust and cooperation (Borgatti et al., 2018).

The most common distribution of nodes and edges that one observes in all types of relational data is called a Pareto distribution, or a power-law distribution. Imagine a graph that describes the number of

connections – that is, edges – that each node holds in a particular population. If the nodes were independent individuals or events, one can expect the distribution to follow a ‘normal’ pattern, according to the principles of the mathematical central-limit theorem. However, a Pareto distribution is heavily skewed with a fat tail, describing that relatively few nodes have a high number of connections and many nodes have a very low number of connections. The relationship between nodes and edges is exponential.

In mathematics, a graph is an object that describes the nodes and edges of a network. For a mathematician, a network is simply a graph,  $G = \{V, E\}$ , where  $G$  stands for *graph*,  $V$  stands for *vertices*, also called nodes, and  $E$  stands for *edges*. Using the mathematical set-theory notation to describe when an object is an element of a set or group ( $\in$ ), one can describe that node  $i$  and node  $j$  are connected in graph  $G$ :  $(i, j) \in G$ . For the analysis, one can use specific terms to describe relationships within graphs using the following common language:

- When two nodes are connected, one says they are *adjacent* which means they share a tie.
- If node  $A$  and node  $B$  are connected and node  $A$  and node  $C$  are connected, one says the two edges are *incident* upon  $A$ .
- The number of edges incident on a node is its *degree*.
- A node not connected to any other nodes is an *isolate*.
- A sequence of adjacent nodes forms a *path*. Paths are unique; one cannot revisit a node more than once in a path. One might model the contagion of a virus as a path if death or immunity prevents nodes from being re-infected.
- A sequence that revisits nodes but not edges is a *trail*. One might graph the flow of gossip as a trail because gossip may come back around to the same person, or node, but from a different source, or edge.
- A sequence of adjacent nodes without restrictions on revisiting nodes or edges is a *walk*. One might graph the flow of a single dollar bill as a walk because it can travel to anyone and repeat its steps.

- The shortest path between two nodes is a *geodesic*.
- A set of nodes in which every node can reach every other node by some path is a *component*.
- An edge that connects two components that would otherwise be separated is a *bridge*.

The aforementioned terminology helps identify and describe relational data in precise ways. Doing so is a critical part of engaging in social scientific inference. If the political scientist and analyst can identify data as having dependencies but she or he does not attempt to identify and account for those dependencies in the analytical strategy, then the observers run the risk of drawing incorrect conclusions. Once a researcher determines that a set of data has network properties, or dependencies, the researcher must be deliberate about identifying the nodes and the connections between them. Many political networks contain multiple networks between nodes (referred to a *multiplex*). For example, where countries are nodes, they can be described as being connected through trade, conflict, treaty, economic investment, or any number of possible ties. It is helpful to think of each possible tie as its own network graph, so the properties of the graph can be studied; however, sometimes we seek to aggregate ties or explicitly model a multiplex.

## A SHORT PRACTICAL GUIDE TO NETWORK MATRICES

A traditional data structure is ‘rectangular’, where rows represent cases or observations, and columns represent variables or attributes of those cases. Network data is relational – it shows the relationships between nodes, observations, or cases – and is often described as ‘square’. A matrix that shows the relationships between students who take classes together, for example, might be presented as a square matrix ( $N \times N$ ), where the rows and columns are identical lists of the population

of students, and the cells represent how many classes each dyad of students have in common. The main diagonal of the matrix (where  $i = j$ ) can show the number of classes in which each individual student is enrolled. Most software for analyzing network data can read relational data in a variety of formats, so it may not be necessary for researchers to explicitly organize their data into square matrices; but much of the graph-theory mathematics that underlie network analyses assume that network data fit this basic shape.

Network data can generally be described as being ‘one-mode’ or ‘two-mode’. In *one-mode* data, the data are square; the data can be arranged as an  $N \times N$  adjacency matrix, with the same number of rows and columns. The rows and columns both refer to the same entities (nodes), and the entries in the cells describe the relationships between them. In *two-mode* data, the rows represent nodes (for example, cases, entities, individuals), and the columns represent some event, characteristic, or attribute of the nodes. A two-mode matrix, also called *bipartite*, is similar to a traditional rectangular data matrix, but the columns in a two-mode matrix refer to some common group.

For example, one might have a two-mode affiliation matrix of members of Congress (rows) and legislation (columns). The values in the cells might represent votes or bill co-sponsorship. One can convert the two-mode affiliation matrix to a one-mode adjacency matrix by multiplying a matrix by its transpose (switch the positions of the rows and columns). The following is a generic example of this procedure:

$$AA^T = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} a & c \\ b & d \end{bmatrix} \\ = \begin{bmatrix} aa + bb & ac + bd \\ ca + db & cc + dd \end{bmatrix}$$

These mathematical properties are highly convenient for working with relational or network data. For example, one can easily

calculate how many co-sponsored bills two members of congress have in common. Or, one can calculate how many co-sponsoring legislators have one or several bills in common. The summation feature of the matrix multiplication operation means that we can readily move between two-mode and one-mode data, all the while obtaining useful information about the pairs or dyads.

The following is an overly simplified example of the utility of the properties of matrix algebra for manipulating relational data. One can easily scale this for larger, more realistic data sets. Here, I have a matrix of legislators and bill co-sponsorships, called *A*.

Matrix *A*

	HR1	HR2	HR3
RYAN	1	0	1
MCCARTHY	1	0	0
PELOSI	0	1	1

The transpose of *A*, or *A*<sup>T</sup> looks like this:

	RYAN	MCCARTHY	PELOSI
HR1	1	1	0
HR2	0	0	1
HR3	1	1	1

If one multiplies *A* by *A*<sup>T</sup> one gets the *N* × *N* adjacency matrix of shared co-sponsorship by legislator. *A* × *A*<sup>T</sup> =

	RYAN	MCCARTHY	PELOSI
RYAN	2	2	1
MCCARTHY	2	2	1
PELOSI	1	1	2

If one multiplies *A*<sup>T</sup> by *A* one gets the *M* × *M* adjacency matrix of shared legislators by bill.

*A*<sup>T</sup> × *A* =

	HR1	HR2	HR3
HR1	2	0	2
HR2	0	1	1
HR3	2	1	3

In addition, it is important to distinguish between *directed* networks and *non-directed* networks. The network example above is non-directed because the pair McCarthy–Pelosi is the same as the pair Pelosi–McCarthy. Think of the matrix as being split along the main diagonal of the matrix, from top left to bottom right. In a non-directed network, the matrix is symmetrical, and thus, the top half and bottom half of the matrix – on either side of the main diagonal – are identical. However, sometimes one’s network data are directed. In directed networks, also called a digraph, the edge that connects A to B is characteristically different than the edge that connects B to A. Directed networks are sensible for data that include interactions, such as a sender and receiver or a buyer and seller. If a network is directed, then the top and bottom half of the network will not likely be symmetrical.

In directed networks that are in square or adjacency format, one is able to distinguish between the *indegree* and *outdegree* of a node. The *indegree* are the number of edges that connect or point toward the node, and the *outdegree* are the number of edges that point away from the node. One can envision *outdegree* as the number of nodes that a given node connects to and *indegree* as the number of nodes that are connected to a given node. In a square directed matrix, the rows represent the ‘senders’ and the columns represent the ‘receivers’. In adjacency format, an observer is able to calculate the indegree and outdegree of directed data by simply calculating the sum of the rows (outdegree) or columns (called indegree). In non-directed data, the degree of a node is the sum of its row or column; row sums and column sums will be equal in non-directed data.

**MEASUREMENT, SAMPLING, AND MISSINGNESS IN NETWORK DATA**

Scholars trained in traditional social science research-methods theory and techniques will be familiar with challenges related to

measurement theory. The social scientist is consistently concerned about the reliability and validity of their measures and seeks to avoid making so-called Type I and Type II errors in the inferences they draw from the analysis.

It is crucial that when using network methods to pursue social scientific questions, the researcher is particularly diligent about such concerns.

Recall that the *reliability* of a measure describes the extent to which it produces the same result upon repeated trials, while the *validity* of a measure refers to the correspondence between a measure and the concept it is meant to capture. If a measurement strategy is at risk for violating these principles, the researcher is more likely to produce false positives, known as a Type I error, or false negatives, which is called a Type II error. In network studies, one must consider the possibility of developing measures or using data collection strategies that result in false-positive nodes, false-negative nodes, false-positive edges, false-negative edges, falsely aggregated nodes, or falsely disaggregated nodes. Lack of careful attention to these details can result in incorrect inference or erroneous descriptions of a network.

A false-positive node may occur if one inadvertently includes a node in a network to which it does not belong. In such a case, one would underestimate the density of the data, which could pose a threat to validity of network measures.

A false-negative node may occur if one excludes a node from a network to which it belongs. If the researcher excludes an important node, measures of centrality and power will be negatively affected. However, if the excluded node is of low degree or importance, there may not be much consequence to the inference drawn from network properties.

A false-positive edge occurs when a network measure includes a connection that does not exist. This would lead a researcher to overestimate the density of the network, but it is unlikely to have too many other consequences.

Likewise, a false-negative edge would occur when the researcher fails to observe a connection between two nodes that *is* actually present. The omission can lead the researcher to underestimate density as well as other measures of centrality and importance. The severity of such an omission will vary depending on the importance of the omitted edge.

A falsely aggregated node would occur if, for example, you counted Senator Bill Nelson (a Democrat from Florida) and Senator Ben Nelson (a Democrat from Delaware) as the same node, because perhaps you only recorded the last name and party of each node. Such an error can lead a researcher to overestimate the degree distribution of the data and can lessen the sensitivity available in the data. Likewise, a falsely disaggregated node would occur if a researcher counted South Carolina Senator Timothy Eugen Scott and Senator Tim Scott, Republican legislator of South Carolina, as different nodes, when in fact they are the same person. If this occurs, a researcher will underestimate the degree distribution of their data.

Issues of aggregation and identification are critical in network studies because network analysis typically draws on the properties of a network and misidentifying the nodes and edges of network can impact one's observation of the properties. In a traditional study, misidentification of this type may lead to an error in the number of observations, perhaps one that is less critical.

It is imperative that the network scholar pays meticulous attention to detail in data collection for this reason.

Another significant difference between network analysis and traditional frequentist inference is the use of sampling. Many social scientists rely on a process of random sampling to produce a manageable data set on which to perform analyses. The quality of the sample is directly related to the quality of the inference the statistician can make about the population of interest. However, with network data, sampling is often not feasible. Many networks of interest are unbounded. One literally cannot define the limits of the

population of interest, so how can one reasonably sample from it?

Consider a researcher who seeks to study political discussion networks of a particular community. The researcher can take a random sample of members of that community to study, but if the community members talk about politics with people outside of their community (likely), then the sample will have omitted relevant nodes. The difference is that in traditional statistics, one seeks to draw inference about an individual or a group of individuals in an identified population, but in network studies the unit of interest is often the relationship itself. It may be unreasonable to define the population of possible connections that exist if nodes can reasonably be connected to nodes outside the primary population. In this way, it can be nearly impossible to sample from networks because the boundaries of the network are undefined.

In addition, if a network scholar seeks to know about relationships in a particular community and the scholar samples that community to study its relationships, it is not possible to discern all of the network properties without some significant statistical error. Unsampled nodes and edges is akin to missing data that can affect one's understanding of network density, centrality of nodes, or component structures. In short, missingness is a significantly more pernicious problem in network studies than it is in traditional statistical studies (Burt, 1987). Typically, when observations are interdependent, missingness is not randomly distributed. Missing data is therefore a much more serious threat to inference in network studies than it is in non-network analyses (Kossinets, 2006).

## NETWORK VISUALIZATION

One of the more attractive features of network data are the data-visualization possibilities that accompany an analysis. Almost by definition, most network data are complex.

If one is engaged in a relational analysis of data, one is attempting to convey or analyze complexities of data. If one can do this in a way that provides a compelling narrative, the analysis is more meaningful. Visualization of network data is a vital component of the enterprise of network analysis precisely because all data analysis is ultimately about storytelling. If a visualization can convey a story easily and compellingly, it is a useful visualization.

As described by Pfeffer (2017), there are three critical design elements in network data visualizations: substance, design, and algorithm. First, to create a compelling visualization, one must have a compelling story to tell, and that entails knowing the story before one tries to create the graph. Sometimes visualizations are a good means of exploring data, but if the exploration reveals something compelling, then the visualization will likely be part of the reporting stage as well. It is therefore critical to have a substantive narrative in the data that one seeks to convey through a visualization.

Second, the design allows for a meaningful arrangement of nodes and edges that reveal the substance and structure of the network. Careful attention to the details of the arrangements of nodes and edges can help to convey the story. Finally, the algorithm used to produce a layout of a graph is the most critical choice an analyst will make. Different algorithms produce different shapes, effects, and emphases. In these details, one finds the elements of the narrative.

In general, the criteria for a good layout of a graph include the following:

- Show the structure and shape of the network from a bird's-eye perspective.
- Optimize the distribution of details; do overlapping nodes simplify or misrepresent the data?
- Maximize the data-ink ratio and avoid added elements or details that convey no information to a graph.
- Optimize line lengths, since short lines imply connectedness.
- Minimize crossing lines because they create confusion.

- Maximize angles; acute angles lead to crossing lines and less space between lines.
- Optimize path distances; adjacent nodes should be drawn close together.
- Quantitative data is best represented in order by position, size, saturation, hue, and shape.
- Nominal data is best represented in order by position, hue, saturation, shape, and size.

Other considerations with visualization are similar to the type of choices one would make for non-network graphs. For example, when varying the size of nodes, size should reflect a meaningful quantitative metric; however, humans have a difficult time distinguishing one size from one another, especially when objects are not directly adjacent, so it is unwise to expect a reader to draw too much inference from a node's size.

In color selection, humans can readily distinguish red, blue, and green, but too many colors or insufficient contrast between colors will render the technique less effective. Varying saturation, such as light and dark intensity, can be an effective means of conveying the levels of ordinal data.

Typically, curved lines are more aesthetically pleasing than straight lines, particularly for sparse networks. When networks are very dense, it is best to bundle or aggregate edges or nodes in some meaningful way or prune the network using a logical threshold.

## USEFUL NETWORK PROPERTIES FOR POLITICAL SCIENTISTS

The overall shape of a network and the features of its structure have many significant implications for questions of interest by social scientists. This section reviews the network properties of cohesion, reciprocity, transitivity, and centrality, and it shows how these physical properties convey concepts of interest to political scientists. For example, cohesive networks have an easier time solving coordination and collective-action problems – the primary challenge of all questions of politics – than less

cohesive networks. Those with more reciprocal ties and greater transitivity engender greater trust among participants, and the network concept of centrality is akin to the political concept of power. Translating political concepts into network properties provides a researcher with analytical leverage.

Herein, when the chapter talks about the cohesiveness of a network, it generally looks to its density. Most simply, the density of a network is the proportion of ties that exist out of all possible ties. In a non-directed graph, the number of possible ties is  $\frac{n(n-1)}{2}$ , where  $n$  is the number of nodes in the network. Density is a descriptive statistic that does not have inferential interpretation. Whether a density of 0.41 is high or low depends on the context of the network or a network to which you seek comparison. If ties represent armed conflicts between nation-states, then 0.41 is very high, but if it is the communication network among a group of six friends, it seems low.

When interpreting density, it is also important to consider whether ties are positive, such as within friendships, or negative, such as conflict between or among adversaries. Typically, networks with positive ties have a higher density than those with negative ties. If one seeks to interpret the density of a network by making a comparison to some other network, it is best if the two networks have the same overall size or the same number of possible connections. A common denominator always makes things easier to compare.

Another common way to understand cohesion is to calculate the average degree of a network. This simply amounts to calculating the degree or number of ties for each node, such as row sums, and then taking the average across these values. The density of a network and its average degree are related in that average degree is density times one less than the number of nodes. Some scholars find average degree easier to interpret than density, but in skewed distributions, which is typical of networks, the average degree may not be a good

representation of the central tendency of the degree distribution.

Large networks often have multiple components in them. One way to gain a sense of the cohesiveness of a network is to look at the size of its main component – the larger the main components, the greater the cohesion of the network. There are two useful statistics that may be useful in this case. First, a component ratio is an inverse measure of cohesion because larger values indicate less cohesion. The component ratio will be equal to one if every node is an isolate and equal to zero if there is only one component in the network

$$\text{Component ratio} = \frac{(c-1)}{(n-1)},$$

where  $c$  is the number of components and  $n$  is the number of nodes. However, the component ratio is not particularly sensitive or intuitively descriptive.

An alternative method to calculate cohesion is to measure the extent of fragmentation in a graph and to infer the connectedness from its complement. David Krackhardt (1994) developed an elegant measure of fragmentation that describes the proportion of dyads that are *not* located in the same component. Fragmentation is equal to one minus connectedness:

$$F = 1 - \frac{\sum r_{i,j}}{n(n-1)} = 1 - \text{Connectedness},$$

where  $r_{i,j}$  are dyads in the same component ( $r$  stands for reachability).

One can also get at the idea of cohesion by simply calculating the average path length, known as the *geodesic distance*. This is a common measure that is often used to describe how efficiently things might flow through a graph or get from one point to another. If the graph has multiple components, however, this measure is less useful. To gather the same type of information in a graph with multiple components, one can measure the *compactness* of the graph. Compactness is a variant of the connectedness measure described above, except

the numerator is replaced by the inverse geodesic distance between each dyad. In this way, compactness allows for paths to be weighed inversely by their length, which provides a much more sensitive measure of cohesion than any of the others described here.

$$\text{Compactness} = \frac{\sum_{i \neq j} \left( \frac{1}{d_{i,j}} \right)}{n(n-1)}$$

Finally, one might think of cohesion as a form of robustness. Robustness describes what happens to the network when individual nodes are removed. If the number of components increases by removing only a few nodes, then the network is not robust. Scholars can experiment with the removal of individual nodes and observe changes in the number of components or level of compactness in a network to gain a sense of its overall robustness.

In short, there are many ways to measure the cohesiveness of a network, and cohesion is a useful concept in politics. Many questions in politics come down to understanding how a group of actors will decide or solve a problem and then identifying the conditions under which the actions occur. Cohesive groups have less conflict and can more easily solve problems than non-cohesive groups. Studying the cohesiveness of networks is a natural methodological tool in the study of political relationships.

In addition to these various measures of cohesion, in directed graphs, one often seeks to understand the degree of reciprocity. In a reciprocal tie, when A is connected to B, then B is connected to A. Reciprocity is a natural concept to convey in terms of politics. Many political interactions deal with exchanges between two entities. If the exchange is positive, then reciprocity demonstrates cooperation. The act of giving and receiving is naturally human and social; measuring it through rates of reciprocity can give a researcher a handle on the amount of generosity, gratitude, or willingness to engage one observes in a network. Of course, reciprocity is a dyad-level feature of a

network – meaning it occurs between two nodes – and our ability to draw inference from it is limited to this level.

Beyond reciprocity, one can observe the level of transitivity and clustering in a network to help one to gauge the amount of trust between the actors. In a transitive relationship, if A is connected to B, and B is connected to C, then A is also connected to C. A, B, and C form a triad. When networks have many triads, they tend to be clumpy, and a triad census can help the political observer and scientist discern the types of triads and their frequency in a network. Triadic relationships tend to breed trust, because they have built-in mechanisms of accountability and monitoring. To illustrate with a colorful example, if political actor A tells a lie to B and the truth to C, then B checks the information with C, A gets caught in the lie and may suffer social consequences. The triadic nature of the relationships encourages truthful and positive interactions. Studying the triads and the extent of triadic closure in a network is useful for political scientists who want to understand the nature of trusting relationships between actors.

The final set of structural considerations about a network to consider here is centrality. There are multiple forms of centrality that one can use to describe a network. In general, centrality is a useful characteristic of a network for a political scientist to study because it describes the position of a particular node relative to other nodes. Nodes that are identified as more central to a graph can be thought of as having more power or influence on the graph. However, being central in a graph is not necessarily a good thing. If the network you study, for example, explains the spread of some misinformation or disease, then being central is much less desirable.

The paragraphs that follow describe five types of centrality: degree, eigenvector, beta, betweenness, and closeness. *Degree centrality* is a straightforward measure based on each node's degree. The downside to degree centrality is that it does not consider each

node's position relative to other nodes. Some might argue that in this way, degree centrality is not a measure of 'centrality' because it does not account for the whole network's properties, but it measures only a degree of competition between individual nodes.

Similar to degree centrality, *eigenvector centrality* shows that each node is weighed by the centrality of the nodes adjacent to it. Eigenvector centrality reminds the observer that centrality cannot be calculated without information about the entire graph, yet the measure itself is node level. Depending on the context of the data, eigenvector centrality can be thought of as popularity or exposure. A downside of eigenvector centrality is that it will not count the members of smaller components, so a political scientist would not want to use an eigenvector centrality measure on a densely clumpy network of many disconnected clusters.

*Beta centrality* measures the total influence a node can have on all other nodes through direct and indirect influence. Mathematically, it is the weighted sum of the total number of walks between each pair of nodes. In a network with many long chains of connections, this measure may be a good choice to use; however, the user must select the weighting factor, which gives the measure a sense of arbitrariness. The centrality measure is sensitive to the weight selected so it is essential to have a theory about how much to weigh the indirect influence of long chains of interactions.

*Closeness centrality* is an inverse measure of the sum of geodesic distances from a node to all other nodes. Large values indicate a node that is highly peripheral, so a normalized version is more intuitive. This measure is not well adjusted for networks with low variance or for networks with many components.

Finally, the phrase *betweenness centrality* explains how often a given node falls along the shortest path between two other nodes. A node's betweenness is zero when it is never along the shortest path between any two other nodes. Betweenness centrality is typically interpreted in terms of its potential for controlling flows through a network. Nodes with



high betweenness are in a position to threaten network disruptions, for example, in the way that an airline hub going offline is highly disruptive, because hubs have high betweenness centrality.

## CONCLUSION

This chapter is a concise primer on network analysis for political scientists. It demonstrates a few ways that network theory and methods have made significant social science contributions to the understanding of US politics, public policy, public administration, international relations, and other sub-disciplines. The literature in political networks is flourishing, as scholars increasingly apply the logic and methods of network analysis to the relational questions in politics and the vast interconnectedness of human interactions as they govern and make decisions about and for one another (Victor et al., 2017).

The basic concepts explained in the chapter draw from sociology, mathematics, computer science, and political science to showcase the ease with which scholars can learn and incorporate them into their analysis and study of politics. Taking care to collect data with network methods in mind, organize and manage data for network analysis, and apply network methods is a matter of careful training and practice. While this chapter has been only an introduction on the topic, it conveys how accessible, useful, and applicable the ideas are for political science observers, analysts, and scholars.

## REFERENCES

- Albert, Réka, Hawoong Jeong and Albert-László Barabási. 1999. 'Diameter of the World-Wide Web'. *Nature* 401 (6749): 130–1.
- Banerjee, Sikata. 2012. *Make Me a Man!: Masculinity, Hinduism, and Nationalism in India*. Albany: SUNY Press.
- Barabási, Albert-László. 2009. 'Scale-Free Networks: A Decade and Beyond'. *Science* 325 (5939): 412–3.
- Barabási, Albert-László. 2014. *Linked: How Everything Is Connected to Everything Else and What It Means for Business, Science, and Everyday Life*. 3/30/03 edition. New York: Basic Books.
- Berardo, Ramiro and Mark Lubell. 2016. 'Understanding What Shapes a Polycentric Governance System'. *Public Administration Review* 76 (5): 738–51.
- Berelson, Bernard R. 1954. *Voting: A Study of Opinion Formation in a Presidential Campaign*. Chicago: University of Chicago Press.
- Bond, Robert M., Christopher J. Fariss, Jason J. Jones, Adam D. I. Kramer, Cameron Marlow, Jaime E. Settle and James H. Fowler. 2012. 'A 61-Million-Person Experiment in Social Influence and Political Mobilization'. *Nature* 489 (7415): 295–8.
- Borgatti, Stephen P., Martin G. Everett and Jeffrey C. Johnson. 2018. *Analyzing Social Networks*. Second edition. Los Angeles: Sage.
- Bowers, Jake, Mark M. Fredrickson and Costas Panagopoulos. 2013. 'Reasoning about Interference Between Units: A General Framework'. *Political Analysis* 21 (1): 97–124.
- Bratton, Kathleen A. and Stella M. Rouse. 2011. 'Networks in the Legislative Arena: How Group Dynamics Affect Cosponsorship: Networks in the Legislative Arena'. *Legislative Studies Quarterly* 36 (3): 423–60.
- Broadbent, Jeffrey. 2017. 'Comparative Climate Change Policy Networks'. In *The Oxford Handbook of Political Networks*, Jennifer Nicoll Victor, Alexander H. Montgomery and Mark Lubell, Eds. New York: Oxford University Press.
- Burkett, Tracy and John Skvoretz. 2001. *Political Support Networks Among US Senators: Stability and Change from 1973 to 1990*. Unpublished Manuscript, College of Charleston 3123.
- Burt, Ronald S. 1985. 'General Social Survey Network Items'. *Connections* 8 (1): 19–23.
- Burt, Ronald S. 1987. 'A Note on Missing Network Data in the General Social Survey'. *Social Networks* 9 (1): 63–73.
- Campbell, Angus, Philip E. Converse, Warren E. Miller and Donald E. Stokes. 1980. *The*

- American Voter: Unabridged Edition*. Chicago: University of Chicago Press.
- Christakis, Nicholas A. and James H. Fowler. 2007. 'The Spread of Obesity in a Large Social Network over 32 Years'. *New England Journal of Medicine* 357 (4): 370–9.
- Christakis, Nicholas A. and James H. Fowler. 2008. 'The Collective Dynamics of Smoking in a Large Social Network'. *New England Journal of Medicine* 358 (21): 2249–58.
- Corbetta, Renato and William J. Dixon. 2005. 'Danger beyond Dyads: Third-Party Participants in Militarized Interstate Disputes'. *Conflict Management and Peace Science* 22 (1): 39–61.
- Eilstrup-Sangiovanni, Mette. 2017. 'Global Governance Networks'. In *The Oxford Handbook of Political Networks*, Jennifer Nicoll Victor, Alexander H. Montgomery and Mark Lubell, Eds. New York: Oxford University Press.
- Erdos, Paul and Alfréd Rényi. 1960. 'On the Evolution of Random Graphs'. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences* 5 (1): 17–60.
- Eulau, Heinz. 1962. 'Bases of Authority in Legislative Bodies: A Comparative Analysis'. *Administrative Science Quarterly*, 7 (3): 309–21.
- Faul, Moira V. 2016. 'Networks and Power: Why Networks Are Hierarchical Not Flat and What Can Be Done About It'. *Global Policy* 7 (2): 185–97.
- Fischer, Manuel. 2011. 'Social Network Analysis and Qualitative Comparative Analysis: Their Mutual Benefit for the Explanation of Policy Network Structures'. *Methodological Innovations Online* 6 (2): 27–51.
- Grinberg, Nir. 2018. 'Identifying Modes of User Engagement with Online News and Their Relationship to Information Gain in Text'. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web – WWW '18*, 1745–54. Lyon: ACM Press.
- Hafner-Burton, Emilie M. and Alexander H. Montgomery. 2006. 'Power Positions: International Organizations, Social Networks, and Conflict'. *Journal of Conflict Resolution* 50 (1): 3–27.
- Hafner-Burton, Emilie M., Miles Kahler and Alexander H. Montgomery. 2009. 'Network Analysis for International Relations'. *International Organization* 63 (3): 559–92.
- Horowitz, Michael C. and Philip B. K. Potter. 2013. 'Allying to Kill Terrorist Intergroup Cooperation and the Consequences for Lethality'. *Journal of Conflict Resolution* 58 (2): 199–225.
- Huckfeldt, Robert and John Sprague. 1987a. 'Networks in Context: The Social Flow of Political Information'. *The American Political Science Review* 81 (4): 1197–216.
- Huckfeldt, Robert and John Sprague. 1987b. 'Networks in Context: The Social Flow of Political Information'. *The American Political Science Review* 81 (4): 1197–216.
- Kapiszewski, Diana, Lauren M. MacLean and Benjamin L. Read. 2015. *Field Research in Political Science: Practices and Principles*. New York: Cambridge University Press.
- Keck, Margaret E. and Kathryn Sikkink. 1997. 'Transnational Advocacy Networks in the Movement Society'. In *The Social Movement Society*, David S. Meyer and Sidney Tarrow, Eds. Lanham: Rowman & Littlefield Publishers.
- Kinsella, David and Alexander H. Montgomery. 2017. 'Arms Supply and Proliferation Networks'. In *The Oxford Handbook of Political Networks*, Jennifer Nicoll Victor, Alexander H. Montgomery and Mark Lubell, Eds. New York: Oxford University Press.
- Kirkland, Justin H. 2011. 'The Relational Determinants of Legislative Outcomes: Strong and Weak Ties Between Legislators'. *The Journal of Politics* 73 (3): 887–98.
- Klofstad, Casey A., Anand Edward Sokhey and Scott D. McClurg. 2013. 'Disagreeing about Disagreement: How Conflict in Social Networks Affects Political Behavior'. *American Journal of Political Science* 57 (1): 120–34.
- Knoke, David and Tetiana Kostiuhenko. 2017. 'Power Structures of Policy Networks'. In *The Oxford Handbook of Political Networks*, Jennifer Nicoll Victor, Alexander H. Montgomery, and Mark Lubell, Eds. New York: Oxford University Press.
- Koger, Gregory and Jennifer Nicoll Victor. 2009. 'Polarized Agents: Campaign Contributions by Lobbyists'. *PS: Political Science & Politics* 42 (3): 485–8.
- Kossinets, Gueorgi. 2006. 'Effects of Missing Data in Social Networks'. *Social Networks* 28 (3): 247–68.
- Crackhardt, David. 1994. 'Graph Theoretical Dimensions of Informal Organizations'.

- In *Computational Organization Theory*, Kathleen Carley and Michael J. Prietula, Eds. Hillsdale, New Jersey: Psychology Press.
- LaZarsfeld, Paul Felix, Bernard Berelson and Hazel Gaudet. 1948. *The People's Choice: How the Voter Makes Up His Mind in a Presidential Campaign*. New York: Columbia University Press.
- Lazer, David M., Anand E. Sokhey, Michael A. Neblo, Kevin M. Esterling and Ryan Kennedy. 2015. 'Expanding the Conversation: Multiplier Effects From a Deliberative Field Experiment'. *Political Communication* 32 (4): 552–73.
- McClurg, Scott D. 2006. 'The Electoral Relevance of Political Talk: Examining Disagreement and Expertise Effects in Social Networks on Political Participation'. *American Journal of Political Science* 50 (3): 737–54.
- Montgomery, Jacob and Brendan Nyhan. 2017. 'The Effects of Congressional Staff Networks in the US House of Representatives'. *Journal of Politics* 79 (3): 745–761.
- Murdie, Amanda and Marc Polizzi. 2017. 'Human Rights and Transnational Advocacy Networks'. In *The Oxford Handbook of Political Networks*, Jennifer Nicoll Victor, Alexander H. Montgomery and Mark Lubell, Eds. New York: Oxford University Press.
- Mutz, Diana C. 2002. 'Cross-Cutting Social Networks: Testing Democratic Theory in Practice'. *American Political Science Review* 96 (1): 111–26.
- Nickerson, David W. 2008. 'Is Voting Contagious? Evidence from Two Field Experiments'. *American Political Science Review* 102 (1): 49–57.
- Patterson, Samuel C. 1959. 'Patterns of Interpersonal Relations in a State Legislative Group: The Wisconsin Assembly'. *Public Opinion Quarterly* 23 (1): 101–9.
- Pfeffer, Jürgen. 2017. 'Visualization of Political Networks'. In *The Oxford Handbook of Political Networks*, Jennifer Nicoll Victor, Alexander H. Montgomery and Mark Lubell Eds. New York: Oxford University Press.
- Porter, Mason A., Peter J. Mucha, Mark EJ Newman and Casey M. Warmbrand. 2005. 'A Network Analysis of Committees in the US House of Representatives'. *Proceedings of the National Academy of Sciences of the United States of America* 102 (20): 7057–62.
- Putnam, Robert D. 1966. 'Political Attitudes and the Local Community'. *The American Political Science Review* 60 (3): 640–54.
- Razo, Armando. 2017. 'Bringing Networks into Comparative Politics'. In *The Oxford Handbook of Political Networks*, Jennifer Nicoll Victor, Alexander H. Montgomery and Mark Lubell, Eds. New York: Oxford University Press.
- Reinicke, Wolfgang H., Francis Deng, Jan Martin Witte, Thorsten Benner, Beth Whittaker and John Gershman, Eds. 2000. *Critical Choices. The United Nations, Networks, and the Future of Global Governance*. First edition. Ottawa: IDRC Books.
- Ringe, Nils and Jennifer Nicoll Victor. 2013. *Bridging the Information Gap: Legislative Member Organizations as Social Networks in the United States and the European Union*. Ann Arbor, MI: University of Michigan Press.
- Rogowski, Jon C. and Betsy Sinclair. 2012. 'Estimating the Causal Effects of Social Interaction with Endogenous Networks'. *Political Analysis* 20 (3): 316–28.
- Rogowski, Jon C. and Betsy Sinclair. 2017. 'Causal Inference in Political Networks'. In *The Oxford Handbook of Political Networks*, Jennifer Nicoll Victor, Alexander H. Montgomery and Mark Lubell, Eds. New York: Oxford University Press.
- Rolfe, Meredith. 2013. *Voter Turnout: A Social Theory of Political Participation*. Reprint edition. Cambridge: Cambridge University Press.
- Routt, Garland C. 1938. 'Interpersonal Relationships and the Legislative Process'. *Annals of the American Academy of Political and Social Science*, 195 (1): 129–36.
- Schiffer, Eva and Jennifer Hauck. 2010. 'Net-Map: Collecting Social Network Data and Facilitating Network Learning through Participatory Influence Network Mapping'. *Field Methods* 22 (3): 231–49.
- Shakya, Holly B., Nicholas A. Christakis and James H. Fowler. 2017. 'An Exploratory Comparison of Name Generator Content: Data from Rural India'. *Social Networks* 48 (January): 157–68.
- Sinclair, Betsy. 2012. *The Social Citizen: Peer Networks and Political Behavior*. Chicago: University of Chicago Press.

- Sokhey, Anand Edward and Scott D. McClurg. 2012. 'Social Networks and Correct Voting'. *The Journal of Politics* 74 (3): 751–64.
- Sokhey, Anand E., Andy Baker and Paul A. Djupe. 2015. 'The Dynamics of Socially Supplied Information: Examining Discussion Network Stability Over Time'. *International Journal of Public Opinion Research* 27 (4): 565–87.
- Staniland, Paul. 2012. 'Organizing Insurgency: Networks, Resources, and Rebellion in South Asia'. *International Security* 37 (1): 142–77.
- Tam Cho, Wendy K. and James H. Fowler. 2010. 'Legislative Success in a Small World: Social Network Analysis and the Dynamics of Congressional Legislation'. *The Journal of Politics* 72 (1): 124–35.
- VanderWeele, Tyler J. 2011. 'Sensitivity Analysis for Contagion Effects in Social Networks'. *Sociological Methods & Research* 40 (2): 240–55.
- Vera, Eugenia Roldán and Thomas Schupp. 2006. 'Network Analysis in Comparative Social Sciences'. *Comparative Education* 42 (3): 405–29.
- Victor, Jennifer Nicoll and Gregory Koger. 2016. 'Financing Friends: How Lobbyists Create a Web of Relationships among Members of Congress'. *Interest Groups & Advocacy* 5 (3): 224–62.
- Victor, Jennifer Nicoll, Alexander H. Montgomery and Mark Lubell, Eds. 2017. *The Oxford Handbook of Political Networks*. Oxford and New York: Oxford University Press Inc.
- Watts, Duncan J. 1999. *Small Worlds: The Dynamics of Networks between Order and Randomness*. Princeton: Princeton University Press.

# Network Modeling: Estimation, Inference, Comparison, and Selection

John P. Schoeneman and Bruce A. Desmarais

## INTRODUCTION

Which international trade relationships are most likely to break down over the next decade? What predicts the level of cross-party collaboration in a legislature? Do online political discussion networks constitute as echo chambers? What these questions have in common is that they address systems of political relationships (trade, lawmaking, political communication) – systems that can be represented as networks in which political actors constitute the nodes, and the relationships constitute the edges between the nodes. Inferential network analysis represents a methodological class that can be drawn upon to answer questions like these. Most statistical models used in the social sciences rely on the assumption that the observations in a dataset (or at least clusters of observations) are drawn independently from a common data generating process. However, when analyzing networks, it is often inappropriate to assume that observations are independent

(Hafner-Burton et al., 2009; Harris, 2013). Indeed, the network analyst is often interested in studying the ways in which observations depend upon each other (Ogburn, 2018) (e.g., do friends influence each others' choices to vote? (Bond et al., 2012); do legislators reciprocate support for legislation? (Kirkland and Williams, 2014)).

The network-scientific approach has proven fruitful across most subfields of political science. For example, many, perhaps even most, of the empirical phenomena of interest to scholars of international relations can be represented as international networks. Paired with the availability of accessible software implementations, the conceptual appropriateness of inferential network analysis in international relations has led to its widespread application over the last 10–15 years. Networks that have been studied through the use of inferential network analysis include, but are not limited to, trade (Ward and Hoff, 2007; Fagiolo and Mastrorillo, 2014; Chu-Shore, 2010; Chyzh, 2016), conflict

(Ward et al., 2007; Cranmer and Desmarais, 2011; Gallop, 2016; Dorff and Ward, 2013), alliances (Cranmer et al., 2012), transnational terrorism (Desmarais and Cranmer, 2013; Metternich et al., 2013; Bush and Bichler, 2015; Asal et al., 2016), sanctioning (Cranmer et al., 2014; Dorff and Minhas, 2017), international governmental organizations (Cao, 2012; Davis and Pratt, 2016; Lupu and Greenhill, 2017), and non-governmental organizations (Atouba and Shumate, 2015). This research has led to several innovative findings. For example, Kinne (2018) finds that defense cooperation agreements (DCAs) between states are self-exciting in that they become more attractive as more states sign on, leading to patterns of triadic closure (i.e., a friend of a friend is a friend) and preferential attachment (i.e., popular states gain more ties) in the formation of networks through DCAs. Duque (2018), in a study of international diplomatic networks, finds similar relational dynamics in that the diplomatic relations of a state with one partner state affects its relations with other partner states.

Methods for inferential network analysis come primarily in the form of probabilistic models for networks, which are fit to data using common estimation frameworks (e.g., maximum likelihood estimation, Bayesian inference). Inferential network analysis is now far too broad of a methodological area to effectively review all of the models in any detail in this chapter. For a general overview of inferential models, the book by Carrington et al. (2005) is a good place to start. Additionally, Luke (2015) provides an overview of the available models using R software packages, with tutorials on how to implement them. However, it should be noted that methodology in social network analysis is a rapidly developing field and therefore most book length projects will not include the most recent advancements. In this chapter, we will review three modeling approaches in detail. These include the latent space model (LSM) (Dorff et al., 2016), the exponential random graph model (ERGM) (Cranmer and

Desmarais, 2011), and the stochastic block model (SBM) (Latouche et al., 2011; Sweet, 2015). The first model, the LSM, is a class of models in which each node is attributed with a position in a latent space of features. The latent space of features is parameterized to represent the structure of the network. The second model, the ERGM is a model that can be customized to represent networks with any quantifiable characteristic (e.g., a high number of reciprocated ties, a strong degree of clustering). The third model we discuss in detail is the stochastic blockmodel (SBM), which is a model in which the ties among units (i.e., nodes) in the network are reduced to relationships between groups of nodes. We chose the first two models, the LSM and the ERGM, on the basis that they are commonly used in the political science literature and therefore relevant to readers of a political science handbook. The SBM, however, has seen little use in political science. We make the case that the SBM should be used more in political science as an alternative to ERGM and LSM due to its focus on community detection, which is often of interests to political scientists (Desmarais et al., 2015; Freelon et al., 2015; Majó-Vázquez et al., 2019). After presenting a review of these three models, we provide a brief overview of extensions of these models to deal with longitudinal and weighted edge data. Next, we replicate the analysis in Wojcik (2017), which was originally completed using ERGMs, and analyze the replication data using both the LSM and the SBM.

## MODEL DESCRIPTION

The three models we review in detail, the LSM, ERGM, and SBM, share a common starting point in the case of dichotomous ties/edges (i.e., the edge does or does not exist). Consider a research problem in which the researcher is interested in using covariates that can be measured on the directed pair of

nodes  $(i, j)$  – where  $i$  is the index of the potential sender of an edge, and  $j$  is the index of the potential receiver of the edge – to model whether the edge  $(i, j)$  exists. Denote the edge indicator  $y_{i,j} = 1$  if there is an edge from node  $i$  to node  $j$ , and 0 otherwise. Let  $x_{i,j}$  be a covariate that can be measured on the directed dyad (e.g., if the nodes are states, the distance between their capital cities) or mapped to the dyad (e.g.,  $x_{i,j}$  is the GDP of the sender state,  $i$ ). A standard approach to modeling the dyadic variable  $y$  as a function of  $x$  would be to estimate a logistic regression in which  $y$  is the dependent variable and  $x$  is the independent variable. The LSM, ERGM, and SBM all reduce to a standard logistic regression model in which edge indicators (or values) are regressed on covariates in the case where the network components do not contribute to the fit of the models (Lubbers and Snijders, 2007; Raftery et al., 2012; Sweet, 2015). As with common regression models, model fit can be assessed through information criteria, prediction experiments, and simulation of network structural quantities (e.g., the distribution of the number of ties to which nodes are incident, the number of triangles in the network). Stated succinctly, logistic regression is a special case of LSM, ERGM, and SBM when modeling a dichotomous tie (i.e., absent or present) network.

In the case of each model, network structure is layered on top of the regression model. The particular form of network structure added varies across the models. In the LSM, each node is represented by some number of continuous-valued features (e.g., coordinates in a Euclidean space), and the probability of a tie is given by some function of the features (e.g., the Euclidean distance between nodes). We refer to the network structure incorporated in the LSM framework as ‘selection’, since the function defined on the latent feature serves as a partner selection function for nodes in the network. The network component under ERGM is designed to capture the prevalence of subnetworks that are theoretically

interesting or otherwise distinct. Examples of these subnetworks include triangles (i.e., triples of nodes  $(i, j, k)$ , in which there is an edge connecting each pair of nodes) and mutual dyads (i.e., dyads, in which there is an edge from  $i$  to  $j$  and from  $j$  to  $i$ ). In the ERGM, subnetwork prevalence is controlled through the specification of dependence relationships among the edges. For example, if a network is modeled to have a relatively high number of mutual dyads, the formation of an edge from  $i$  to  $j$  increases the likelihood that an edge will form from  $j$  to  $i$ . Since the basic building blocks of the network component of the ERGM reflect different forms of dependence, we refer to the network structure in the ERGM as ‘dependence’. The network structure of the SBM is built on a common interest in identifying the communities, clusters, or groups that define the main blocks of ties in the networks. Under the SBM’s network structure, the blocks of which two nodes are members determines whether those two nodes will form an edge. Since the network component of the SBM is focused on segmenting the nodes into a set of groups, we refer to the network structure in the SBM as segmentation.

### ***Latent Space Modeling of Networks***

The notion of feature-based selection is deeply rooted in political science literature. For example, in the study of legislative networks, ties between legislators have been repeatedly found to be driven by ideology and the geographic proximity of districts (Bratton and Rouse, 2011; Clark and Caro, 2013; Osei and Malang, 2018). The probability of a tie between two legislators is a function of the features of the two legislators – primarily ideology and geography. The LSM reflects this very structure – the probability of a tie depending on a combination of functions of node (e.g., legislator) features – with exception being that the features inferred in

the LSM are latent (i.e., unmeasured). That is, given one or more functions according to which the latent features affect the probability of an edge, in the LSM, the nodes' feature values are inferred to be those that best explain the pattern of edges observed in the data.

Under the LSM,

$$Pr(y_{i,j} = 1) = \text{logit} \left[ \beta_0 + \sum_{p=1}^P \beta_p x_{i,j}^{(p)} + d(z_i, z_j) \right],$$

where  $\text{logit}(a) = 1/(1 + \exp(-a))$ ,  $\beta_0$  is an intercept that controls the overall density of the network,  $\sum_{p=1}^P \beta_p x_{i,j}^{(p)}$  models the effect of measured covariates on the probability of an edge forming,  $z_i$  and  $z_j$  are latent coordinates of nodes  $i$  and  $j$ , and  $d()$  is a distance function that maps the coordinates into a metric space. In the LSM, the  $z$  is inferred as parameters. They are latent attributes that capture the structure of the network but are not observed. Example functions that have been used for  $d()$  include the negated Euclidean distance (Mahmood and Sismeiro, 2017),

$$d(z_i, z_j) = -\sqrt{\sum_{k=1}^K (z_i^{(k)} - z_j^{(k)})^2},$$

and the bilinear metric (Vance et al., 2009),

$$d(z_i, z_j) = \sum_{k=1}^K z_i^{(k)} \times z_j^{(k)}.$$

The latent variables account for network structure, which is not modeled effectively using the observed covariates.

Selection of the particular distance function used can be based on a few different factors. First, the researcher may have a substantive reason to prefer one distance function over another. For example, the Euclidean function models the effect of latent variables such that the log-odds of a tie is inversely

proportional to the straight-line distance between two nodes in the latent space, whereas under the bilinear function, the log-odds of a tie is inversely proportional to the angle between the two coordinate vectors. In these two examples, the Euclidean distance relies solely on proximity as a fixed effect, where zero represents the highest likelihood of a tie, but the bilinear metric is a mean zero random effect, where two nodes being close to one another does not necessarily increase the likelihood of a tie since they need to be in the same direction to increase the likelihood of a tie (Dorff et al., 2016). The researcher may consider one distance function to be most substantively appropriate in their application based on this difference. Second, one function may fit the data most effectively. Based on information theoretic measures such as the Bayesian Information Criterion (BIC) (Wang, 2009) or visual assessments of model fit, the researcher may determine that one of the distance functions provides the best fit to the data. Third, if the network is very large, computational properties of estimation may make the use of one of the distance functions more feasible. If the Markov Chain Monte Carlo approach to Bayesian inference in the LSM converges much faster under one distance function than another, the researcher may choose to go with the model that converges faster.

### The Exponential Random Graph Model

Theories of dependence in political science have been developing rapidly in recent years. Example findings include a strong tendency towards rapid retaliation in the issuance of international economic sanctions (Cranmer et al., 2014), bilateral preferential trade agreement networks' formation of triads and/or four-cycles to share costs among larger partner groups (Milewicz et al., 2018), and that influence networks among policy actors are hierarchical – exhibiting a tendency towards



transitive tie formation (Christopoulos and Ingold, 2015). The ERGM opens up a new class of hypotheses that can be tested relative to regression models. In addition to studying how covariates effect the absence/presence of edges, the ERGM permits researchers to test hypotheses about how the edges affect each other (e.g., is the enemy of an enemy a friend?). The key components in using an ERGM are to develop hypotheses regarding the ways in which edges depend on each other and conceptualize these dependencies in terms of subnetwork structures such as dyads and triads.

Under the ERGM, the probability (likelihood function) of observing the entire network  $Y$  is

$$P(Y, \theta) = \frac{\exp\{\theta'Y\}}{\sum_{Y^* \in \mathcal{Y}} \exp\{\theta'Y^*\}}$$

where  $h(Y)$  is the vector of network statistics selected to specify the model,  $\theta$  is the vector of effects of the network statistics on the probability of observing a particular configuration of the network,  $\exp\{\theta'h(Y)\}$  is the positive weight associated with the respective network configuration,  $\mathcal{Y}$  is the set of all possible network configurations, and  $\sum_{Y^* \in \mathcal{Y}} \exp\{\theta'Y^*\}$  is the normalizing constant that assures that the probability distribution over networks sums to one. The ERGM is specified through the selection of  $h(Y)$  and is very flexible. This flexibility comes at two costs when it comes to application. First, since  $\mathcal{Y}$  is so large –  $2^{n*(n-1)}$  for a directed network with no self-ties, where  $n$  is the number of nodes in the network – the likelihood function cannot be computed exactly, which means estimation involves either approximation of the likelihood or method of moments criteria (Hummel et al., 2012; He and Zheng, 2016; Schmid and Desmarais, 2017). Second, the researcher may specify a model that cannot be effectively fit to the observed data, the consequence of which is model degeneracy. Model

degeneracy is a condition under which the model does not converge to a model that can effectively approximate even the number of edges in the network, the result of which is a model that places nearly all of the probability on either the completely full or completely empty network (Schweinberger, 2011). An active methodological literature has made substantial progress towards solving these two problems (DeMuse et al., 2018) and state-of-the-art open-source software implements many of these solutions (Hunter et al., 2008).

The ERGM has been used to study several different forms of dependence in political networks – structural patterns that go beyond the effects of covariates, of which the ERGM is uniquely capable of modeling. These dependence patterns concern the ways in which ties in the network are related to each other and are modeled through the specification of corresponding network statistics in  $h(Y)$ . Peng et al. (2016) apply the ERGM to study reciprocity in the online communication networks among members of the US Congress. Osei and Malang (2018) use the ERGM to analyze the degree to which political discussion is characterized by triadic closure (i.e., the friend of a friend is a friend) among members of the Ghanaian legislature. Gerber et al. (2013) use ERGMs to study the degree to which ties beget ties (i.e., tie formation is driven by popularity dynamics) in organizational collaboration networks surrounding local regional planning in California. Guerrero et al. (2015) analyze a multilevel collaboration network among organizations working on agricultural conservation in Australia. The levels include individual property owners, sub regional organizations (e.g., local governments) and supra-regional organizations (e.g., NGOs). They use the ERGM to study the prevalence of nodes that bridge multiple levels in the network. Dependence patterns such as the ones cited in this paragraph represent an exciting and important component of the function of

political systems, and a class of structural patterns that can only be studied through the use of inferential network analysis.

### The Stochastic Blockmodel

Concepts in political science related to group and system polarity (Baldassarri and Bearman, 2007; Cranmer et al., 2015) fall within the orbit of the SBM formulation. Under the SBM there is a finite number of node types or ‘blocks’, and each block is defined by the probabilities according to which nodes within the block form ties with each other and with nodes in other blocks. If there are  $n$  nodes in the network, the SBM reduces the complexity of interactions among those  $n$  nodes to interactions among and within a set of  $b$  blocks, where  $b$  is typically much smaller than  $n$ . The SBM with covariates, to our knowledge, has not been applied in the published political science literature. That is, perhaps, because it was only recently extended by Sweet (2015) to incorporate covariates – unlike the ERGM and the LSM, which were developed in part with the objective of modeling the effects of nodal and dyadic covariates. As such, unlike our review of the ERGM and the LSM, our presentation of the SBM represents an introduction of the model to the political science research methods literature.

Under the SBM, the probability of a tie from  $i$  to  $j$  is given by

$$Pr(y_{i,j} = 1) = \frac{1}{1 + \exp\left[-\left(\text{logit}\left(\pi_{B(i),B(j)}\right) + \sum_{p=1}^P \beta_p x_{i,j}^{(p)}\right)\right]}$$

where  $\pi_{B(i),B(j)}$  is the baseline probability with which a member of  $i$ 's block ( $B(i)$ ) sends a tie to a member of  $j$ 's block. The probability of a tie is also affected by a set of covariates. The blockmodel component of the SMB is defined by a  $k \times k$  matrix,  $\Pi$ , of probabilities,

where  $k$  is the number of blocks to which nodes are assigned, and the  $i, j$  element of  $\Pi$  gives the probability that a member of block  $i$  sends a tie to a member block  $j$ . If the network is undirected,  $\Pi$  is a symmetric matrix, and each element gives the probability of a tie between members of two blocks (or within the same block, if on the diagonal of  $\Pi$ ). Under the basic version of the SBM, the researcher sets the number of blocks, and each node is assigned to one block. The SBM, like the LSM, represents a latent variable approach to accounting for the network structure that cannot be modeled with the observed covariates. Also, like the LSM, the SBM cannot be used to model dependence effects and is most appropriate for applications in which the goal is to model the effects of covariates while accounting for network structure.

The specification of the SBM with covariates is, as is the specification in many modeling frameworks, driven by two factors – theoretical considerations and model fit. The selection of the number of blocks to which vertices are assigned is typically done using a measure of model fit such as the BIC (Robinson et al., 2015), or the Deviance Information Criterion (DIC) (Sohn and Park, 2017). In the application below, we select the model with the number of blocks that yields the lowest DIC. In terms of covariate specification, the SBM can model the effects of any covariates that one would include in a standard logistic regression, ERGM and/or LSM. As in logistic regression, and the standard versions of both the ERGM and the LSM, the coefficient associated with a covariate yields the change in the log odds of a tie due to a one unit increase in the covariate value.

### Additional Methods in Brief

The LSM, ERGM, and SBM reviewed above represent the basic forms of the most

commonly used models for statistical inference with networks. However, there is a substantial, decades-old, literature that expands upon these core model forms and has resulted in methods that are appropriate for studying networks characterized by ties that take on many different value types (e.g., continuous, count, textual) and are appropriate for time-stamped network data. In this section, we provide a brief review of the different extensions that have been built upon the foundations of the LSM, ERGM, and SBM, with a specific focus on networks with non-binary ties and those with temporal structure.

Any relationship that can be measured in a binary fashion can probably also be attributed with additional, and perhaps more theoretically interesting, quantitative information. For example, Cranmer and Desmarais (2011) and Baller (2017) use the ERGM to study cosponsorship networks between legislators by applying a threshold such that a tie between legislator  $i$  and  $j$  indicates that their cosponsorship relationship exceeded a set threshold. However, it may be more informative to study the network with weighted ties as a function of the number of times  $i$  cosponsored bills sponsored by  $j$  or the number of times two nodes cosponsored the same bills, as has been done in recent applications with cosponsorship networks (Kirkland, 2012; Signorelli and Wit, 2018). Both of the R packages commonly used for latent space modeling of networks – *amen* (Hoff et al., 2017) and *latentnet* (Krivitsky and Handcock, 2017) – include extensions of the binary LSM to allow modeling of continuous (*amen/latentnet*), count (*latentnet*), and ordinal (*amen*) edge weights. Krafft et al. (2012) developed a model for text-valued networks (e.g., e-mail text among a set of nodes) that integrates the LSM with topic modeling. The ERGM has been extended to model quantitatively valued and rank-ordered ties (Wyatt et al., 2009; Desmarais and Cranmer, 2012; Krivitsky, 2012; Krivitsky and Butts, 2017). The Weighted Stochastic Blockmodel,

introduced by Aicher et al. (2014), can be used to model edge values using any exponential family probability distribution. Though it is illustrative to consider each model in the case of simple binary ties, there are plenty of modeling options for working with ties that are attributed with more information.

Particularly in political science – where networks are often constructed from long-running historical records (e.g., international conflict (Li et al., 2017)), court/judicial processes (Box-Steffensmeier and Christenson, 2014), and legislative records (Ringe et al., 2013) – it is common for network data to be time-stamped and form long periods. The basic versions of the LSM, ERGM, and SBM presented above are appropriate for modeling single-time-point networks. However, each model has been extended to accommodate time series network data. There are two broad types of time-stamped data encountered in network analysis – relational state data and relational event data. Relational state data is network data in which the ties reflect durable relationship conditions (e.g., two states are or are not at war for sustained periods of time, two legislators are or are not serving on the same committee for sustained periods of time). Relational event data is network data in which relationships manifest as instantaneous (e.g., one government official sends an email to another) or very short-term (e.g., two political activists attend a rally together) interactions. Relational event data can be, and often is, converted to relational state data by indicating the existence of a tie between two nodes if the frequency and/or intensity of relational events within a given time period exceeds some threshold. For example, in the previous paragraph, we discussed ways in which researchers have measured cosponsorship networks, using them to indicate relational states over legislative session. Cosponsorship data is observed as relational events. Brandenberger (2018) analyzes time-stamped cosponsorship activity from the perspective of relational event network analysis. The LSM has been extended to model

relational state data (Sewell and Chen, 2015) but, to our knowledge, has not been developed for relational events. The ERGM has been extended for both relational state data (Hanneke et al., 2010) and relational events (Perry and Wolfe, 2013). The SBM has also been extended for both relational state data (Xu and Hero, 2014) and relational events (DuBois et al., 2013).

### **APPLICATION OF THE ERGM, LSM, AND SBM**

In this section we provide an example of using all three models of inferential network analysis covered in this chapter. This application is intended to demonstrate that the models produce comparable results with regard to covariate inferences, as they all have the same basic underlying regression framework, but also highlight how the different methods of approximating network structure shape differences in conclusions. The example also highlights differences such as the ability to include network structure in the ERGM and visually examine the latent space for the LSM and the block assignments in the SBM. For the application, we replicate a recent study of the factors that shape directed connections between legislators in Brazilian legislative networks (Wojcik, 2017). Wojcik argues that, despite high levels of personalism in Latin American politics, political parties are effective in shaping actor behavior, indicating that there are higher levels of party discipline than previously thought. By addressing this problem from a network framework, the author offers new evidence that the relationships between legislators exhibit complex forms of dependence.

We use the R software package *ergm* (Handcock et al., 2018) to replicate the ERGM models, *latentnet* (Krivitsky and Handcock, 2017) for the LSMs, and *CIDnetworks* (Adhikari et al., 2015) for the SBMs. The

ERGMs were replicated as originally specified and used the same random seed as the original paper. For the LSMs, we specify them the same way as the ERGMs, but without transitive ties, and run models with varying numbers of latent space dimensions and clusters, reporting the result from the models that had the lowest BIC. For the SBM, we fit models specified the same with two to five blocks, again reporting the model results that has the lowest DIC. In general, we find that results from all the models match the general conclusions of the paper. Below, we explain the data from the study, the model specifications, and the differences in results in more detail.

### **Data**

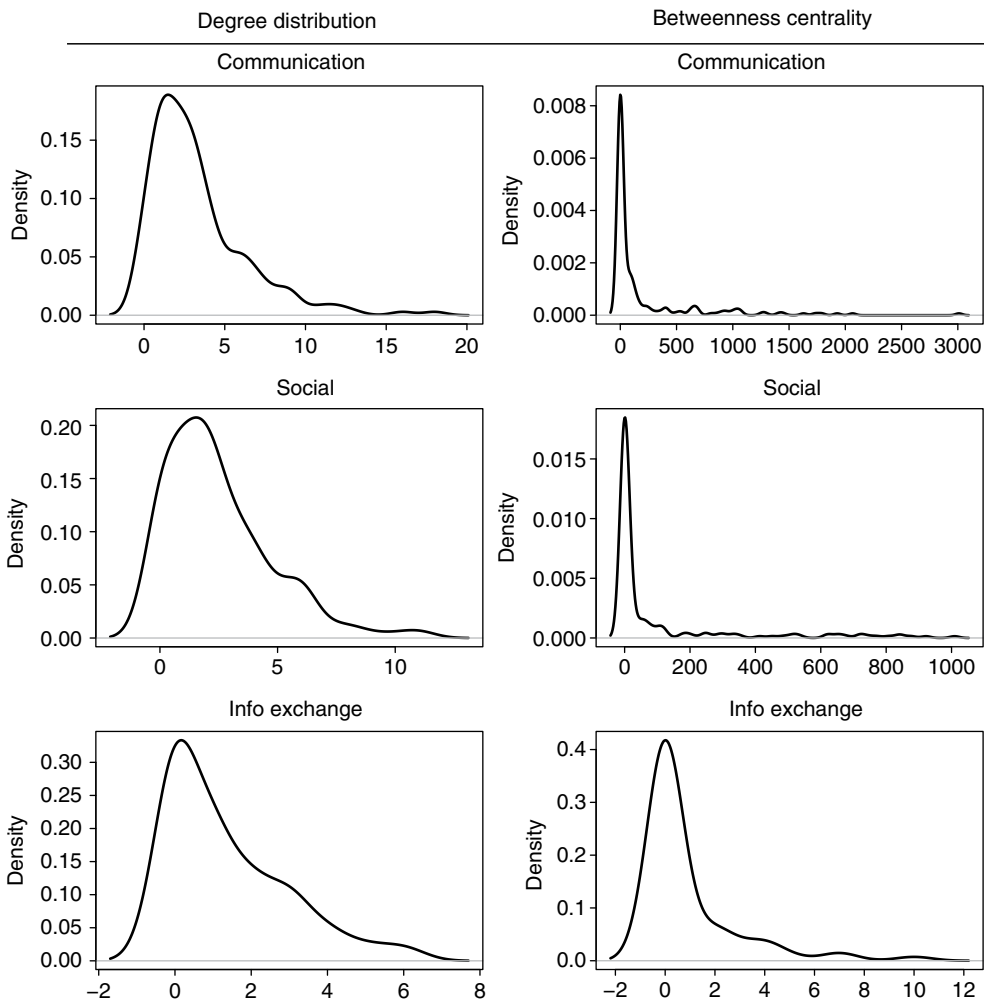
To collect the network data, Wojcik distributed an interactive digital survey to office managers from the offices of the lower chamber of the Brazilian national legislature. The three different types of connections surveyed were communication on legislative issues, socialization about non-legislative issues, and information gathering to resolve questions about legislative proposals and procedures. Wojcik also collected biographic information on those who completed the survey and compared traits of the sample to that of the entire Congress, finding no identifiable differences. The survey had a 25% response rate, and missing data was filled in using results from a pen-and-paper survey from 2013.<sup>1</sup>

Summary statistics of graph-level network measures for the three networks are presented in Table 46.1. All of the networks' edge densities – the proportion of possible ties that exist, with a range of zero to one – have fairly low values. This is interesting because, for this application, it indicates that legislative communications are sparse. This means that central actors are more important for assuring that information is spread throughout the network. Betweenness centrality of a node is the number of shortest paths between nodes

**Table 46.1 Summary statistics: networks**

Network	Measurement	Value
Communication	Edge density	0.008
Communication	Reciprocity	0.255
Communication	Transitivity	0.108
Social	Edge density	0.006
Social	Reciprocity	0.203
Social	Transitivity	0.119
Info exchange	Edge density	0.009
Info exchange	Reciprocity	0.157
Info exchange	Transitivity	0.135

in the network on which the node sits (Zhu and Watts, 2010) and is a common measure of how critical a node is to connect different parts of a network. Depicted in Figure 46.1, the long tails of the betweenness centrality distributions also indicate that there are only a handful of influentially positioned nodes. However, it should be noted that while the degree distributions – the distributions of the number of ties to which nodes are incident – are skewed right, the mean degree is high enough in the communication and social



**Figure 46.1 Density plots for degree distribution and betweenness centrality for all three networks**

networks to indicate that these networks are more spread out relative to the information exchange network. Additionally, the drop in reciprocity for the information exchange network, with an increase of transitivity, indicates that this network, relative to the others, mainly forms connections in one direction. Largely one-directional networks are interpreted as hierarchical (Liu et al., 2012). These substantial structural differences across networks drive home the usefulness of using inferential network analyses instead of models that assume independence, as they help explain the formation in ties beyond the exogenous covariates.

The independent variables, summarized in Table 46.2, have very similar values for all three networks. This is expected since they are drawn from the same population, but it does indicate that there are no major biases in the samples, despite the networks differing in size. We also see that most nodes have the same education level and that this lack of variation helps explain why education is not always a significant predictor of ties.

**Hypotheses**

The main hypothesis of the paper is that political parties actively work to form ties among members and thus counter the individualistic tendencies of politicians to increase their own electoral share/power at

the expense of the party. The network then creates trust and cohesion among the actors, while also facilitating monitoring. Wojcik expects that political party affiliation will be the stronger driver of tie formation, relative to individual traits.

The secondary hypotheses of the paper are based on the idea that politicians’ individual traits will significantly impact their likelihood to form a tie with one another, despite these effects being weaker than shared party membership. Wojcik includes five different individual traits with different combinations of assortative and disassortative hypotheses. The first is that when two politicians are both in leadership positions, they will need to interact more – but when neither are in leadership positions they will not need to interact with one another and can rely on party leaders. The second is that when politicians have similar ages, they will form connections based on shared experiences. The third is that when politicians are from the same state, they will form connections since there is a proportional voting system in Brazil, and that means that coordination among members in the same states is necessary to exchange information about local needs. The fourth is that politicians with the same education level are likely to have similar knowledge and social class, which increases the chance of network ties. The fifth is that when politicians work on the same floor this creates opportunity for interaction.

**Table 46.2 Summary statistics: independent variables**

Variable	Proportion		
	Communication	Social	Info exchange
Same education level	71%	70%	74%
Same floor	10%	10%	12%
Both leadership	6%	7%	6%
Both non-leadership	10%	10%	14%
Same state	10%	9%	12%
Same party	46%	50%	43%
	Average		
Δ Age	12	12	13

## Replication Design

The ERGMs are specified in the article to include an intercept term, the exogenous covariates, and a term for transitive ties. Transitive ties is a network statistic, specific to ERGM, and is defined as the number of ties from  $i$  to  $j$  for which there is at least one third node  $k$  such that there is a tie from  $i$  to  $k$  and a tie from  $j$  to  $k$ . The coefficient attached to the transitive ties term can be interpreted as the effect on the log odds of a tie from  $i$  to  $j$  of there being at least one indirect path of length two from  $i$  to  $j$  (i.e., some third node  $k$ , such that there is an edge from  $i$  to  $k$  and from  $k$  to  $j$ ).

For the LSM and the SBM, the specification is the same, except it does not include a transitive ties term. This is because the models do not have network statistics as an ERGM model does, but instead we specify the latent space that models higher order dependencies, including transitivity. For our extensions, we use the Euclidean distance model term that is equal to the negative Euclidean distance between actors in the unobserved social space. We estimated the term along two dimensions and across two or three clusters, depending on best model fit using the BIC. The SBMs do not include a constant term, as the block or group effect takes the place of the intercept term. It also does not have network terms like transitive ties as it accounts for dependencies using the block membership for the nodes.

## Results

The results regarding the direction and signs of the covariate effects are fairly consistent across models (Table 46.3). However, the coefficients in the LSMs, while signed the same, are mostly larger in magnitude. One exception is the node-match variables for leadership – a covariate that is equal to 1 if  $i$  and  $j$  have the same leadership status, and 0 otherwise. In Model 1, the effect size is

slightly larger, but in Model 3, the effect size is nearly one-third smaller. Moreover, the changes in magnitude are generally smaller for the leadership variables than the other variables. The latent space plots (Figure 46.2) show that the clusters are not distinct from one another, but rather there are inside groups that form clusters and each subsequent group of nodes are less central to the graph. Note, in the LSM, the clusters are post-hoc groupings of nodes, and do not play a significant role in the tie generation of the networks, unlike in the SBM.

The results from the SBM present a contrast to those of the LSM, as the coefficients for exogenous covariates, while signed the same, are mostly smaller in magnitude, indicating that block membership is capturing the impact of many of the variables. This is reasonable since we would expect block membership to be associated with variables such as party membership. Also, the coefficient estimates have smaller standard errors relative to the size of the estimate. In Figure 46.3, block likelihoods for each node assignment shows that for communication ties in Model 1, we see that the blocks are fairly evenly distributed but that membership is not distinct and that many of the nodes have spread out likelihoods for membership assignment. In Model 2, we see that this shared membership is even greater for the social networks. However, for the information network in Model 3, membership likelihood is sharply separated and concentrated in one block for each node.

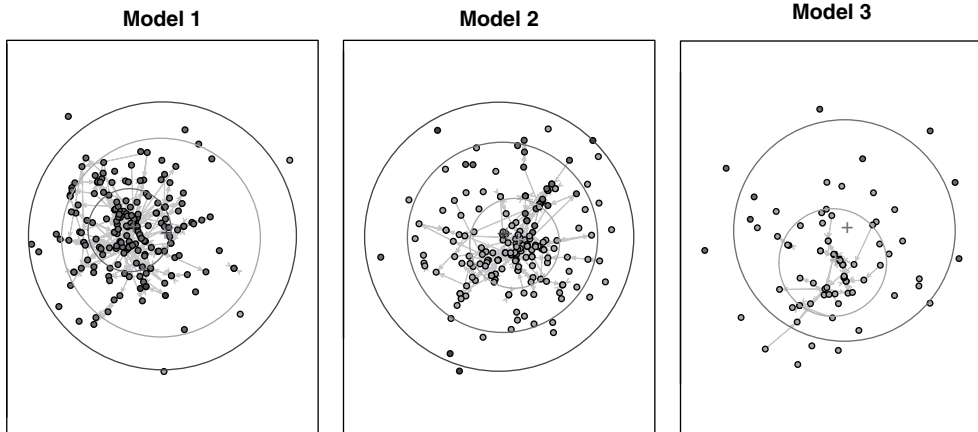
It is not immediately obvious which model performs best for this application, but, in general, there are a couple factors that influence which model is used for final analysis. The first is that theory should drive the decision of which model to use. All three models control for network dependence but differ in how they allow the researcher to explore the different forms. For example, the ERGM focuses on graph level dependence and if a scholar is interested in studying the presence and magnitude of a specific type of graph-level

**Table 46.3 Replication results**

	ERGM	LSM	SBM	ERGM	LSM	SBM	ERGM	LSM	SBM
	Communication networks			Social networks			Information exchange		
Edges (intercept)	-5.67*** (0.16)	-3.10*** (0.09)	--	-6.07*** (0.19)	-3.00*** (0.29)	--	-5.74*** (0.45)	-3.12*** (0.62)	--
Δ Age	-0.02*** (0.01)	-0.03*** (0.01)	-0.01*** (<0.01)	-0.02*** (0.01)	-0.04*** (0.01)	-0.01*** (<0.01)	-0.06*** (0.02)	-0.08*** (0.03)	-0.04*** (0.01)
Same education level	0.164 (0.125)	0.22*** (0.031)	0.143** (0.057)	0.161 (0.144)	0.35*** (0.084)	0.081 (0.070)	0.600* (0.334)	0.93** (0.433)	0.293** (0.150)
Same floor	0.76*** (0.15)	1.02*** (0.15)	0.32*** (0.07)	1.37*** (0.14)	1.53*** (0.01)	0.63*** (0.07)	0.79*** (0.39)	1.16** (0.49)	0.43** (0.17)
Both non-leadership	-0.66*** (0.13)	-0.82*** (0.05)	-0.28*** (0.05)	-0.24* (0.14)	-0.28** (0.42)	-0.13** (0.06)	-0.81** (0.34)	-0.82** (0.11)	-0.41*** (0.14)
Both leadership	0.534*** (0.16)	0.61*** (0.01)	0.21*** (0.07)	0.47** (0.21)	0.44 (0.30)	0.20** (0.10)	0.84** (0.40)	0.64 (0.5)	0.34** (0.17)
Same state	2.12*** (0.12)	2.75*** (0.11)	1.05*** (0.06)	2.11*** (0.14)	2.56*** (0.18)	1.04*** (0.07)	0.99** (0.43)	1.31* (0.67)	0.46** (0.20)
Same party	2.32*** (0.12)	2.86*** (0.11)	1.09*** (0.05)	2.05*** (0.14)	2.30*** (0.03)	0.99*** (0.07)	2.96*** (0.34)	3.47*** (0.41)	1.36*** (0.14)
Transitive ties	0.66*** (0.12)	--	--	0.90*** (0.14)	--	--	0.57 (0.40)	--	--
Latent dimensions	--	2	--	--	2	--	--	2	--
n blocks/clusters	--	3	3	--	3	2	--	2	3
N	39,402	39,402	39,402	36,290	36,290	36,290	5550	5550	5550

Note: \*p < 0.1; \*\*p < 0.05; \*\*\*p < 0.01.





**Table 46.4** Plots of latent space positions for the LSMs along two dimensions. The gray lines represent the directed edges between nodes. Each cluster has a unique grayscale shade that is used for the nodes and line that encloses the nodes in the cluster

network dependence, such as the transitive ties statistic in this application, the ERGM is the best suited model. If, however, a scholar is interested in understanding relationships between clusters of observations and the likelihood that an actor in one cluster will form ties with an actor in another cluster, the SBM is best suited for examining this. From the stochastic block matrices, researchers are also able to investigate network structure hypotheses such as the existence of overlapping or distinct communities, the rates of within-community tie formation, and how communities are connected to other communities (e.g. core-periphery). The LSM, on the other hand, is able to control for dependence, but does so in a way that embeds network structure in continuously valued attributes of nodes. Thus, it may work best for scholars that want to control for the dependence but are not interested in how network structure is driven by individual node attributes.

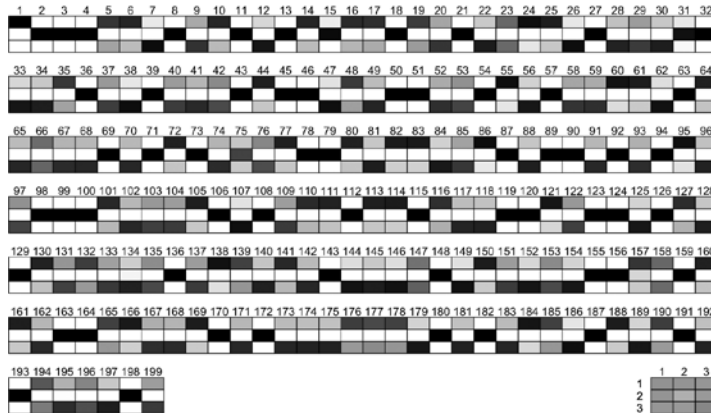
The second issue is practicality. Any scholar that has used ERGMs, has likely dealt with model degeneracy. For this reason, and if the scholar wants to account for network dependence, but does not want to work through specifying relevant dependence

terms, the LSM and SBMs are both viable options. As this replication has shown, the flexibility of each method often makes it possible that, if for nothing but robustness testing, scholars can implement more than one model type and gain insight through comparing model results. In the case of the current application, the central conclusion regarding party organization is very robust. The log odds of a tie is one to three units higher when two legislators are in the same party than when they are not in the same party, across networks and modeling approaches.

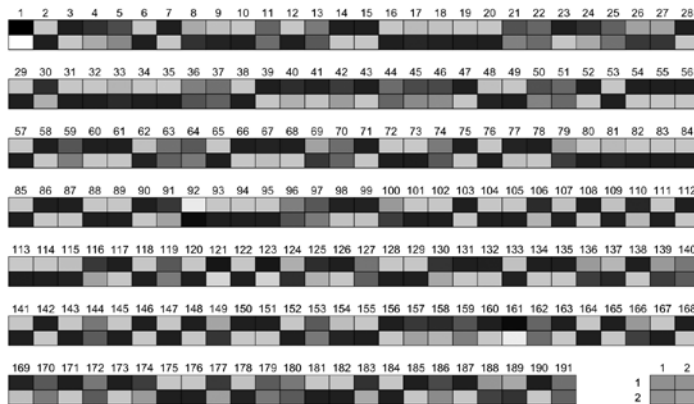
## DISCUSSION

Advances in methodology and software availability have allowed inferential network analysis to increasingly be used in political science. Inferential network analysis has opened the door to the empirical evaluation of new types of theoretical claims and improved the rigor with which political scientists approach dyadic data analysis. With the inherent relational structure of much data in political science, especially international

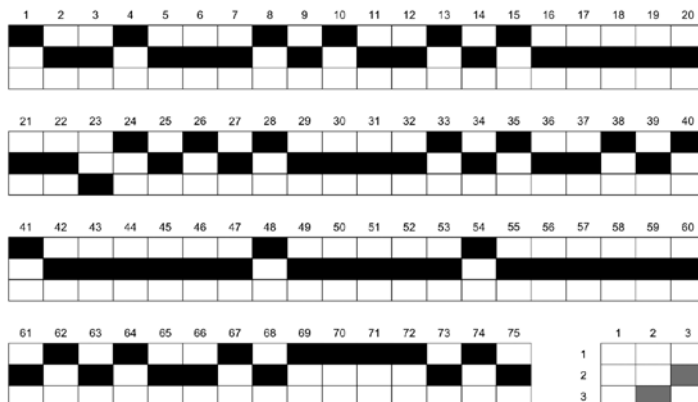
### Model 1



### Model 2



### Model 3



**Table 46.5 SBM block assignment plots. Each column represents one node and the rows are the block possibilities for each node. The blocks are shaded from white to black for the probability of block membership, with white representing a probability of approximately zero and black representing a likelihood of approximately one. The bottom right box is shaded according to the value of the block matrix value**

relations, inferential network analysis will continue to make an impact in political science methodology.

We hope that this chapter will help guide scholars in comparing and selecting available inferential network analysis tools, and understanding the methodology behind them, as well as their usefulness and their limitations. In this chapter, we have reviewed three of the most commonly used statistical network models, but it should be noted that the toolkit for inferential network analysis is broader than what we have reviewed and is constantly growing. As for the application, it provides a brief overview of network descriptive statistics and model specification options. There are many other useful network descriptive measures not included that scholars can employ, as well as other dependence terms to include in model specification, especially when using the ERGM. Furthermore, data-driven model selection across modeling frameworks presents an important and open methodological problem in political science. Lastly, regarding research software, R is the software of choice for inferential network analysis and contains a vast library of packages for static, time-dependent, and weighted networks. Other software packages for network analysis, albeit with limited functionality for inferential methods, include the Python library NetworkX (Hagberg et al., 2008), the Java-based program Gephi (Bastian et al., 2009), and user-friendly GUI-based software like PNet (Wang et al., 2009) and UCINET (Borgatti et al., 2014).

## Note

- 1 Network data, just like most types of data, are often found to be missing data. Dealing with this is an active and promising area of research. For example, for the ERGM model, researchers have had success using an ERGM-based imputation (Wang et al., 2016). Additionally, from the LSM and SBM, it is straightforward to estimate the probability of a tie from partial datasets, from which missing edge data can be imputed.

## REFERENCES

- Adhikari, Samrachana, Beau Dabbs, Brian Junker, Mauricio Sadinle, Tracy Sweet and A C Thomas. 2015. *CIDnetworks: Generative Models for Complex Networks with Conditionally Independent Dyadic Structure*. R package version 0.8.1. URL: <https://CRAN.R-project.org/package=CIDnetworks>.
- Aicher, Christopher, Abigail Z Jacobs and Aaron Clauset. 2014. Learning latent block structure in weighted networks. *Journal of Complex Networks* 3(2): 221–248.
- Asal, Victor H, Hyun Hee Park, R Karl Rethemeyer and Gary Ackerman. 2016. With friends like these... Why terrorist organizations ally. *International Public Management Journal* 19(1): 1–30.
- Atouba, Yannick C and Michelle Shumate. 2015. International nonprofit collaboration: Examining the role of homophily. *Nonprofit and Voluntary Sector Quarterly* 44(3): 587–608.
- Baldassarri, Delia and Peter Bearman. 2007. Dynamics of political polarization. *American sociological review* 72(5): 784–811.
- Baller, Inger. 2017. Specialists, party members, or national representatives: Patterns in co-sponsorship of amendments in the European Parliament. *European Union Politics* 18(3): 469–490.
- Bastian, Mathieu, Sebastien Heymann and Mathieu Jacomy. 2009. Gephi: An open source software for exploring and manipulating networks. In *Third International AAAI Conference on Weblogs and Social Media*.
- Bond, Robert M, Christopher J Fariss, Jason J Jones, Adam DI Kramer, Cameron Marlow, Jaime E Settle and James H Fowler. 2012. A 61-million-person experiment in social influence and political mobilization. *Nature* 489(7415): 295.
- Borgatti, Stephen P, Martin G Everett and Linton C Freeman. 2014. Ucinet. In Reda Alhajj and Jon Rokne (eds.) *Encyclopedia of social network analysis and mining*. New York: Springer, pp. 2261–2267.
- Box-Steffensmeier, Janet M and Dino P Christenson. 2014. The evolution and formation of amicus curiae networks. *Social Networks* 36: 82–96.

- Brandenberger, Laurence. 2018. Trading favors? Examining the temporal dynamics of reciprocity in congressional collaborations using relational event models. *Social Networks* 54: 238–253.
- Bratton, Kathleen A and Stella M Rouse. 2011. Networks in the legislative arena: How group dynamics affect cosponsorship. *Legislative Studies Quarterly* 36(3): 423–460.
- Bush, Stacy and Gisela Bichler. 2015. Measuring disruption in terrorist communications. Disrupting criminal networks: network analysis in crime prevention. *Crime Prevention Studies* 28: 177–208.
- Cao, Xun. 2012. Global networks and domestic policy convergence: A network explanation of policy changes. *World Politics* 64(3): 375–425.
- Carrington, Peter J, John Scott and Stanley Wasserman. 2005. *Models and Methods in Social Network Analysis*. Vol. 28, New York: Cambridge university press.
- Christopoulos, Dimitris and Karin Ingold. 2015. Exceptional or just well connected? Political entrepreneurs and brokers in policy making. *European Political Science Review* 7(3): 475–498.
- Chu-Shore, Jesse. 2010. Homogenization and specialization effects of international trade: Are cultural goods exceptional? *World Development* 38(1): 37–47.
- Chyzh, Olga. 2016. Dangerous liaisons: An endogenous model of international trade and human rights. *Journal of Peace Research* 53(3): 409–423.
- Clark, Jennifer Hayes and Veronica Caro. 2013. Multimember districts and the substantive representation of women: An analysis of legislative cosponsorship networks. *Politics & Gender* 9(1): 1–30.
- Cranmer, Skyler J and Bruce A Desmarais. 2011. Inferential network analysis with exponential random graph models. *Political Analysis* 19(1): 66–86.
- Cranmer, Skyler J, Bruce A Desmarais and Elizabeth J Menninga. 2012. Complex dependencies in the alliance network. *Conflict Management and Peace Science* 29(3): 279–313.
- Cranmer, Skyler J, Elizabeth J Menninga and Peter J Mucha. 2015. Kantian fractionalization predicts the conflict propensity of the international system. *Proceedings of the National Academy of Sciences* 112(38): 11812–11816.
- Cranmer, Skyler J, Tobias Heinrich and Bruce A Desmarais. 2014. Reciprocity and the structural determinants of the international sanctions network. *Social Networks* 36: 5–22.
- Davis, Christina L and Tyler Pratt. 2016. *The Forces of Attraction: How Security Interests Shape Membership in Economic Institutions*. Paper presented at the annual meeting of the American Political Science Association Annual Meeting, Philadelphia. Available at [https://scholar.harvard.edu/files/cldavis/files/davispratt\\_forces\\_july2018.pdf](https://scholar.harvard.edu/files/cldavis/files/davispratt_forces_july2018.pdf) (Accessed 15 January, 2020).
- DeMuse, Ryan, Danielle Larcomb and Mei Yin. 2018. Phase transitions in edge-weighted exponential random graphs: Near-degeneracy and universality. *Journal of Statistical Physics* 171(1): 127–144.
- Desmarais, Bruce A, Raymond J La Raja and Michael S Kowal. 2015. The fates of challengers in US house elections: The role of extended party networks in supporting candidates and shaping electoral outcomes. *American Journal of Political Science* 59(1): 194–211.
- Desmarais, Bruce A and Skyler J Cranmer. 2012. Statistical inference for valued-edge networks: The generalized exponential random graph model. *PLoS one* 7(1): e30136.
- Desmarais, Bruce A and Skyler J Cranmer. 2013. Forecasting the locational dynamics of transnational terrorism: A network analytic approach. *Security Informatics* 2(1): 8.
- Dorff, Cassy and Michael D Ward. 2013. Networks, dyads, and the social relations model. *Political Science Research and Methods* 1(2): 159–178.
- Dorff, Cassy and Shahryar Minhas. 2017. When do states say uncle? Network dependence and sanction compliance. *International Interactions* 43(4): 563–588.
- Dorff, Cassy, Shahryar Minhas and Michael D Ward. 2016. Latent networks and spatial networks in politics. In Victor, Jennifer Nicoll, Alexander H. Montgomery and Mark Lubell, (eds.) *The Oxford Handbook of Political Networks*. New York: Oxford University Press, pp. 249–276.

- DuBois, Christopher, Carter Butts and Padhraic Smyth. 2013. Stochastic blockmodeling of relational event dynamics. *Proceedings of Machine Learning Research* 31: 238–246.
- Duque, Marina G. 2018. Recognizing international status: A relational approach. *International Studies Quarterly* 62(3): 577–592.
- Fagiolo, Giorgio and Marina Mastrorillo. 2014. Does human migration affect international trade? A complex-network perspective. *PLoS one* 9(5): e97331.
- Freelon, Deen, Marc Lynch and Sean Aday. 2015. Online fragmentation in wartime: A longitudinal analysis of tweets about Syria, 2011–2013. *The ANNALS of the American Academy of Political and Social Science* 659(1): 166–179.
- Gallop, Max B. 2016. Endogenous networks and international cooperation. *Journal of Peace Research* 53(3): 310–324.
- Gerber, Elisabeth R, Adam Douglas Henry and Mark Lubell. 2013. Political homophily and collaboration in regional planning networks. *American Journal of Political Science* 57(3): 598–610.
- Guerrero, Angela M, Ryan RJ Mcallister and Kerrie A Wilson. 2015. Achieving cross-scale collaboration for large scale conservation initiatives. *Conservation Letters* 8(2): 107–117.
- Hafner-Burton, Emilie M, Miles Kahler and Alexander H Montgomery. 2009. Network analysis for international relations. *International Organization* 63(3): 559–592.
- Hagberg, Aric, Pieter Swart and Daniel S Chult. 2008. *Exploring Network Structure, Dynamics, and Function Using NetworkX*. Technical report Los Alamos National Lab (LANL): Los Alamos, NM.
- Handcock, Mark S., David R. Hunter, Carter T. Butts, Steven M. Goodreau, Pavel N. Krivitsky and Martina Morris. 2018. *ergm: Fit, Simulate and Diagnose Exponential-Family Models for Networks*. The Statnet Project (<http://www.statnet.org>). R package version 3.9.4. URL: <https://CRAN.R-project.org/package=ergm>
- Hanneke, Steve, Wenjie Fu, Eric P Xing et al. 2010. Discrete temporal models of social networks. *Electronic Journal of Statistics* 4: 585–605.
- Harris, Jenine K. 2013. Communication ties across the national network of local health departments. *American Journal of Preventive Medicine* 44(3): 247–253.
- He, Ran and Tian Zheng. 2016. Estimating exponential random graph models using sampled network data via graphon. In Ravi K., James C. and Hanghang T. (eds.) *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. San Francisco, CA: pp. 112–119. Available at <https://dl.acm.org/action/showFmPdf?doi=10.5555%2F3192424> (Accessed on 15 January, 2020).
- Hoff, Peter, Bailey Fosdick, Alex Volfovsky and Yanjun He. 2017. *amen: Additive and Multiplicative Effects Models for Networks and Relational Data*. R package version 1.3. URL: <https://CRAN.R-project.org/package=amen>
- Hummel, Ruth M, David R Hunter and Mark S Handcock. 2012. Improving simulation-based algorithms for fitting ERGMs. *Journal of Computational and Graphical Statistics* 21(4): 920–939.
- Hunter, David R, Mark S Handcock, Carter T Butts, Steven M Goodreau and Martina Morris. 2008. *ergm: A package to fit, simulate and diagnose exponential-family models for networks*. *Journal of Statistical Software* 24(3): nihpa54860.
- Kinne, Brandon J. 2018. Defense cooperation agreements and the emergence of a global security network. *International Organization* 72(4): 799–837.
- Kirkland, Justin H. 2012. ‘Multimember districts’ effect on collaboration between US state legislators. *Legislative Studies Quarterly* 37(3): 329–353.
- Kirkland, Justin H and R Lucas Williams. 2014. Partisanship and reciprocity in cross-chamber legislative interactions. *The Journal of Politics* 76(3): 754–769.
- Krafft, Peter, Justin Moore, Bruce Desmarais and Hanna M Wallach. 2012. Topic-partitioned multinet network embeddings. In Pereira F, Burges C, Bottou L and Weinberger K (eds.) *Advances in Neural Information Processing Systems*. Red Hook, NY: Curran Associates, Inc, pp. 2807–2815.
- Krivitsky, Pavel N. 2012. Exponential-family random graph models for valued networks. *Electronic Journal of Statistics* 6: 1100.
- Krivitsky, Pavel N and Carter T Butts. 2017. Exponential-family random graph models for

- rank-order relational data. *Sociological Methodology* 47(1): 68–112.
- Krivitsky, Pavel N. and Mark S. Handcock. 2017. *latentnet: Latent Position and Cluster Models for Statistical Networks*. The Statnet Project (<http://www.statnet.org>). R package version 2.8.0. URL: <https://CRAN.R-project.org/package=latentnet>
- Latouche, Pierre, Etienne Birmelé and Christophe Ambroise. 2011. Overlapping stochastic block models with application to the French political blogosphere. *The Annals of Applied Statistics* 5(1): 309–336.
- Li, Weihua, Aisha E Bradshaw, Caitlin B Clary and Skyler J Cranmer. 2017. A three-degree horizon of peace in the military alliance network. *Science Advances* 3(3): e1601895.
- Liu, Yang-Yu, Jean-Jacques Slotine and Albert-László Barabási. 2012. Control centrality and hierarchical structure in complex networks. *Plos One* 7(9): e44459.
- Lubbers, Miranda J and Tom A B Snijders. 2007. A comparison of various approaches to the exponential random graph model: A reanalysis of 102 student networks in school classes. *Social Networks* 29(4): 489–507.
- Luke, Douglas A. 2015. *A User's Guide to Network Analysis in R*. London: Springer.
- Lupu, Yonatan and Brian Greenhill. 2017. The networked peace: Intergovernmental organizations and international conflict. *Journal of Peace Research* 54(6): 833–848.
- Mahmood, Ammara and Catarina Sismeyro. 2017. Will they come and will they stay? Online social networks and news consumption on external websites. *Journal of Interactive Marketing* 37: 117–132.
- Majó-Vázquez, Sílvia, Rasmus K Nielsen and Sandra González-Bailón. 2019. The backbone structure of audience networks: A new approach to comparing online news consumption across countries. *Political Communication* 36(2): 227–240.
- Metternich, Nils W, Cassy Dorff, Max Gallop, Simon Weschle and Michael D Ward. 2013. Antigovernment networks in civil conflicts: How network structures affect conflictual behavior. *American Journal of Political Science* 57(4): 892–911.
- Milewicz, Karolina, James Hollway, Claire Peacock and Duncan Snidal. 2018. Beyond trade: The expanding scope of the nontrade agenda in trade agreements. *Journal of Conflict Resolution* 62(4): 743–773.
- Ogburn, Elizabeth L. 2018. Challenges to estimating contagion effects from observational data. In Lehmann, S. and Yong-Yeol A. (eds.) *Complex Spreading Phenomena in Social Systems*. New York: Springer, pp. 47–64.
- Osei, Anja and Thomas Malang. 2018. Party, ethnicity, or region? Determinants of informal political exchange in the parliament of Ghana. *Party Politics* 24(4): 410–420.
- Peng, Tai-Quan, Mengchen Liu, Yingcai Wu and Shixia Liu. 2016. Follower-follower network, communication networks, and vote agreement of the US members of congress. *Communication Research* 43(7): 996–1024.
- Perry, Patrick O and Patrick J Wolfe. 2013. Point process modelling for directed interaction networks. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 75(5): 821–849.
- Raftery, Adrian E, Xiaoyue Niu, Peter D Hoff and Ka Yee Yeung. 2012. Fast inference for the latent space network model using a case-control approximate likelihood. *Journal of Computational and Graphical Statistics* 21(4): 901–919.
- Ringe, Nils, Jennifer N Victor and Christopher J Carman. 2013. *Bridging the Information Gap: Legislative Member Organizations as Social Networks in the United States and the European Union*. Ann Arbor (MI): University of Michigan Press.
- Robinson, Lucy F, Lauren Y Atlas and Tor D Wager. 2015. Dynamic functional connectivity using state-based dynamic community structure: Method and application to opioid analgesia. *NeuroImage* 108: 274–291.
- Schmid, Christian S and Bruce A Desmarais. 2017. Exponential random graph models with big networks: Maximum pseudolikelihood estimation and the parametric bootstrap. In *Big Data (Big Data), 2017 IEEE International Conference on*. IEEE. Boston, MA, pp. 116–121.
- Schweinberger, Michael. 2011. Instability, sensitivity, and degeneracy of discrete exponential families. *Journal of the American Statistical Association* 106(496): 1361–1370.
- Sewell, Daniel K and Yuguo Chen. 2015. Latent space models for dynamic networks. *Journal of the American Statistical Association* 110(512): 1646–1657.

- Signorelli, Mirko and Ernst C Wit. 2018. A penalized inference approach to stochastic block modelling of community structure in the Italian Parliament. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 67(2): 355–369.
- Sohn, Yunkyu and Jong Hee Park. 2017. Bayesian approach to multilayer stochastic block-model and network changepoint detection. *Network Science* 5(2): 164–186.
- Sweet, Tracy M. 2015. Incorporating covariates into stochastic blockmodels. *Journal of Educational and Behavioral Statistics* 40(6): 635–664.
- Vance, Eric A, Elizabeth A Archie and Cynthia J Moss. 2009. Social networks in African elephants. *Computational and Mathematical Organization Theory* 15(4): 273.
- Wang, Cheng, Carter T Butts, John R Hipp, Rupa Jose and Cynthia M Lakon. 2016. Multiple imputation for missing edge data: a predictive evaluation method with application to add health. *Social Networks* 45: 89–98.
- Wang, Lan. 2009. Wilcoxon-type generalized Bayesian information criterion. *Biometrika* 96(1): 163–173.
- Wang, Peng, Garry Robins and Philippa Pattison. 2009. PNet. *Program for the Simulation and Estimation of Exponential Random Graph ( $p^*$ ) Models: User Manual*. Melbourne: University of Melbourne.
- Ward, Michael D and Peter D Hoff. 2007. Persistent patterns of international commerce. *Journal of Peace Research* 44(2): 157–175.
- Ward, Michael D, Randolph M Siverson and Xun Cao. 2007. Disputes, democracies, and dependencies: A reexamination of the Kantian peace. *American Journal of Political Science* 51(3): 583–601.
- Wojcik, Stefan. 2017. Why legislative networks? Analyzing legislative network formation. *Political Science Research and Methods* 7(3): 1–18.
- Wyatt, Danny, Tanzeem Choudhury and Jeff Bilmes. 2009. Dynamic multi-valued network models for predicting face-to-face conversations. In *NIPS workshop on Analyzing Networks and Learning with Graphs*. Vancouver, Canada.
- Xu, Kevin S and Alfred O Hero. 2014. Dynamic stochastic blockmodels for time-evolving social networks. *IEEE Journal of Selected Topics in Signal Processing* 8(4): 552–562.
- Zhu, Bin and Stephanie A Watts. 2010. Visualization of network concepts: The impact of working memory capacity differences. *Information Systems Research* 21(2): 327–344.



# Bayesian Methods in Political Science

Jong Hee Park and Sooahn Shin

## INTRODUCTION

Bayesian methods refer to a suite of statistical methods that consistently use Bayes' formula to update researchers' beliefs about statistical quantities of interest using observed data. The consistent use of Bayes' formula requires two major inputs – a prior distribution and a generative model of (observed and unobserved) data – and an efficient estimation method that can find statistical quantities of a new probability density (called a posterior distribution) from the product of multiple probability densities.

The beauty of Bayesian methods lies in the fact that if researchers have a theory of a real-world process that can be represented by a generative statistical model, Bayesian inference is the most consistent way to represent this belief using the theory of probability. This point is formally proved by Bruno de Finetti (1974). The re-discovery of Markov chain Monte Carlo (MCMC) methods in the 1970s and 1980s made it possible to

materialize de Finetti's theoretical conjecture in statistical inference.

Political scientists have been interested in the idea of Bayesian inference since the 1950s. At the time, political scientists applied Bayesian philosophy to decision problems following prominent work by Savage (1954). Bayesian inference as a data analytic methodology had to wait until the rediscovery of MCMC methods in the 1970s and 1980s, and the spread of personal computers in the 1990s. In the 1990s, Bayesian inference emerged as a vibrant research area in many different fields of science and the number of publications using Bayesian inference has exploded since then.

One distinct aspect of Bayesian inference in political science is that major developments have taken the form of 'model developments' accompanied by computational improvement. This parallel development in model development and computational improvement is due, in part, to the advent of powerful open-source software such as R (R Core Team, 2018), BUGS, python, and



STAN. This parallel development enhances the transparency of Bayesian methods, facilitates the wide usage of Bayesian methods in applied research, and helps methodologists develop a more sophisticated Bayesian method. During this progress, political methodologists find various Bayesian techniques such as data augmentation, parameter expansion, Gibbs sampling, Metropolis–Hasting sampling, and variational inference to be highly useful to address methodological challenges political scientists encounter. The most notable applications of the Bayesian technique in political science include roll-call data analysis, survey data analysis, discrete data analysis, time series cross sectional data analysis, network analysis, and text analysis.

Despite many impressive innovations, Bayesian inference in political science faces new challenges. The most critical challenge comes from big data and subsequent developments in machine learning/artificial intelligence methods. How does Bayesian inference adapt to the explosion of data size and model complexity while properly incorporating the uncertainty involved in model selection and parameter estimation? How does Bayesian computation satisfy the demand for scalability without sacrificing advantages of a fully probability inference? Although there is no question that Bayesian inference will continue to provide a consistent and complete answer to many problems political scientists face, a certain level of adaptation would be required to address these new inferential, as well as computational, challenges.

## WHAT IS A BAYESIAN INFERENCE?

A concise definition of Bayesian inference can be found in Gelman et al. (2012: 1): ‘Bayesian inference is the process of fitting a probability model to a set of data and summarizing the result by a probability distribution on the parameters of the model and on unobserved quantities such as predictions for

new observations’. A longer definition can be found in *Encyclopædia Britannica* (2016):

a method of statistical inference that allows one to combine prior information about a population parameter with evidence from information contained in a sample to guide the statistical inference process. A prior probability distribution for a parameter of interest is specified first. The evidence is then obtained and combined through an application of Bayes's theorem to provide a posterior probability distribution for the parameter. The posterior distribution provides the basis for statistical inferences concerning the parameter.

In both definitions, the central element of Bayesian inference lies in its consistent use of probability distribution in statistical inference.

The consistent nature of Bayesian inference arises from de Finetti's theorem. de Finetti shows the existence of a general representation form of Bayesian inference to any type of exchangeable observations. Let  $\theta$  be a parameter vector,  $y$  be a response vector,  $p(y_i|\theta)$  be a density of data  $i$  given  $\theta$ , and  $p(\theta)$  be a distribution on  $\theta$ . de Finetti's theorem says:

- de Finetti's Theorem: *For exchangeable random variables  $y_1, y_2, \dots, y_n$ ,*

$$p(y_1, y_2, \dots, y_n) = \int \prod_{i=1}^n p(y_i | \theta) p(\theta) d\theta.$$

de Finetti's theorem tells us that if we want to learn about data using parameters, we need to put some form of distribution on  $\theta$ . Thus, de Finetti's theorem justifies the hierarchical structure (i.e., parameters in the likelihood function first come from their own distributions and then follow the distribution of data) of Bayesian inference in its simplest form.

Bayesian inference uses Bayes theorem to learn parameter distributions after seeing data:

$$\underbrace{p(\theta | y, \mathcal{M})}_{\text{posterior distribution}} = \frac{\overbrace{p(y | \theta, \mathcal{M})}^{\text{likelihood}} \overbrace{p(\theta | \mathcal{M})}^{\text{prior}}}{\underbrace{p(y | \mathcal{M})}_{\text{marginal density of data}}}.$$

The above notation is different from a conventional Bayes theorem notation. We specifically denote the fact that all (model) parameters are dependent upon a certain data generative model ( $\mathcal{M}$ ). This expanded notation is highly useful when we think about what prior density, marginal density of data, and the probability of a model are.

First, prior density depends either directly or indirectly on the chosen data generative model ( $p(\theta|\mathcal{M})$ ). When prior density belongs to the same family of distribution with data density, the whole inference becomes straightforward, the property of which is called *conjugacy*. Second, the data density (or likelihood) is directly dependent upon the chosen parametric model ( $p(y|\theta, \mathcal{M})$ ). The density of data is read as the probability of observing data given a specific parametric form of a data generative model. Last, the marginal density of data is independent of the chosen parametric form but is not independent from a data generative model:  $p(y|\mathcal{M})$ . Thus, the marginal density of data should be read as the probability of observing data given a chosen data generative model. Since we want to know about  $\theta$  in most cases,  $p(y|\mathcal{M})$  is often ignored except when we assess model evidence given data ( $p(\mathcal{M}|y)$ ).

To summarize, Bayesian inference consists of the following probability distributions:

- $p(y|\theta, \mathcal{M})$  is a likelihood function or the density of data given parameters and a specific data generative model  $\mathcal{M}$ . This is a major concern of interest in model development.
- $p(\theta|\mathcal{M})$  is a prior distribution of parameters arising from a specific data generative model  $\mathcal{M}$ . We usually choose prior distributions of parameters in conjunction with  $p(y|\theta, \mathcal{M})$  in model development.
- $p(\theta|y, \mathcal{M})$  is a posterior distribution that represents our updated belief about the distribution of  $\theta$  given a specific data generative model  $\mathcal{M}$  and data  $y$ . This is the major quantity of interest in model fitting.
- $p(y|\mathcal{M})$  is a conditional density of data given a specific data generative model. It is also called

model evidence or marginal likelihood.  $p(y|\mathcal{M})$  approximates the probability of observing actually observed data over the entire prior distribution given a specific data generative model  $\mathcal{M}$ . Thus, without specifying a finite sample space for a prior probability distribution, it is difficult to define and interpret this quantity.  $p(y|\mathcal{M})$  is different from  $p(y)$ , which is an *unconditional* density of data and serves just as a normalizing constant in Bayesian model comparison.

- $p(\mathcal{M}|y)$  is a probability distribution of a specific data generative model. When we compare different models, this is the quantity we would like to compare different data generative models.

We can learn about all the above-mentioned quantities using the law of probability, a family of probability distributions, and data. For example, if researchers wish to know about predictive distributions of data  $p(y^{\text{pred}}|y)$ ,

$$\begin{aligned} p(y^{\text{pred}}|y, \mathcal{M}) &= \int p(y^{\text{pred}}, \theta|y, \mathcal{M}) d\theta \\ &= \int p(y^{\text{pred}} | \theta, y, \mathcal{M}) p(\theta | y, \mathcal{M}) d\theta \\ &= \int p(y^{\text{pred}} | \theta, \mathcal{M}) p(\theta | y, \mathcal{M}) d\theta. \end{aligned}$$

If we replace  $y^{\text{pred}}$  by  $y^{\text{missing}}$  and  $y$  by  $y^{\text{observed}}$ , the above formula is equivalent to a missing data imputation method.

If one wishes to assess the model probability of  $\mathcal{M}$ ,

$$p(\mathcal{M}|y) = \frac{p(y|\mathcal{M})p(\mathcal{M})}{p(y)} \propto p(y|\mathcal{M})p(\mathcal{M}).$$

The rightmost approximation indicates that  $p(y)$ , the unconditional data density, is just a normalizing constant and hence it can be ignored in the computation. Another interesting result from the above formula is that if we assume the constant probability for each  $\mathcal{M}$ , comparing  $p(y|\mathcal{M})$  is sufficient to check the model probability.

In many cases of political science research, a statistical model contains multiple parameters, and it is often difficult to sample directly from the joint posterior distribution. In such cases, there exists two general MCMC methods for approximating the posterior distribution: the *Metropolis–Hastings (MH) algorithm* and the *Gibbs sampler*. A probabilistic MCMC algorithm should satisfy the local reversibility condition (or the detailed balance) to make the chain converge to an invariant distribution:

$$p(\theta, \theta^*)f(\theta) = p(\theta^*, \theta)f(\theta^*).$$

The MH algorithm takes advantage of the idea that we can make irreversible transition kernels reversible by multiplying a proper acceptance rate. That is,

$$p(\theta, \theta^*)f(\theta)\alpha(\theta^*, \theta) = p(\theta^*, \theta)f(\theta^*).$$

Then, the acceptance rate is

$$\alpha(\theta^*, \theta) = \min \left[ \frac{p(\theta, \theta^*)f(\theta)}{p(\theta^*, \theta)f(\theta^*)}, 1 \right].$$

This idea can be applied to any posterior distribution. When we can directly sample from the full conditional distributions (i.e., the full conditional distributions have a known distributional form), the proposal distribution becomes the full conditional posterior distributions:

$$\alpha(\theta^*, \theta) = \min \left[ \frac{p(\theta^*)p(\theta)}{p(\theta)p(\theta^*)}, 1 \right] = 1.$$

The implementation of the Gibbs sampler is very similar to the MH algorithm except from the fact that the acceptance rate does not need to be computed because the full conditional distributions follow a known distributional form. Suppose a statistical model includes parameters  $\theta_1, \dots, \theta_k$  and let  $\theta$  denote the vector of length  $k$  that consists of all the model parameters. Here, we want to obtain  $n$  samples from the joint posterior distribution

$p(\theta|y)$ . The Gibbs sampler generates samples by following steps:

- 1 Initialize with  $\theta^{(0)} = (\theta_1^{(0)}, \dots, \theta_k^{(0)})$
- 2 For  $i = 1, \dots, n$  repeat: sample  $\theta_k^{(i)} \sim p(\theta_k | \theta_1^{(i)}, \dots, \theta_{k-1}^{(i-1)}, y)$

The samples generated by this algorithm constitute a dependent sequence of vectors  $\{\theta^{(0)}, \dots, \theta^{(n)}\}$ . Note that the probability of each vector depends only on the previous vector; that is, it is a Markov chain. Under some conditions, those samples approximate the joint posterior distribution.

Because of the computational simplicity, the Gibbs sampler is one of the most widely used MCMC methods. In particular, the Gibbs sampler becomes a major workhorse of Bayesian inference with the popularity of the data augmentation method that transforms an intractable conditional posterior distribution into a tractable one by augmenting latent parameters (Tanner and Wong, 1987; van Dyk and Meng, 2001).

Although the MH algorithm and the Gibbs sampler work well in most statistical models, they often show poor performance in high-dimensional models with many parameters. Many scholars have proposed various MCMC methods to improve the performance of existing MCMC methods. Examples include slice sampling, parallel tempering, particle filtering, and Hamiltonian Monte Carlo methods (see Gordon et al., 1993; Neal, 2003; Earl and Deem, 2005; Neal, 2011 for more details).

## A SHORT HISTORY OF BAYESIAN INFERENCE IN POLITICAL SCIENCE

In this section, we briefly review academic publications on Bayesian methods in political science with a focus on journal articles and open source software contributions. One caveat in this analysis is that we apply the

territorial principle, instead of the personal principle, in judging what constitutes ‘Bayesian inference in political science’. That is, we delimit our discussion to published work in political science journals, omitting a body of work published outside political science journals.

### **Journal Publications**

The word ‘Bayes formula’ appeared in political science publications as early as 1950 (Girshick and Lerner, 1950). However, between the 1950s and 1980s, Bayesian inference meant a decision theoretic analysis following the classical work of Savage (1954). The only exception is Achen (1978), who used the modern form of Bayesian inference to theoretically justify his measures.

The beginning of Bayesian analysis as a data analytical tool in political science was started by King and Gelman (1991). In this paper, King and Gelman show how Bayesian inference can be applied to estimate incumbency advantage in US Congressional elections. First, they showed why a probabilistic model should be preferred to deterministic methods. A probabilistic model is less susceptible to arbitrary decisions, allows our inference to be more realistic, and provides measures of uncertainty. Then, they built a Bayesian hierarchical model of district vote shares capturing variations across congressional districts and election years. In this way, the paper demonstrates how Bayesian inference can be used to quantify substantively important quantities of interest in political science such as partisan bias, electoral responsiveness, and incumbent advantages.

Since the publication of King and Gelman (1991), papers that use or develop Bayesian methods have increased dramatically in political science journals. For example, Jackman (1994) developed a Bayesian measurement model of electoral bias and responsiveness using summaries of posterior distributions.

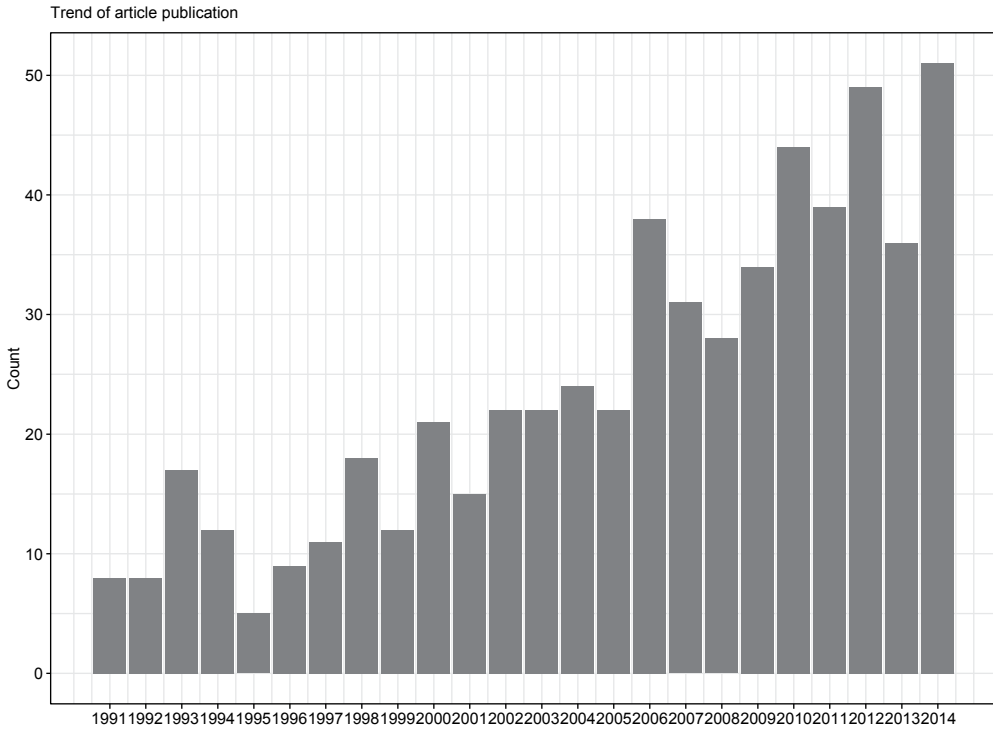
Figure 47.1 shows the number of political science journal articles between 1991 and 2014 containing ‘Bayesian’ or ‘Bayes’ as a keyword. In this search, we excluded game theory papers. We also limited our search up to 2014 because the Jstor engine ([www.jstor.org](http://www.jstor.org)) does not have publication information for some journals after 2014. Figure 47.1 clearly demonstrates that the number of ‘Bayesian’ publications increased dramatically during the last two decades.

Table 47.1 summarizes the Bayesian publication between 1991 and 2014 by journals. *Political Analysis* is ranked at the top (137 articles), followed closely by the *American Journal of Political Science* (116 articles). *American Political Science Review*, the top flagship journal of political science, has as many as 51 publications during this period. Overall, the publication records clearly showed that Bayesian methods have been well established as a major data analytic paradigm during the last two decades.

Then, what types of topics have been covered by those publications? Table 47.2 summarizes the topics of Bayesian publications between 1991 and 2014. We fit a Latent Dirichlet Allocation model to a text dataset of journal abstracts for all the Bayesian publications between 1991 and 2014. Topic 1 represents Bayesian inference of voting behavior and electoral studies. Topic 2 covers political economy and public policy literature. Topic 3 concerns technical papers and time series analysis using Bayesian methods. Topic 4 corresponds to international politics applications of Bayesian inference. Topic 2 includes Bayesian inference of opinion poll analysis. Overall, the range of topics clearly shows that Bayesian methods have been widely applied to various sub fields in political science.

### **Software**

Another way to review a history of Bayesian inference in political science is to



**Figure 47.1** The number of journal articles containing Bayesian as a keyword excluding game theory papers. The search was done for journal articles between 1991 and 2014 using the JSTOR search engine

**Table 47.1** Summary by journals: the number of journal articles containing 'Bayesian' in the abstract excluding game theory-related articles

Journal	Count
<i>Political Analysis</i>	137
<i>American Journal of Political Science</i>	116
<i>The American Political Science Review</i>	57
<i>The Journal of Politics</i>	47
<i>The Journal of Conflict Resolution</i>	38
<i>International Studies Quarterly</i>	32
<i>Political Research Quarterly</i>	29
<i>PS: Political Science and Politics</i>	29
<i>International Organization</i>	25
<i>The Public Opinion Quarterly</i>	25
<i>British Journal of Political Science</i>	23
<i>Legislative Studies Quarterly</i>	12
<i>Comparative Politics</i>	4
<i>Political Science Quarterly</i>	2

look at software developments. We focus on R packages contributed by political scientists to CRAN (2018). To identify Bayesian packages, we used the information in Bayesian Inference Task View (<https://cran.r-project.org/web/views/Bayesian.html>) and individual package descriptions and function contributions in package documents. Some packages cover a wide range of models and have multiple contributors.

As of September 2018, we identified 10 Bayesian packages contributed by political scientists to CRAN. The Bayesian software contributions in R do not closely match the increasing number of Bayesian publications in political science journals. One reason may be that political scientists prefer online version control platforms such as github for code sharing and repository instead of developing

**Table 47.2 Summary by topics: top 15 words sorted by the topic-word probability using text2vec package (Selivanov and Wang, 2018)**

<i>Topic 1</i>	<i>Topic 2</i>	<i>Topic 3</i>	<i>Topic 4</i>	<i>Topic 5</i>
Elections	Research	Statistical	Economic	Public
Presidential	Policy	Estimators	Conflict	Opinion
Voter	Politics	Data	International	Social
Candidates	Government	Estimation	Democracy	Parties
Partisanship	Rates	Time	War	Court
Congressional	Behavior	Parametric	Theory	Polls
Legislators	Design	Series	Wars	Supreme
Electoral	Spending	Analysis	Relations	Point
Parties	Domestic	Inference	Resolution	Surveys
Campaigns	Empirical	Coefficients	Forecasting	Liberalism
States	Learning	Science	Civil	Control
Incumbents	Unemployment	Model	Military	Spatial
Turnout	Labor	Bias	Trade	Groups
Election	Evidence	Multilevel	Dyadic	Estimates
Party	Economic	Regression	Effects	Conservatism

**Table 47.3 Bayesian R Packages: Bayesian packages hosted in CRAN**

<i>R package</i>	<i>Description</i>	<i>Author</i>
BaM	Companion of Bayesian methods: A social and behavioral sciences approach	Jonathan Homola, Danielle Korman, Jacob Metz, Miguel Pereira, Mauricio Vela, and Jeff Gill
EBMAforecast	Ensemble Bayesian model averaging model	Jacob M. Montgomery, Florian M. Hollenbach, Michael D. Ward
eco	Bayesian methods for ecological inference	Kosuke Imai, Ying Lu, and Aaron Strauss
MCMCpack	MCMC algorithms for a wide range of models	Andrew D. Martin, Kevin M. Quinn, Jong Hee Park
MNP	Bayesian multinomial probit model	Kosuke Imai and David van Dyk
NetworkChange	Bayesian changepoint analysis of network models	Jong Hee Park and Yunkyu Sohn
pscl	Bayesian analysis of IRT models	Simon Jackman
rstan	R interface to Stan	Jiqiang Guo, Jonah Gabry, Ben Goodrich
SimpleTable	Bayesian inference for causal effects from $2 \times 2$ tables K	Kevin M. Quinn
sparsereg	Sparse Bayesian models for regression, subgroup analysis, and panel data	Marc Ratkovic and Dustin Tingley.

a stand-alone package. Another reason is that more and more political scientists are interested in applied Bayesian research using existing software instead of developing their own models/algorithms and companion packages.

Beside journal publications and software development, we briefly mention books on

Bayesian methods published by political scientists. First, Gill's (2007) *Bayesian Methods: A Social and Behavioral Sciences Approach* is the first general social science introduction to Bayesian methods with a companion R package (BaM). Later, Jackman (2009) published *Bayesian Analysis for the Social*

*Sciences* as a general reference to Bayesian methods for social sciences. Gelman et al.'s (2012) *Bayesian Data Analysis* is the most comprehensive reference to Bayesian methods beyond political science and social science audience.

## MAJOR DEVELOPMENTS IN BAYESIAN INFERENCE IN POLITICAL SCIENCE

In this section, we briefly review major developments of Bayesian methods in political science in last two decades. Given the limitation of this review, we had to choose a small number of published work that could be classified by the categories of measurement models, hierarchical models, time series models, and new developments.

### *Measurement Models*

The measurement model refers to a statistical model that infers latent variables using observed data. Among several Bayesian approaches to the measurement problem, item-response theory (IRT) models have been widely used by political scientists. In this section, we briefly overview why Bayesian inference has been so successful in ideal point estimation (Jackman, 2001; Martin and Quinn, 2002; Clinton et al., 2004; Quinn, 2004; Bafumi et al., 2005; Barberá, 2015; Imai et al., 2016) and how this framework was transformed to measure different latent variables in political science (Treier and Jackman, 2008; Pemstein et al., 2010).

### *Measuring Ideal Points*

There is no doubt that ideal point estimation is one of the most successful applications of

Bayesian methods in political science. We will forward readers to Chapter 48 of this *Handbook* for the details of Bayesian ideal point estimation method. Instead, here, we highlight three factors that contributed to the success of Bayesian inference in ideal point estimation.

The first factor is the introduction of IRT models as a data generative model ( $\mathcal{M}$ ) of legislative voting. IRT was originally developed in psychometrics as a dimension-reduction technique of high-dimensional testing results. Poole and Rosenthal (1997) was the first important groundwork that connected a statistical model with a spatial voting model in the MLE framework. However, the IRT model was re-discovered by Bayesian statisticians (Albert, 1992) and political scientists as a better statistical framework for modeling legislative voting. The close connection between parameters of the (two-parameter) IRT model and a spatial voting model makes the IRT model a natural choice. Clinton et al., (2004) present a standard Bayesian implementation of the IRT model for ideal point estimation.

The second factor is the computational power of MCMC algorithms. Albert (1992) and Clinton et al., (2004) provide an efficient Gibbs sampling algorithm to update all the parameters of the (two-parameter) IRT model. The MCMC estimation of the IRT model can deal with large data and provide proper measures of estimation uncertainty. Easy to use Bayesian software such as *pscl*, *BUGS*, and *MCMCpack* allows users to fit Bayesian IRT models to their data without much programming skill.

The third, and most important, factor for the success of Bayesian inference in ideal point estimation is its modeling flexibility. Unlike other non-Bayesian frameworks, Bayesian IRT models prove to be highly flexible to incorporate additional layers of model complexity. For example, Martin and Quinn (2002) added a dynamic linear process of ideal points within a Bayesian IRT model to estimate US Supreme Court

justices' ideal point changes over their service time. Martin and Quinn's (2002) model

can be understood as a special type of

two-parameter IRT model with prior densities of subject-specific, first-order Markov processes:

$$z_{ijk} = \beta_i \theta_{j,t} - \alpha_i + \varepsilon_{ijt}, \quad \varepsilon_{ijt} \sim N(0,1) \quad [\text{Likelihood}]$$

$$\theta_{j,t} \sim N(\theta_{j,t-1}, \sigma_j^2) \quad \text{for } t > 1 \text{ and } j = 1, \dots, J \quad [\text{First-order Markov process}]$$

Quinn (2004) also developed a Bayesian measurement model for mixed multivariate responses. The model estimates latent factor components from ordinal and continuous response data. Recently, Barberá (2015) developed a simple model of network ideal point where the Twitter user  $i$ 's latent propensity to follow politician  $j$ 's Twitter is a decreasing function of the distance between ideal points of Twitter user  $i$  and politician  $j$ .

corresponds to the *item discrimination parameter* in the sense that it indicates how much the item  $j$  helps in distinguishing different levels of democracy. Pemstein et al., (2010) also proposed a Bayesian approach for measuring democracy. They assume that rating  $t_{ij}$  is generated by the following process:  $t_{ij} = z_i + e_{ij}$  where  $z_i$  is the latent level of democracy in country-year  $i$ .

Schnakenberg and Fariss (2014) presented a dynamic ordinal IRT (DO-IRT) model to measure temporal dependence in human rights levels using jags. They showed that a dynamic measurement of human rights levels using Bayesian methods outperforms a static measurement model.

### Measuring Democracy and Human Rights Levels

Based on the success of Bayesian IRT models in ideal point estimation, Treier and Jackman (2008) developed the ordinal IRT model to measure the level of democracy from the Polity indicators. Let  $i = 1, \dots, n$  index country-years,  $j = 1, \dots, m$  index the Polity indicators, and  $k = 1, \dots, K_j$  index the (ordered) response categories for item  $j$  – then the proposed model is

$$\begin{aligned} \Pr(y_{ij} = 1) &= F(\tau_{j1} - x_i \beta_j) && \vdots \\ &\vdots && \vdots \\ \Pr(y_{ij} = k) &= F(\tau_{jk} - x_i \beta_j) - F(\tau_{j,k-1} - x_i \beta_j) && \vdots \\ &\vdots && \vdots \\ \Pr(y_{ij} = K_j) &= 1 - F(\tau_{j,K_j-1} - x_i \beta_j) \end{aligned}$$

where  $x_i$  is the latent level of democracy in country-year  $i$ ,  $y_{ij}$  is the  $i$ th country-year's score on indicator  $j$ ,  $\tau_j$  is a vector of unobserved thresholds for item  $j$ , and  $F(\cdot)$  is here defined as the logistic CDF; that is,  $F(\tau_{jk} - z_i) = 1/(1 + \exp(-\tau_{jk} + z_i))$ . Note that  $\beta_j$

### Hierarchical Models

The idea of Bayesian inference as a scientific research method can be appealing in many different ways. What we have emphasized so far was the consistent use of probability distributions in the estimation, prediction, and diagnostics of a model. Moreover, we have mentioned that the de Finetti theorem tells us that Bayesian inference is a highly general framework to model any type of exchangeable data.

Another important appeal of Bayesian inference comes from the property of Bayesian estimates. The first important property is that the mean of the posterior distribution minimizes the mean square error (MSE). When we use a conjugate prior, the posterior mean can be decomposed into a weighted average of prior information and information from data.



Let  $y_i|\theta \sim N(\theta, \sigma^2)$  and  $\theta \sim N(\mu, \tau^2)$ . Then, the posterior mean is

$$E(\theta | y) = \bar{y} + \omega(\mu - \bar{y})$$

where  $\omega = \frac{\tau^{-2}}{n\sigma^{-2} + \tau^{-2}}$ .  $\omega$  is a shrinkage factor that pushes the posterior mean toward the prior mean as it approaches to 1.  $\omega$  approaches 1 as the sample variance of data ( $\sigma^2/n$ ) approaches to infinity. Then, we learn almost nothing from data and  $E(\theta|y)$  becomes the prior mean ( $\mu$ ). Instead, as the sample variance of data decreases,  $\omega$  approaches 0 and  $E(\theta|y)$  becomes  $\bar{y}$ .

The same can be said to a multilevel (or hierarchical) model. Let  $y_{ij}|\theta_j \sim N(\theta_j, \sigma^2)$  for  $i = 1, \dots, n_j$  and  $j = 1, \dots, J$ .  $i$  is an individual unit and  $j$  is a group indicator containing multiple individuals. We assume  $\theta_j \sim N(\mu, \tau^2)$ . Therefore, the posterior mean for group  $j$  is

$$E(\theta_j | y, \mu, \tau^2) = \bar{y}_j + \omega_j(\mu - \bar{y}_j)$$

where  $\bar{y}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ij}$  and  $\omega_j = \frac{\tau^{-2}}{\sigma_j^{-2} + \tau^{-2}}$ . Again, the posterior mean is a weighted average of the prior mean and the sample mean of group  $j$ 's observations. Since we assume a known variance,  $\sigma_j^{-2} = \frac{\sigma^2}{n_j}$ . Then, as group  $j$ 's observations increase,  $\omega_j$  becomes smaller. In other words, the Bayesian mean estimate weighs prior mean and data mean based on the amount of information in each group.

This property of Bayesian estimates is known as a partial pooling or shrinkage estimation. As we have mentioned above, partial pooling or shrinkage estimates are the minimum MSE estimates in most cases and this property is particularly useful to make a statistical inference from complex data with multiple layers. For example, Western (1998) shows that Bayesian hierarchical models can be used to test a multiplicative hypothesis in social sciences as Bayesian hierarchical

models are a general version of a full interaction model. Gelman et al. (2007) use the idea of Bayesian partial pooling to illustrate the puzzle of switching signs of income and voting at the individual and state levels. Gelman et al. (2007) model individual vote choices to be affected by multi layered factors at the individual and state level. Then, they show that while individual income affects individual vote choices similarly across states, the association of average state income with state-level vote propensities varies significantly across states.

The method of partial pooling is particularly useful to aggregate multiple sources of polls into a national trend for election forecast and media bias estimation. For example, Jackman (2005) develops a hierarchical model of election polls where partially pooled estimates of pollster effects can be considered as pollster bias. Lock and Gelman (2010) also use Bayesian inference to forecast election outcomes using data collected from various sources (actual past election outcomes, pre-election polls, and model estimates).

Last, Hoff and Ward (2004) introduce a Bayesian hierarchical modeling approach to high-order network dependences. This hierarchical framework for network data significantly influences network analysis in international relations (Ward and Hoff, 2007, Minhas et al., 2016).

### ***Time-series Models***

Another important area in the development of Bayesian methods in political science is time series models. Although classical frequentist methods have dominated time series models for stationary processes in political science and economics for a long time, many scholars find Bayesian methods to be highly useful to model non-stationary stochastic processes in political science data. Martin and Quinn (2002) is the important early work that shows the power of Bayesian methods in

modeling temporal dynamics in a measurement model.

Bayesian time series developments in political science can be divided into three different areas. The first is change-point models (or discrete hidden Markov models) (Spirling, 2007; Park, 2010, 2011, 2012; Blackwell, 2018). The change-point problem can be stated as a problem to find break points of parameter heterogeneity in sequentially observed data. Let  $y = y_1, \dots, y_t$  be sequentially observed response data and  $x$  be a vector of covariates. Then,

$$y_t = \begin{cases} x'_t \beta_1 + \varepsilon_t, & \varepsilon_t \sim N(0, \sigma_1^2) & \text{for } t_0 \leq t < \tau_1 \\ \vdots & \vdots & \vdots \\ x'_t \beta_M + \varepsilon_t, & \varepsilon_t \sim N(0, \sigma_M^2) & \text{for } \tau_{M-1} \leq t < T. \end{cases}$$

The change-point problem is to find  $\tau = \{\tau_1, \dots, \tau_{M-1}\}$  for  $M$  changepoints and estimate regime-specific model parameters. Western and Kleykamp (2004) first showed the appeal of change-point analysis in political science research and provided a simple BUGS code to fit a single change-point model based on Carlin et al. (1992). Spirling (2007) further extended Carlin et al.'s (1992) single change-point model into discrete data cases. Based on these successful demonstrations of Bayesian change-point analysis, Park (2010) introduced a Poisson regression change-point model with multiple change-points based on Chib (1998) and Frühwirth-Schnatter and Wagner (2004). Park (2011) further developed a probit regression change-point model for binary and ordinal time series data. Recently, Blackwell (2018) presented change-point models for count data that allow infinitely many change-points through Dirichlet process prior.

The second area of time series modeling using Bayesian methods is the Bayesian implementation of vector autoregression (VAR) models (Brandt and Freeman, 2006, 2009; Brandt and Sandler, 2012). Brandt and Freeman (2006, 2009) introduced a Bayesian VAR model into political science and showed its utility to model macro political processes

with complex endogeneity and contemporaneous shocks. Brandt and Sandler (2012) further extended the Bayesian VAR into count data using a Poisson model.

The third important area of time series modeling using Bayesian methods is time-series cross sectional (TSCS) models. Pang (2010) developed a Bayesian generalized linear model for TSCS data with  $p$ th-order autoregressive error process. Pang (2014) further developed a Bayesian multilevel multifactor error structure model with a  $p$ th-order autoregressive process in linear, probit, and logistic specifications to account for complex cross-sectional dependence in TSCS data. While these models address unobserved time-varying shocks at the individual levels, another concern in TSCS analysis is regime-dependent shocks at the individual level or regime-dependent common shocks. To address these confounders, Park (2012) developed a hidden Markov fixed-effects model and a hidden Markov random-effects model.

## NEW DEVELOPMENTS AND DISCUSSIONS

Before we finalize our review, we would like to highlight a couple of important developments that use Bayesian models as a data generative model while using machine-learning techniques to estimate Bayesian models.

The first example is Bayesian models for text analysis. Grimmer (2010) presented a Bayesian hierarchical topic model for measuring expressed agenda from text data. In Grimmer's (2010) model, each US senator's expressed agenda is a random draw from a Senate-level multinomial distribution. Conditional upon the drawn agenda, a topic of each press release document is a random draw from document-level multinomial distribution. Then, word counts are normalized to have unit length and modeled to follow the von Mises–Fisher distribution. Due to the

complexity of the model, Grimmer (2010) employs variational Bayes methods.

The second example is a structural topic model (STM) (Roberts et al., 2014). Conventional LDA-based topic models use a Bayesian hierarchical structure to make inference about topic and word distributions. STM adds another layer to explain the path from exogenous covariates (or meta information) to topic and word distributions. STM significantly widens the scope of text data analysis in political science, as discussed in Roberts et al. (2014) and Lucas et al. (2015). Estimation of STM is based on variational expectation-maximization (EM).

The third example is a sparse Bayesian regression model (Ratkovic and Tingley, 2017). Ratkovic and Tingley (2017) proposed a Bayesian estimator of a sparse model by replacing the shape parameter of the global shrinkage parameter in Bayesian lasso by calibrated parameters. Ratkovic and Tingley (2017) show that the resulting sparse estimates are consistent and satisfy the Oracle Property and Oracle Inequality.

The last example is Imai et al. (2016). Imai et al. (2016) innovate the estimation method of various Bayesian item response theory models developed by political scientists using EM and variational Bayes methods while keeping the model structure intact. Imai et al. (2016) reported significant reductions in computation time compared to conventional MCMC-based estimation. Their estimation method is particularly useful when researchers analyze large-size data.

These three examples have commonality in using Bayesian models as a generative model for data generating process. However, they avoid conventional MCMC methods to improve computational performance and to account for a more sophisticated model structure.

There is no doubt that political scientists are witnessing an unprecedented inflow of data in the 21st century and these innovations

will serve as important resources to this new challenge. However, we would like to make some cautionary notes regarding this new trend.

First, computational improvements using EM or variational Bayes applies solely or mainly to point estimation at the *mode* of parameters. For example, EM algorithm finds maximum a posteriori (MAP) estimates of parameters, which tend to have a larger mean squared error (MSE) than posterior means (Kyung et al., 2010). When a model has multiple modes, as in mixture models, the MSE difference between MAP and posterior means could be sizable and leads to erroneous inference. Moreover, EM algorithm does not generate measures of uncertainty by itself. Thus, the estimation of uncertainty needs to be done separately after the MAP estimation. The case of Imai et al. (2016) is a good example. Although EM and variational Bayes-based algorithms significantly reduce the computing time to get point estimates of various IRT models, they need to re-run their models (i.e., bootstrap) to get standard errors after. These re-runs for standard error calculation are not just computationally expensive (maybe not that much compared to a full MCMC run) but also inferentially costly because we do not know how much information is lost in this two-step (point estimation and bootstrapping) process.

Generally speaking, most approximate Bayesian methods involve a fundamental shortcoming of understating parameter uncertainty. For example, Grimmer (2011) notes that ‘factorized approximations will always understate the variability in the posterior’ (Grimmer, 2011: 6). A troubling fact is that it is difficult to precisely gauge the level of understatement and the difficulty becomes more pronounced as model complexity increases.

‘Getting uncertainty right’ has been the most important vantage point of Bayesian methods and this is why MCMC methods

that provide proper Monte Carlo draws of a posterior distribution were hailed as a major breakthrough by many applied researchers. With this in mind, we should think carefully about the trade-off involved in recent computational innovations in Bayesian literature.

The second cautionary note is the lack of principled methods to check the model uncertainty in recent computational innovations. For example, EM and variational Bayes-based algorithms transform the Bayesian estimation from stochastic integration to numerical optimization through analytical reconstructions of the posterior distribution. In doing so, it becomes difficult to compute the posterior model probability ( $p(\mathcal{M}|y)$ ) using outputs of EM and variational Bayes-based algorithms. In the case of Grimmer (2010), researchers may want to learn the number of topics from data instead of assuming it to be known and fixed, as in Grimmer (2010). In the case of Imai et al. (2016), researchers may want to compare one data generative model for legislative voting with others to check different utility functions, different parametric forms, and different voting models. All of these questions require a proper evaluation of the posterior model probability ( $p(\mathcal{M}|y)$ ). It would be an interesting future research topic to develop a method to compute the posterior model probability ( $p(\mathcal{M}|y)$ ) using computationally efficient algorithms.

Bayesian inference in political science has contributed significantly to empirical research in political science through its representational consistency, computational power, and modeling flexibility. The advent of big data and innovative machine learning methods pose both opportunities and challenges to Bayesian methods. However, parameter and model uncertainty is essential information for social scientists in theory testing, variable selection, and model comparison. With that in mind, expanding the frontiers of Bayesian methods for

social scientists would be an exciting adventure of social science methodology in the 21st century.

## REFERENCES

- Achen, Christopher H. 1978. Measuring Representation. *American Journal of Political Science* 22(3): 475–510.
- Albert, James H. 1992. Bayesian Estimation of Normal Ogive Item Response Curves Using Gibbs Sampling. *Journal of Educational Statistics* 17(3): 251–69.
- Bafumi, Joseph, Andrew Gelman, David K. Park, and Noah Kaplan. 2005. Practical Issues in Implementing and Understanding Bayesian Ideal Point Estimation. *Political Analysis* 13(2): 171–87.
- Barberá, Pablo. 2015. Birds of the Same Feather Tweet Together: Bayesian Ideal Point Estimation Using Twitter Data. *Political Analysis* 23(1): 76–91.
- Blackwell, Mathew. 2018. Game Changers: Detecting Shifts in Overdispersed Count Data. *Political Analysis* 26(2): 230–39.
- Brandt, Patrick T., and John R. Freeman. 2006. Advances in Bayesian Time Series Modeling and the Study of Politics: Theory Testing, Forecasting, and Policy Analysis. *Political Analysis* 14(1): 1–36.
- Brandt, Patrick T., and John R. Freeman. 2009. Modeling Macro Political Dynamics. *Political Analysis* 17(2): 113–42.
- Brandt, Patrick T., and Todd Sandler. 2012. A Bayesian Poisson Vector Autoregression Model. *Political Analysis* 20(3): 292–315.
- Carlin, Bradley P., Alan E. Gelfand, and Adrian F. M. Smith. 1992. Hierarchical Bayesian Analysis of Change-point Problems. *Applied Statistics* 41(2): 389–405.
- Chib, Siddhartha. 1998. Estimation and Comparison of Multiple Change-Point Models. *Journal of Econometrics* 86(2): 221–41.
- Clinton, Joshua, Simon Jackman, and Douglas Rivers. 2004. The Statistical Analysis of Legislative Behavior: A Unified Approach. *American Political Science Review* 98(2): 355–70.

- CRAN. 2018. Bayesian Task View. <https://CRAN.R-project.org/view=Bayesian>.
- de Finetti, Bruno. 1974. *Theory of Probability*. New York: John Wiley & Sons.
- Earl, David J, and Michael W Deem. 2005. Parallel Tempering: Theory, Applications, and New Perspectives. *Physical Chemistry Chemical Physics* 7(23): 3910–16.
- Encyclopædia Britannica. 2016. 'Bayesian analysis' Encyclopædia Britannica, inc. URL: <https://www.britannica.com/science/Bayesian-analysis> (Access Date 11-20-2019)
- Frühwirth-Schnatter, Sylvia, and Helga Wagner. 2004. Data Augmentation and Gibbs Sampling for Regression Models for Small Counts. *Student* 5(3–4): 201–34.
- Gelman, Andrew, Boris Shor, Joseph Bafumi, and David Park. 2007. Rich State, Poor State, Red State, Blue State: What's the Matter with Connecticut? *Quarterly Journal of Political Science* 2: 345–67.
- Gelman, Andrew, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. 2012. *Bayesian Data Analysis*. 3rd ed. New York: CRC Press.
- Gill, Jeff. 2007. *Bayesian Methods: A Social and Behavioral Sciences Approach*. 2nd ed. Chapman: Hall/CRC.
- Girshick, M. A., and Daniel Lerner. 1950. Model Construction in the Social Sciences—an Expository Discussion of Measurement and Prediction. *Public Opinion Quarterly* 14(4): 710–28.
- Gordon, N. J., D. J. Salmond and A. F. M. Smith. 1993. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proceedings F - Radar and Signal Processing*, 140(2): 107–13
- Grimmer, Justin. 2010. A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases. *Political Analysis* 18(1): 1–35.
- Grimmer, Justin. 2011. An Introduction to Bayesian Inference via Variational Approximations. *Political Analysis* 19(1): 32–47.
- Hoff, Peter D., and Michael D. Ward. 2004. Modeling Dependencies in International Relations Networks. *Political Analysis* 12(2). SPM-PMSAPSA: 160–75.
- Imai, Kosuke, James Lo, and Jonathan Olmsted. 2016. Fast Estimation of Ideal Points with Massive Data. *American Political Science Review* 110(4): 632–56.
- Jackman, Simon. 1994. Measuring Electoral Bias: Australia, 1949–93. *British Journal of Political Science* 24(3): 319–57.
- Jackman, Simon. 2001. Multidimensional Analysis of Roll Call Data via Bayesian Simulation: Identification, Estimation, Inference and Model Checking. *Political Analysis* 9(3): 227–41.
- Jackman, Simon. 2005. Pooling the Polls over an Election Campaign. *Australian Journal of Political Science* 40(4): 499–517.
- Jackman, Simon. 2009. *Bayesian Analysis for the Social Sciences*. New York: John Wiley & Sons.
- King, Gary, and Andrew Gelman. 1991. Systemic Consequences of Incumbency Advantage in U.S. House Elections. *American Journal of Political Science* 35(1): 110–38.
- Kyung, Minjung, Jeff Gill, Malay Ghosh, and George Casella. 2010. Penalized Regression, Standard Errors, and Bayesian Lassos *Bayesian Analysis* 5(2): 369–412.
- Lock, Kari, and Andrew Gelman. 2010. Bayesian Combination of State Polls and Election Forecasts. *Political Analysis* 18(3): 337–48.
- Lucas, Christopher, Richard A. Nielsen, Margaret E. Roberts, Brandon M. Stewart, Alex Storer, and Dustin Tingley. 2015. Computer-Assisted Text Analysis for Comparative Politics. *Political Analysis* 23(2): 254–77.
- MacKay, David J. C. 2003. *Information Theory, Inference and Learning Algorithms*. Cambridge: Cambridge University Press.
- Martin, Andrew D., and Kevin M. Quinn. 2002. Dynamic Ideal Point Estimation via Markov Chain Monte Carlo for the U.S. Supreme Court, 1953–1999. *Political Analysis* 10(2): 134–53.
- Minhas, Shahryar, Peter D. Hoff, and Michael D. Ward. 2016. A New Approach to Analyzing Coevolving Longitudinal Networks in International Relations. *Journal of Peace Research* 53(3): 491–505.
- Neal, Radford M. 2003. Slice Sampling. *Annals of Statistics* 31(3): 705–41.
- Neal, Radford M. 2011. Ch. 5. MCMC Using Hamiltonian Dynamics. In *Handbook of Markov Chain Monte Carlo*, eds. Steve Brooks, Andrew Gelman, Galin L. Jones, and

- Xiao-Li Meng, New York: Chapman & Hall/CRC, pp. 113–62.
- Pang, Xun. 2010. Modeling heterogeneity and serial correlation in binary time-series cross-sectional data: A Bayesian multilevel model with AR(p) errors. *Political Analysis* 18(4): 470–98.
- Pang, Xun. 2014. Varying Responses to Common Shocks and Complex Cross-Sectional Dependence: Dynamic Multilevel Modeling with Multifactor Error Structures for Time-Series Cross-Sectional Data. *Political Analysis* 22(4): 464–96.
- Park, Jong Hee. 2010. Structural Change in U.S. Presidents' Use of Force. *American Journal of Political Science* 54(3): 766–82.
- Park, Jong Hee. 2011. Change-point Analysis of Binary and Ordinal Probit Models: An Application to Bank Rate Policy Under the Inter-war Gold Standard. *Political Analysis* 19(2): 188–204.
- Park, Jong Hee. 2012. A Unified Method for Dynamic and Cross-Sectional Heterogeneity: Introducing Hidden Markov Panel Models. *American Journal of Political Science* 56(4): 1040–54.
- Pernstein, Daniel, Stephen A. Meserve, and James Melton. 2010. Democratic Compromise: A Latent Variable Analysis of Ten Measures of Regime Type. *Political Analysis* 18(4): 426–49.
- Poole, Keith T., and Howard Rosenthal. 1997. *Congress: A Political-Economic History of Roll-Call Voting*. Oxford: Oxford University Press.
- Quinn, Kevin M. 2004. Bayesian Factor Analysis for Mixed Ordinal and Continuous Responses. *Political Analysis* 12(4): 338–53.
- R Core Team. 2018. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. <https://www.R-project.org/>.
- Ratkovic, Marc, and Dustin Tingley. 2017. Sparse Estimation and Uncertainty with Application to Subgroup Analysis. *Political Analysis* 25(1): 1–40.
- Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G. Rand. 2014. Structural Topic Models for Open-Ended Survey Responses. *American Journal of Political Science*, 58: 1064–82.
- Savage, Leonard J. 1954. *The Foundations of Statistics*. New York: John Wiley & Sons.
- Selivanov, Dmitriy, and Qing Wang. 2018. text2vec: Modern Text Mining Framework for R. R package version 0.5.1. <https://CRAN.R-project.org/package=text2vec>
- Schnakenberg, Keith E., and Christopher J. Fariss. 2014. Dynamic Patterns of Human Rights Practices. *Political Science Research and Methods* 2(1): 1–31.
- Spirling, Arthur. 2007. Bayesian Approaches for Limited Dependent Variable Change Point Problems. *Political Analysis* 15(4): 387–405.
- Tanner, M. A., and Wong, W. H. 1987. The Calculation of Posterior Distributions by Data Augmentation (with discussion). *Journal of the American Statistical Association* 82: 528–50.
- Treier, Shawn, and Simon Jackman. 2008. Democracy as a Latent Variable. *American Journal of Political Science* 52(1): 201–17.
- van Dyk, David A. and Xiao-Li Meng. 2001. The Art of Data Augmentation, *Journal of Computational and Graphical Statistics*, 10(1): 1–50.
- Ward, Michael D, and Peter D Hoff. 2007. Persistent Patterns of International Commerce. *Journal of Peace Research* 44(2): 157–75.
- Western, Bruce. 1998. Causal Heterogeneity in Comparative Research: A Bayesian Hierarchical Modelling Approach. *American Journal of Political Science* 42(4): 1233–59.
- Western, Bruce, and Meredith Kleykamp. 2004. A Bayesian Change Point Model for Historical Time Series Analysis. *Political Analysis* 12(4): 354–74.

# Bayesian Ideal Point Estimation

Shawn Treier

## INTRODUCTION

Ideology is a fundamental concept of politics, and this is especially true of legislatures. The popular understanding is that a politician makes policy decisions based on their discernible ideology, with relative comparisons of these politicians easily placed in common conversation. This is consistent with an academic conceptualization of ideology as a set of constrained belief systems, represented geometrically by low-dimensional ‘basic spaces’ where legislators’ ideal positions in this space determine their observed stances on the concrete policy proposals they regularly face in legislative session. This characterization, best enunciated by Poole (2005), implies that the secret to understanding how a legislature operates is through the configuration of these ideal points in the basic ideological space, demanding empirical measurement. This invitation was successfully answered with the ground-breaking NOMINATE-based

estimates, developed by Keith Poole and Howard Rosenthal (1985, 1991, 1997) and greatly facilitated by the presence of large amounts of recorded roll calls – and a lack of strict party discipline. The widespread distribution of these measures has advanced the study of Congress – and other legislatures and political institutions – immensely.<sup>1</sup>

The focus of this chapter is a survey of work that utilizes a new set of statistical tools to pursue the same measurement issues as well as address new questions. The increase in computing power by the 1990s made feasible the application of Bayesian approaches to the estimation of many types of models. Bayesian estimation of the ideal point model was first broadly promulgated by Clinton et al. (2004), who outline the advantages. These include the relatively simple simultaneous estimation of all parameters, a full characterization of uncertainty of the ideal points, and the immediate calculation of complex auxiliary quantities. These qualities also make

the models easily extendable, allowing for a straightforward incorporation of the ideal points into regression models that account for uncertainty or implementation of direct tests of theories based on the configuration of preferences.

This chapter will outline these developments and survey the many applications and extensions. First, the basics of the Clinton–Jackman–Rivers (CJR) IDEAL model (estimation, inference, identification, and model assessment) are presented, as well as comparisons to NOMINATE (including the Bayesian version). Next, many various extensions to the model are covered. Bayesian models are naturally modular through the specification of hierarchical structures, simplifying the inclusion of ideal points as dependent or independent variables in regression models. Many researchers also address the complicated identification issue of comparability over time and/or institutions. Some of these solutions are specifically Bayesian; most notably, the dynamic models behind the Martin Quinn Supreme Court ideal point estimates, which allow small movements over time in any direction. Others, such as identifying ‘bridge’ votes (e.g., votes on identical worded bills in the House and Senate) or other public positions (e.g., presidential proclamations), concern only the structure of the data but are easier to assess with Bayesian methods. Models have also been extended in order to include agenda information (tracking bill passages to identify status quo and bill parameters).

The inclusion of data other than roll calls will be a critical extension for the future estimating ideal points. Of all these potential data sources, the most important will be text data. Melding text data with roll calls may create greater distinctions between legislators, provide another basis for establishing comparability over time and institution, and characterize the ideological content of bills directly, with a Bayesian approach remaining the most feasible method, despite many challenges.

## SPATIAL MODEL OF VOTING AND RUM FRAMEWORK

Two methodological approaches familiar to most political scientists are the rational choice theory of spatial voting and the random utility model (RUM) framework applied to logit and probit regressions.

In the spatial model of voting, every vote  $j$  is a choice between two positions: ‘Nay’ (status quo,  $q_j$ ) and ‘Yea’ (bill proposal or amendment,  $p_j$ ). Every legislator’s preferences between these two options are determined by the relative ideological proximity of the individual’s ‘ideal’ point,  $x_i$  (i.e., proposal,  $p_j$ , is the most preferred proposal is closer to  $x_i$ , otherwise  $q_j$ ). Utility is defined for both options in the following one-dimensional representation:

$$U(q_j) = f(|x_i - q_j|) \tag{1}$$

$$U(p_j) = f(|x_i - p_j|) \tag{2}$$

with legislator  $i$ ’s vote on roll call  $j$  ( $y_{ij}$ ) represented as

$$y_{ij} = \begin{cases} \text{‘Yea’} & \text{if } U(p_j) > U(q_j) \\ \text{‘Nay’} & \text{if } U(q_j) > U(p_j) \end{cases} \tag{3}$$

and indifferent when  $U(p_j) = U(q_j)$  (but a probability zero event). The decision can be represented in terms of a threshold. Define the midpoint between  $q_j$  and  $p_j$  as  $c_j = \frac{q_j + p_j}{2}$ .

if  $p_j > q_j$ , legislator  $i$ ’s voting decision is

$$y_{ij} = \begin{cases} \text{‘Yea’} & \text{if } x_i > c_j \\ \text{‘Nay’} & \text{if } x_i < c_j \end{cases} \tag{4}$$

with  $>$  and  $<$  operators reversed if  $q_j > p_j$ .

An essential assumption in the spatial voting model is that only the distance between the ideal point and voting options are relevant. An almost universally applied assumption is that preferences are single-peaked, with utility  $U(p_j)$  [ $U(q_j)$ ] decreasing as  $|x_i - p_j|$  [ $|x_i - q_j|$ ] increases. Furthermore, these



preferences are assumed to be symmetric, regardless of the vote option being less than or greater than  $x_i$ .

The statistical model is characterized by specific parametric forms for  $f(\cdot)$  and the addition of random error in  $U(\cdot)$ . The Clinton et al. (2004) representation specifies  $f(\cdot)$  as the quadratic function, with the utility of each option including a linearly additive random error:

$$U(q_j) = -(x_i - q_j)^2 + \varepsilon_{i,q_j} \quad (5)$$

$$U(p_j) = -(x_i - p_j)^2 + \varepsilon_{i,p_j} \quad (6)$$

where both  $\varepsilon_{i,q_j} \sim N(0,0.5)$  and  $\varepsilon_{i,p_j} \sim N(0,0.5)$  are independent, thus implying  $\varepsilon_{i,p_j} - \varepsilon_{i,q_j} \sim N(0,1)$ , consistent with a traditional RUM specification for a probit model (conversely, each error term could follow a Type I error distribution in order to define a logistic model).

The difference of these utilities can be represented as

$$\begin{aligned} y_{ij}^* &= U_i(p_j) - U_i(q_j) \\ &= -(x_i - p_j)^2 + (x_i - q_j)^2 + (\varepsilon_{i,p_j} - \varepsilon_{i,q_j}) \\ &= -x_i^2 + 2p_jx_i + p_j^2 + x_i^2 \\ &\quad - 2q_jx_i + q_j^2 + (\varepsilon_{i,p_j} - \varepsilon_{i,q_j}) \\ &= 2(p_j - q_j)x_i - (p_j^2 - q_j^2) + (\varepsilon_{i,p_j} - \varepsilon_{i,q_j}) \\ &= \beta_jx_i - \alpha_j + \varepsilon_{ij}, \end{aligned}$$

where  $\beta_j = 2(p_j - q_j)$  and  $\alpha_j = p_j^2 - q_j^2$ . The model reduces to the latent-variable representation of a binary logit or probit, with the following complications: (1) the independent variable  $x_i$  is unobserved and (2) the intercept and slope coefficients  $\alpha_j$  and  $\beta_j$  vary by vote  $j$ . This model also corresponds to the binary item response theory (IRT) models from education and psychology, testing literatures with the following reparameterization:

$$\begin{aligned} y_{ij}^* &= \beta_jx_i - \alpha_j + \varepsilon_{ij} \\ &= \beta_j \left( x_i - \frac{\alpha_j}{\beta_j} \right) + \varepsilon_{ij} \\ &= \beta_j \left( x_i - \frac{p_j^2 - q_j^2}{2(p_j - q_j)} \right) + \varepsilon_{ij} \\ &= \beta_j \left( x_i - \frac{p_j + q_j}{2} \right) + \varepsilon_{ij} \\ &= \beta_j(x_i - \kappa_j) + \varepsilon_{ij} \end{aligned}$$

with  $\beta_j$  as the *discrimination* parameter and  $\kappa_j$  is the *difficulty* parameter. The difficulty parameter is identical to the cutpoint from the spatial voting model, so it can be recovered as  $\kappa_j = -\frac{\alpha_j}{\beta_j}$ , and is the value of  $x_i$  such

that  $\Pr(y_{ij} = 1) = 0.5$ . The discrimination parameter measures the extent to which the question separates higher- and lower-ability students in the testing literature, and – with respect to ideal point models – separates right and left (usually oriented as conservative and liberal) legislators; e.g., the higher the cutpoint, the more conservative the vote (relative to the status quo). If  $\beta_j = 0$ , it implies the vote is unrelated to the ideological positions of legislators.

The multidimensional representation of the CJR model is Euclidean, with equal weights for the dimensions. In two dimensions,

$$\begin{aligned} U(\mathbf{q}_j) &= -(x_{1i} - q_{1j})^2 - (x_{2i} - q_{2j})^2 + \varepsilon_{i,\mathbf{q}_j} \\ U(\mathbf{p}_j) &= -(x_{1i} - p_{1j})^2 - (x_{2i} - p_{2j})^2 + \varepsilon_{i,\mathbf{p}_j} \end{aligned}$$

with a reduced form of

$$y_{ij}^* = x_{1i}\beta_{1j} + x_{2i}\beta_{2j} - \alpha_j + \varepsilon_{ij}$$

More generally,

$$\begin{aligned} y_{ij}^* &= x_{1i}\beta_{1j} + x_{2i}\beta_{2j} + \dots + x_{ki}\beta_{kj} \\ &\quad - \alpha_j + \varepsilon_{ij} = \mathbf{x}\beta_j - \alpha_j + \varepsilon_{ij} \end{aligned}$$

is the reduced form for the difference in the following utilities:

$$\begin{aligned}
 U(\mathbf{q}_j) &= \sum_{k=1}^K -(x_{ki} - q_{kj})^2 + \varepsilon_{i,\mathbf{q}_j} \\
 &= -\|\mathbf{x}_i - \mathbf{q}_j\|^2 + \varepsilon_{i,\mathbf{q}_j} \\
 U(\mathbf{p}_j) &= \sum_{k=1}^K -(x_{ki} - p_{kj})^2 + \varepsilon_{i,\mathbf{p}_j} \\
 &= -\|\mathbf{x}_i - \mathbf{p}_j\|^2 + \varepsilon_{i,\mathbf{p}_j}
 \end{aligned}$$

where  $\beta_j = 2(\mathbf{p}_j - \mathbf{q}_j)$  and  $\alpha_j = \mathbf{p}'_j\mathbf{p}_j - \mathbf{q}'_j\mathbf{q}_j$ .

**ESTIMATION**

The probability that legislator  $i$  on roll call vote  $j$  votes ‘Yea’ is

$$\begin{aligned}
 \Pr(y_{ij} = 1) &= \Pr(y_{ij}^* > 0) \\
 &= \Pr(\varepsilon_{ij} < \beta_j \mathbf{x}_i - \alpha_j) = F(\beta_j \mathbf{x}_i - \alpha_j)
 \end{aligned}$$

and defines the likelihood of the model:

$$\begin{aligned}
 L(\mathbf{x}, \alpha, \beta | \mathbf{Y}) &\propto f(\mathbf{Y} | \mathbf{x}, \alpha, \beta) \\
 &= \prod_{i=1}^n \prod_{j=1}^J F(\beta_j \mathbf{x}_i - \alpha_j)^{y_{ij}} \\
 &\quad [1 - F(\beta_j \mathbf{x}_i - \alpha_j)]^{1-y_{ij}} \tag{7}
 \end{aligned}$$

Like all measurement models, this relies on the local independence assumption: conditional on  $\mathbf{x}_i$ , responses  $y_{ij}$ ,  $y_{ik}$  are independent; the only source of commonality is latent value  $x_i$ . So, not only are decisions by individuals independent (i.e.,  $y_{ij}$ ,  $y_{hj}$ ,  $i \neq h$ ), but each vote decision is distinct from the others. Violations could occur if the dimensionality is underestimated or if votes have a logical connection (such as a log-roll).

The estimation problem is particularly daunting; there are  $nJ$  observations with  $nK + J(K + 1)$  parameters and increases in either

$n$  or  $J$  present difficulties in the asymptotic properties of the model. Joint maximization of the full information likelihood is computationally difficult (and not necessarily consistent). Marginal maximum likelihood has been implemented in similar models of IRT, but a critical aspect of these models is that they *marginalize*  $\mathbf{x}_i$  – average over – from the likelihood, thus not estimating  $x_i$ . *Ex ante posteriori* approaches are used to recover the ideal points, but these employ the same prior distribution assumptions of the Bayesian model. In any case, direct marginalization is often quite difficult, so the E-M algorithm is deployed instead. Finally, the approach used with NOMINATE – alternating conditional maximum likelihood – estimates one set of parameters with the others fixed and repeats the process for the other parameter blocks, as discussed in detail below.

Proponents of a Bayesian approach argue that it provides a less problematic approach to estimation to all parameters. The fundamental difference between the two approaches appears superficial:

$$\begin{aligned}
 \text{Likelihood} : L(\mathbf{x}, \alpha, \beta | \mathbf{y}) &\propto f(\mathbf{y} | \mathbf{x}, \alpha, \beta) \\
 \text{Bayesian} : f(\mathbf{x}, \alpha, \beta | \mathbf{y}) &\propto f(\mathbf{y} | \mathbf{x}, \alpha, \beta) f(\mathbf{x}, \alpha, \beta)
 \end{aligned}$$

In the case of ideal point estimation, the prior distribution for each set of parameters are *a priori* independent —  $f(\mathbf{x}, \alpha, \beta) = f(\mathbf{x})f(\alpha)f(\beta)$  — with assumed distributions

$$\begin{aligned}
 x_i &\sim N(0, 1) \quad \forall i = 1, \dots, n \\
 \beta_j &\sim N(0, \sigma_\beta^2) \quad \forall j = 1, \dots, J \\
 \alpha_j &\sim N(0, \sigma_\alpha^2) \quad \forall j = 1, \dots, J
 \end{aligned}$$

Typically, the variance parameters are extremely diffuse, such as  $\sigma_\alpha^2 = \sigma_\beta^2 = 100$ . Both expressions condition the parameters on observed data  $\mathbf{y}$ , reflecting our focus on the parameters, but  $L(\mathbf{x}, \alpha, \beta | \mathbf{y})$  is not a proper joint probability distribution of the parameters. The addition of the prior completes the posterior distribution

$$f(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta} | \mathbf{y}) = \frac{f(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{y})}{f(\mathbf{y})}$$

$$= \frac{f(\mathbf{y} | \mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}) f(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta})}{f(\mathbf{y})},$$

with

$$f(\mathbf{y}) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(\mathbf{y} | \mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}) f(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}) d\mathbf{x} d\boldsymbol{\alpha} d\boldsymbol{\beta},$$

a complicated but unnecessary component (which, if calculable, is useful for model comparison) and thus the summary of the elements containing the parameters. The additional prior information does not contribute much influence to the posterior (other than set the scale) but does allow us to assign probabilistic interpretations to inferential statements. A second, more practical advantage is the relative ease of computation.

In general, there are three approaches to Bayesian estimation. First, for the simplest problems, the posterior distribution can be directly summarized from the data. When the joint posterior distribution does not define a standard distribution and cannot easily be sampled, there are two common approaches: one is to find the posterior mode and apply an approximation, similar to the maximization approach used for likelihood analysis. Since direct joint optimization in the ideal point model is impractical, the expectation-maximization algorithm can be applied, and such techniques have become more utilized as the size of vote matrices have exploded (see Imai et al., 2016, as an example). With an increased focus on text data, Bayesian variational approximation is another popular choice (Ormerod and Wand, 2010; Grimmer, 2011).<sup>2</sup>

The final approach, and the one most applicable to the Bayesian ideal point problem, is dependent sampling through Markov-chain Monte Carlo (MCMC) methods. The most general approach is Metropolis–Hastings sampling, where one samples from a ‘candidate’ distribution (usually a multivariate

normal or *t* distribution) and ‘adjusts’ the sample by comparing the target and candidate distributions at the sampled value. The algorithm always accepts draws sampled too infrequently from the candidate distribution and probabilistically rejects oversampled regions of the parameter space (and upon rejection, simply replicates another instance of the current value). Applying Metropolis–Hastings sampling to all parameters simultaneously is completely impractical, so Gibbs sampling can be employed instead. If each block of parameters, conditional on the other parameters, forms an easily sampled distribution (usually either an easily recognizable, standard distribution or one-dimensional), then samples are produced iteratively, conditional on other parameters, without rejection. There is also a hybrid approach, where parameter blocks that are less easily sampled employ Metropolis–Hastings sampling, while other parameters are sampled through Gibbs steps. The product is a sample of thousands of draws for the parameters from the posterior distribution. A summary of the distribution based on means, variances, and quantiles involves the simple calculation of summary statistics, as does the construction of summary measures for the posterior distribution of auxiliary quantities.

The multivariate binomial likelihood and normal distribution priors certainly do not combine to form semi-conjugate conditionals. One could apply Metropolis–Hastings sampling on each set of parameters (ideal points and discrimination and difficulty parameters), conditional on current iteration values of the other parameters (see, e.g., Patz and Junker, 1999b), but the simplest approach involves the probit specification of the utility error terms and the sampling approach of Albert (1992), which incorporates the latent utility differences  $y_{ij}^*$  as auxiliary parameters of the posterior. The steps are taken in the following order (following Jackman, 2009: 454–8): (1) sample  $\mathbf{Y}^{*(t)}$  from  $f(\mathbf{Y}^* | \mathbf{X}^{(t-1)}, \boldsymbol{\beta}^{(t-1)}, \boldsymbol{\alpha}^{(t-1)})$ ; (2) sample  $\boldsymbol{\Theta} = [\boldsymbol{\beta}^{(t)}, \boldsymbol{\alpha}^{(t)}]$  from  $f(\boldsymbol{\Theta} | \mathbf{Y}^{*(t)}, \mathbf{X}^{(t-1)})$ ; and (3) sample  $\mathbf{X}^{(t)}$  from

$f(\mathbf{X}|\mathbf{Y}^{*(t)}, \boldsymbol{\beta}^{(t)}, \boldsymbol{\alpha}^{(t)})$ . In one dimension, conditional on model parameters  $\mathbf{x}$ ,  $\boldsymbol{\alpha}$  and  $\boldsymbol{\beta}$ , each  $y_{ij}^*$  is distributed truncated normal:

$$f(y_{ij}^* | y_{ij}, \mathbf{x}_i, \boldsymbol{\beta}_j, \boldsymbol{\alpha}_j) = \begin{cases} N(\boldsymbol{\beta}_j x_i - \boldsymbol{\alpha}_j, 1)\mathbb{I}(-\infty, 0) & \text{if } y_{ij} = 0 \\ N(\boldsymbol{\beta}_j x_i - \boldsymbol{\alpha}_j, 1)\mathbb{I}(0, \infty) & \text{if } y_{ij} = 1 \end{cases}$$

with the implied threshold at 0, since  $y_{ij} = 1$  if  $U(p_j) - U(q_j) \equiv y_{ij}^* > 0$ , and less than 0 if  $y_{ij} = 0$ . Conditional on these latent values, the MCMC problem for  $\boldsymbol{\theta} = [\boldsymbol{\beta} \boldsymbol{\alpha}]$ , with ‘stacked’ prior  $N(\boldsymbol{\tau}_0, \mathbf{T}_0)$  and augmented  $\mathbf{X}^* = [-\mathbf{1} \ \mathbf{x}]$ , simply defines  $J$  Bayesian regressions of  $\mathbf{X}^*$  on  $\mathbf{Y}^*$ , which require draws from normal distributions:

$$f(\boldsymbol{\beta}_j, \boldsymbol{\alpha}_j | \mathbf{X}^*, \mathbf{y}_j^*) = N(\boldsymbol{\mu}_{\theta_j}, \mathbf{V}_{\theta_j})$$

where  $\boldsymbol{\mu}_{\theta_j} = (\mathbf{X}^{*\prime} \mathbf{X}^* + \mathbf{T}_0^{-1})^{-1} (\mathbf{X}^{*\prime} \mathbf{y}^* + \mathbf{T}_0^{-1} \boldsymbol{\tau}_0)$  and  $\mathbf{V}_{\theta_j} = (\mathbf{X}^{*\prime} \mathbf{X}^* + \mathbf{T}_0^{-1})^{-1}$ . Finally, conditional distributions for  $\mathbf{x}$  are defined by a ‘reverse regression’ relationship. Conditional on other parameters, we can rearrange

$$y_{ij}^* + \boldsymbol{\alpha}_j = x_i \boldsymbol{\beta}_j + \boldsymbol{\varepsilon}_{ij}$$

to form  $n$  regressions, each on  $J$  observations of the discrimination parameters  $\boldsymbol{\beta}$  regressed on artificial dependent variable  $w_{ij} = y_{ij}^* + \boldsymbol{\alpha}_j$ . Then,  $\mathbf{x}_i$  is sampled from

$$f(x_i | \mathbf{y}_i^*, \boldsymbol{\alpha}, \boldsymbol{\beta}) = N(\mu_{x_i}, \sigma_{x_i}^2)$$

where  $\mu_{x_i} = (\boldsymbol{\beta}' \boldsymbol{\beta} + 1)^{-1} \boldsymbol{\beta} \mathbf{w}_i$  and  $\sigma_{x_i}^2 = (\boldsymbol{\beta}' \boldsymbol{\beta} + 1)^{-1}$ . In multiple dimensions,  $\mathbf{X}^* = [-\mathbf{1} \ \mathbf{x}_1 \ \mathbf{x}_2 \ \dots \ \mathbf{x}_K]$  in steps 1 and 2, while step 3 involves  $\mathbf{x}'_i = \begin{bmatrix} x_{i1} & x_{i2} & \dots & x_{iK} \end{bmatrix}$  and identity matrix  $\mathbf{I}$  for the prior variance component in the variance.

An additional concern is missing data. Legislators will have missing roll calls due to absences or members not being considered during part of the Congress or Congresses. If the absences are unrelated to the vote, this

poses no particular problems. There are two approaches: the first uses the complete likelihood, where only members’ observed votes contribute to the likelihood. This is straightforward in the ideal point model, since each vote defines a separate probit model. In the Gibbs sampler,  $y_{ij}^*$  is sampled only for those  $j$  where legislator  $i$  votes ‘Yea’ or ‘Nay’. The sampler for  $x_i$  only conditions on  $y_{ij}^*$  and vote parameters  $\boldsymbol{\alpha}_j$  and  $\boldsymbol{\beta}_j$  for votes  $j$  that are observed; i.e., the reverse regression of dimension  $J_i$ . Finally, the sampler for  $\boldsymbol{\theta}_j$  is a normal Bayesian regression but with only  $n_j$  observations (the number of legislators who voted).

The second approach treats the missing votes as parameters to be estimated; at each iteration, a value for  $y_{ij}$  could be generated, then the algorithm proceeds normally (Patz and Junker, 1999a). In practice, the only additional step is in the generation of  $y_{ij}^*$ ; if  $y_{ij} = 1$ , one samples from a normal distribution truncated from below by zero, and if  $y_{ij} = 0$ , it is truncated from above by zero. If  $y_{ij}$  is missing, then the sample is generated from an untruncated normal distribution (Albert, 1992; Albert and Chib, 1993).

In general, considerations of missing data follow the same determinations of MCAR, MAR, and non-ignorability, with the approaches described so far being applicable to data that is MCAR or MAR. Missing data that is non-ignorable will require the specification of the missing-data process and will be problem specific (Mislevy, 2016). Rosas et al. (2015), for instance, posit a model of strategic abstention where actors decide to abstain when their own preferences conflict with those of a principal to whom they must answer (such as party leadership). They adjust the standard model to account for congruence or incongruity (incorporating identified positions of the principal on votes). Their Monte Carlo evidence indicates that this particular form of abstention is unproblematic – unless it occurs at high rates (e.g., certainly not in the US Congress, but it is characteristic of other legislatures and the UN General Assembly).<sup>3</sup>

To identify the model, one must set the *location*, *scale*, and *polarity*. In

$$y_{ij}^* = x_i \beta_j - \alpha_j + \varepsilon_{ij},$$

substituting  $x^* = \left(\frac{1}{c}\right)x_i$  and  $\beta_j^* = c\beta_j$  produces the same likelihood, as does  $x^* = x_i - c$  and  $\alpha_j^* = \alpha_j - \beta_j c$ . More generally,  $\mathbf{x}_i$  can be centered on any  $K$  length vector  $\mathbf{c}$  and/or transformed by any  $K \times K$  matrix  $\mathbf{C}$ , and corresponding shifts in  $\beta$  and  $\alpha$  will produce the same posterior. To identify the model, one could set the mean (location) and variance (scale) of the ideal points – although during estimation, this complicates the last Gibbs step. Furthermore, these restrictions do not address polarity; e.g., estimates where conservatives are located on the right or the left could be produced. This is discussed in CJR and Bafumi et al. (2005) and implemented in *ideal* (Jackman, 2015). Two alternatives exist: first, one could fix ideal points to constants. In one dimension, fix a ‘left’ legislator and a ‘right’ legislator (say, at  $-1$  and  $1$ ). In multiple dimensions,  $K(K + 1)$  linearly independent restrictions on ideal points are sufficient (and, if only considering constraints on the ideal points, necessary); e.g., for two dimensions, fixing three ideal points that form a triangle (Rivers 2003). While making such restrictions in one dimension is simple (pick two legislators who, by reputation, are far left and far right), in multiple dimensions these choices are more difficult. They need to be chosen carefully: if any of the positions are between two modes (say, in 1D, using a moderate as an anchor), selecting a high posterior mode is not guaranteed (Bafumi et al., 2005: 177–8). One may impose these constraints during estimation, applied to each iteration as they are sampled, or one can instead implement these restrictions after estimation through post-processing, since the likelihood does not need to be identified in order to sample from the posterior (de Jong et al., 2003). The convergence properties can also be improved as a result.

Second, a final approach to identification is the imposition of restrictions on the

reduced-form vote parameters  $\beta$ , similar to confirmatory factor analysis. To identify the model through the vote parameters, the items should be permutable into the following lower triangular structure.

$$\beta = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ \beta_{2,1} & 1 & 0 & \dots & 0 \\ \beta_{3,1} & \beta_{3,2} & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots \\ \beta_{j,1} & \beta_{j,2} & \beta_{j,3} & \dots & 1 \\ \vdots & & & & \vdots \\ \beta_{J,1} & \beta_{J,2} & \beta_{J,3} & \dots & \beta_{J,K} \end{bmatrix}$$

Each dimension has a ‘reference’ item, with one item relating to only one dimension, and the upper triangle of zeros indicate that the reference items can only be related to a subset of issues (except for the  $K$ th dimension reference item). This is the *minimum* number of restrictions; in practice, most, if not all, of the reference items should load only on the reference dimension, and any reasonable exclusion restrictions on other votes should be implemented.

### NOMINATE

Without the ground-breaking work by Keith Poole and Howard Rosenthal, we might not be dissecting the topic of this chapter. They synthesized early political-science analysis of roll calls, rational-choice models of voting, psychological research on decision-making and measurement theory, and the cutting edge of scientific computing to produce one of the most consequential developments in political methodology. Their measures of ideal points revolutionized the study of Congress and have been applied to a myriad of legislatures outside the United States. Given the prominence and ubiquity of their techniques (NOMINATE, W-NOMINATE, and DW-NOMINATE), along with Poole’s additional contributions (Common Space

Scores, Optimal Classification), the IDEAL model of CJR and other variants inevitably must address the question ‘What advantage does the Bayesian model offer over the universally applied non-Bayesian approaches?’ These questions are considered in Clinton et al. (2004) and Bafumi et al. (2005), and they are debated extensively between Carroll et al. (2009) and Clinton and Jackman (2009).

For NOMINATE,  $f(\cdot)$  is the Gaussian function instead of a quadratic expression (Poole and Rosenthal, 1985):

$$U(q_j) = \beta e^{\left(\frac{1}{8}(x_i - q_j)^2\right)} + \varepsilon_{i,q_j}$$

$$U(p_j) = \beta e^{\left(\frac{1}{8}(x_i - p_j)^2\right)} + \varepsilon_{i,p_j}$$

where  $\beta$  is a ‘signal to noise ratio’, and the error terms follow Type I extreme value distributions (in NOMINATE and W-NOMINATE) or  $N(0, 0.5)$  distributions (for DW-NOMINATE).

The utility difference does not reduce, with the implemented model reparameterizing the ‘Yea’ and ‘Nay’ outcomes as  $z_{j,p_j} = z_{mj} + d_j$  and  $z_{j,q_j} = z_{mj} - d_j$  in the differences of utilities, estimating cutpoint  $z_{mj} = \frac{p_j + q_j}{2}$  and quantity  $d_j = \frac{p_j - q_j}{2}$ :

$$y_{ij}^* = \beta \left\{ \exp\left(\frac{1}{8}(x_i - z_{j,p_j})^2\right) - \exp\left(\frac{1}{8}(x_i - z_{i,q_j})^2\right) \right\} + \varepsilon_{ij}$$

To estimate ideal points and vote parameters, alternative conditional maximum likelihood is used:

- 1 Obtain start values for  $x_i$ .
- 2 Estimate roll call parameters, holding other parameters fixed (using normal steps in maximum likelihood estimation (MLE)).
- 3 Estimate ideal points, holding other parameters constant.
- 4 Estimate  $\beta$ , holding other parameters constant.
- 5 Repeat steps until convergence.

Each step employs regular MLE techniques for each set of parameters, at each iteration, until convergence. Final convergence occurs when parameters are correlated at 0.99 with previous estimates. The alternating nature of the algorithm does not guarantee the typical properties of MLE, and, in any case, the matrix of second derivatives is too large to be calculated. Clinton and Jackman (2009) provide an extended critique of estimation and inference from NOMINATE, while Carroll et al. (2009) question the sensitivity to prior information. Clinton and Jackman (2009) counter their characterization of the sensitivity of the model. The entire debate between the two camps is worth reading closely, but the primary differences between the NOMINATE and IDEAL models are (1) the functional form specified for utilities and (2) the mode of estimation and inference. While Carroll et al. (2013) demonstrate support for the Gaussian utility form, Carroll et al. (2009) find that the estimates of the ideal points and vote parameters are extremely similar; the choice is between Bayesian and non-Bayesian estimation. Clinton and Jackman (2009) argue that the Bayesian approach, regardless of model type, has advantages regarding (1) inference, (2) model assessment, and (3) model extensibility.

The advantages of (1) are most important and characterize nearly every article covered in this chapter. Means, medians, quantiles, and ranks of parameters – probability of events – and one- and two-dimensional probability interval calculations are all produced from simple summary statistic calculations on MCMC output. Simple calculations of complex quantities open up avenues for theory testing.

On the other hand, while most recognize (2) as being equally important, model assessment is deployed much too infrequently, with researchers rarely exploring item fit, observation outliers, dimensionality and local independence. Information criteria, especially the Deviance Information Criterion (Spiegelhalter et al., 2002), are frequently

used to compare separately estimated models, while Bayesian residual analysis (Albert and Chib, 1995) and posterior predictive model checks (Meng, 1994; Gelman et al., 1996) are underutilized. Sinharay (2016) provides an excellent summary of model assessment for IRT models. This chapter will present extensive examples of (3), which is the ease at which the standard Bayesian ideal point model can be adapted.

**COMPARABILITY**

**Over Time**

The difficulty with over-time comparisons is that *everything* is changing: potentially, both the bills being voted upon and the legislators’ positions change over time. If the agenda is changing, how do we know if legislators’ ideological positions have evolved if they are simply voting on more liberal (or conservative) policies? The restrictions to the agenda and/or the degree to which legislators’ ideal points change over time must be implemented.

The simplest approach is to assume constant preferences for the legislators, which appears in Poole’s Common Space Scores. DW-NOMINATE specifies the following model for ideal points, which parameterizes ideal points changing in terms of a polynomial time trend:

$$x_{it} = x_{i0} + x_{i1}\phi_{t1} + x_{i2}\phi_{t2} + x_{i3}\phi_{t3} + \dots + x_{iv}\phi_{tv}$$

where  $v = 0$  is a constant model ( $x_{it} = x_{i0}$ ) and  $v = 1$  is a linear model actually implemented. Consequently,

$$\begin{aligned} x_{it} &= x_{i0} + x_{i1}\phi_{t1} \\ &= x_{i0} + x_{i1} \left[ -1 + (t - 1) \left( \frac{2}{T_i - 1} \right) \right] \end{aligned}$$

The alternating conditional maximum likelihood procedure follows a slightly

different order for DW-NOMINATE than W-NOMINATE:  $\beta$  and weights  $w$ , roll call parameters, and then ideal points.  $x_{i0}$  lies in the unit sphere for identification. Conditional estimation also holds other dimensions constant and error terms are normal, not logistic.

In this specification, legislators can only change in one direction and by the same amount every year; only the start and end points are estimated, with all other points interpolated. This may not pose too many problems if one simply wants to control for shifts in the legislature over time beyond replacement, but it does render the estimates useless for testing responsiveness in particular elections, since regardless of how factors change from election to election, the member changes by *exactly* the same amount each election.<sup>4</sup> Higher-order polynomial approximations, such as quadratic and cubic specifications, are possible. Voeten (2004) implements a quadratic specification for sessions of the UN General Assembly between 1990 and 2001.

A more general and flexible approach was developed by Martin and Quinn (2002) applying the methods of Bayesian dynamic linear models (West and Harrison, 1997) to the latent-variable representation of the RUM. Their model is the same as CJR, except that the ideal points are also subscripted by time,  $t$ . The utility differences are

$$y_{ijt}^* = \beta_j x_{it} + \epsilon_{ijt}$$

Martin and Quinn (2002) apply this model to cases of the modern US Supreme Court since 1937. Ideal points for each justice across terms follow a random walk time series  $x_{it} = x_{i,t-1} + \delta_{it}$ , where  $\delta_{it} \sim N(0, \Delta_{it})$ . The justice’s first term is set through the prior  $x_{i0} \sim N(m_{i0}, C_{i0})$ . For most justices, this initial prior is  $N(0, 1)$ , with a select number of justices assigned less variable priors ( $C_{i0} = 0.1$ ) and liberal, moderate, or conservative prior ideal point positions (e.g., William Douglas at  $-3.0$ ; Clarence Thomas and Antonin Scalia at  $2.5$ ; William Rehnquist at  $2.0$ ). The evolutionary variance  $\Delta_{it}$  controls the amount

of year-to-year change in the ideal points;  $\Delta_{it} = 0$  implies constant preferences while  $\Delta_{it} \rightarrow \infty$  approaches a specification where the time series of ideal points is independent. The larger this variance, the more likely abrupt shifts in preferences are observed; smaller values result in much smoother trajectories. In Martin and Quinn's (2002) specification, this value is set to 0.1, except for William Douglas, whose value is set at 0.001 (for reasons of consistent ideological extremity and the substantial amount of missing votes near the end of his tenure).

The same Gibbs sampler can be applied to this dynamic model but with one important caveat. Because of the dynamic relation between ideal points, the samples for individual ideal point estimates will be highly correlated and thus mix slowly. Instead, Martin and Quinn (2002) implement the forward-filtering, backward sampling algorithm for dynamic linear models (West and Harrison, 1997: 569–71), which greatly improves convergence over standard Gibbs sampling (Frühwirth-Schnatter, 1994: 194–5). This approach ‘blocks’ the Gibbs sampler by drawing from a multivariate conditional distribution rather than individual conditional distributions. Drawing directly from a potentially large multivariate distribution (likely of unknown form), though, is impractical, so the multivariate distribution is defined by the product of its conditionals (with the structure defined by the random-walk process):

$$f(\mathbf{x}_i | D_T) = f(\mathbf{x}_{iT} | D_T) f(\mathbf{x}_{i,T-1} | \mathbf{x}_{iT}, D_{T-1}) \cdots f(\mathbf{x}_{i,1} | \mathbf{x}_{i,2}, D_1) f(\mathbf{x}_{i,0} | \mathbf{x}_{i,1}, D_0)$$

where  $D_t$  is the information up to time,  $t$  (latent utility differences, vote parameters, and hyperpriors). The functional form for each component distribution is determined through ‘forward-filtering’ (i.e., begin at period 0, combined with prior beliefs), deriving the distribution of each subsequent ideal point, conditional on previous information. These distributions (as depicted in Martin and Quinn, 2002: 142–5) are normal.

The actual sampling is ‘backwards’, beginning with  $x_{iT}$  and ultimately reaching  $x_{i0}$ . The model is implemented in `MCMCpack`, utilizing their own statistical library `Scythe` for computation (Martin et al., 2011; Pemstein et al., 2011).

Easily downloaded from <http://mqscores.lsa.umich.edu/> (accessed 24 November 2019), the Martin–Quinn scores have become as ubiquitous in the study of the Supreme Court as DW-NOMINATE scores are in the study of Congress, and they appear frequently in popular media (such as *The New York Times* and *The Economist*). Measures of uncertainty are available (posterior standard deviations and quantiles), although similar to DW-NOMINATE, are typically not used.

An alternative estimation approach is implemented in Bailey (2007), with the justices’ ideal points following a polynomial trend, where the degree varies (with the longest serving justices modeled with a polynomial of degree four and other justices modeled with quadratic, linear, or constant preferences). Even more critical is the incorporation of agenda information. For instance, justices in the 1960s might be estimated to be the most liberal over all periods, while some justices who served on later Courts might actually be more liberal. The difficulty is that with many liberal precedents settled in law, current liberal challenges are *extremely* liberal, so much so that justices will inadvertently appear more conservative than earlier justices when they vote against the more liberal position. For instance, the decision to support affirmative action in the 1970s is a more liberal position than simply desegregating schools, and the status quo position of no affirmative action is likely more liberal than *Brown v. Board of Education*, thus opposing such measures should not imply that justices would also support segregation.

Bailey (2007) examines cases related by precedent, coding whether later cases result in more liberal or conservative decisions, relative to earlier cases. Substantive knowledge



of decisions and their precedents is used to establish a relative ordering of the relevant cutpoints. A cutpoint that must be more conservative (greater than) or more liberal (less than) compared to another cutpoint is imposed through rejection sampling at each iteration (the model must be parameterized in terms of the cutpoint/difficulty parameter instead of a vote intercept). Justices' comments on earlier Court cases are coded into positions on the case, as if they were sitting judges at the time of decision.

The ordering of cutpoints can certainly be incorporated into the dynamic linear model approach of Martin and Quinn (2002), and Bailey (2013) does exactly that in order to compare these estimates to the Martin–Quinn scores.<sup>5</sup> The general trends appear, especially the liberal shift in the 1960s, but the Court – in Bailey's estimates – does not shift substantially right in the early 1970s (the more intuitive result, given the number of landmark liberal decisions decided by the Court) and is consistently moderately conservative from the 1980s to 2010 (without the surprising liberal shift in the 2000s). There is also no 'end period' issue from the dynamic estimates. There are some differences in estimates – a different prior for initial state, a standard Gibbs sampling routine for the ideal points, and a restriction of cases to those relating to social policy (i.e., crime, civil rights, free speech, religion, abortion, and privacy), while also incorporating non-spatial predictors into the decision model.

The arguments surrounding the estimation of justices' ideal points over time certainly apply to legislators. Bailey (2007) scales the Supreme Court justices jointly with the president and members of Congress, with those members' ideal points assumed to be constant over time (with the stated positions on specific court cases by members also included in the roll call data). Similarly, Bateman et al. (2017) examine a similar problem of comparability concerning civil-rights issues in Congress from 1877 to 2011. Standard estimates incredulously imply that

recent Congresses are as polarized on civil-rights issues as those just following the Civil War, while Bateman et al.'s (2017) estimates that incorporate ordering constraints demonstrate much less polarization between the contemporary parties.

Bateman et al. (2017) also assume that preferences of members of Congress are constant. The Martin and Quinn model for dynamic changes in ideal points has rarely been applied outside of judicial ideal point estimation. One example is Bailey et al. (2017), who estimate a dynamic model of UN states; it does not implement the forward-filtering, backward sampling approach though, and the number of votes, years, and countries are still much smaller than what one faces with congressional estimates. As highlighted by Martin and Quinn (2002), the sampling of ideal points can be *parallelized*; the ideal points for each justice are independent, so the joint distribution can be sampled using the forward-filtering, backward sampling algorithm for each justice or legislator, up to the number of cores available. Lewis and Tausanovitch (2018) implement the IDEAL model, parallelizing on machines using a NVIDIA graphics card or Amazon's EC2 service, and similar implementations are available for dynamic linear models following Gruber and West (2016).

Furthermore, all instances are unidimensional. Bailey and Voeten (2018) do estimate a multidimensional model but treat the ideal points as independent over time; comparability is established through ordering cutpoints, dimension by dimension, like Bailey (2007) and Bateman et al. (2017). A true multidimensional implementation of this model would be a definite advance, paralleling the specification in West and Harrison (1997: 582–6).

### ***Across Institutions***

One long-running dilemma is that DW-NOMINATE scores for the House and

Senate are not fitted to a comparable scale. To compare across chambers, Poole (1998) developed Common Space Scores (CSS), which place the president, House, and Senate on the same latent scale. Comparability is established by (1) assuming legislators have constant ideal points, even when (2) they switch from House to Senate, or to the president. The potential problem is that Senators represent different constituents to a narrower House district, and the president certainly responds to vastly different pressures than a Representative or Senator. Shor et al. (2010) expand the scope to consider state legislatures, using state legislators who became Representatives or Senators as the bridge.

Alternatively, one can focus on the comparability of votes. The House and Senate have separate agendas, but for a law to be enacted, it has to pass both chambers in identical forms. If these votes are recorded (which often fails for the Senate) one can impose the restrictions of  $\alpha_{j,H} = \alpha_{j,S}$  and  $\beta_{j,H} = \beta_{j,S}$  on each 'bridge' vote  $j$ . The implementation involves stacking the roll call matrices to align with the bridge votes (although this creates even larger roll call matrices with many empty cells – something that most implementations process inefficiently). This approach is utilized in Treier (2011) and appears frequently in educational testing (where an overlap of questions between groups of test-takers facilitates comparison). With additional assumptions, one can expand on this over time. Through cosponsorship, Asmussen and Jo (2016) identify positions on bills introduced in multiple Congresses to provide a common bridge.

An alternative (or additional) approach is to incorporate bridge actors who take positions on votes taken in Congress, such as the president, Supreme Court justices, bureaucrats, and interest groups. This occurs with DW-NOMINATE and CSS scores, with the president's 'roll call' record based on *Congressional Quarterly* positions included in the downloadable data from Voteview. It is also a critical component of Bailey and

Chang (2001) and Bailey (2007), who also code positions on court decisions for the president and members of Congress to create a set of bridge votes on issues before the court.<sup>6</sup> Treier (2010) evaluates the *CQ* positions and suggests supplementing these positions. Bertelli and Grose (2011) add cabinet-level secretaries into the estimation through coding their public positions and incorporating them into the roll call matrix. Treier (2011) utilizes the positions of interest groups rating members of Congress (the Americans for Democratic Action (ADA) and the American Conservative Union (ACU)) to supplement comparisons across chambers and time (assuming the groups have constant preferences). Bonica (2013, 2014) utilizes campaign donors as bridge agents, using their contribution decisions to presumably ideologically similar legislators to recover (non-Bayesian) ideal point estimates for candidates (incumbent and non-incumbent) and donors, while Bonica and Sen (2017) apply Bonica's approach to the legal profession.

Shor and McCarty (2011) supplement the approach by Shor et al. (2010) with responses of the National Political Awareness Test (NPAT) from Project Vote Smart, which survey state and national legislative candidates (incumbents and challengers). They scale the NPAT responses and estimate ideal points for each state and federal legislature separately using Bayesian item-response methods. The legislatures are then projected back to the NPAT space through ordinary least squares regression of the respective coordinates. The project is a common form of bridging, especially in connecting legislatures with other institutions or groups (e.g., voters). In particular, Epstein et al. (2007) map Martin–Quinn scores back to Congress (through a non-linear regression connecting the confirmed justices' first year estimated ideal points and the nominating president's CSS). Bonica (2018) also establishes mappings from the donor-based estimates to NOMINATE (Shor and McCarty, 2011: 533–4), and observes that, ideally, one would estimate these models

jointly, but the computational issues are simply too substantial.

Other applications address fundamental questions about the representative relationship, by placing the mass public and public officials on the same scale. Bridging occurs through surveys of both the public and legislators or surveys where some questions mimic important roll calls in Congress. Among the many analyses along these lines are Jessee (2009) and Bafumi and Herron (2010). The results of these studies are an interesting contrast, since they arrive at completely opposite conclusions: in Bafumi and Herron (2010), voters are more moderate than their representatives and senators, while in Jessee (2009), voters are more extreme. There are differences in surveys, bridging assumptions, and modeling (Jessee, 2009, includes an additional variance component for additional heterogeneity).

Jessee (2016) re-evaluates these examples (but using the 2008 CCES), providing some insight into the challenges of bridging observations. The most critical result is that the characterization of the latent dimension is primarily determined by the larger comparison group, leaving the estimates heavily dependent on assumptions that the mass public and legislators view issues the same way, and that this perspective is homogenous.

## IMPLEMENTATION

The most common implementations appear as model-specific R packages written by researchers (R Core Team, 2016). The IDEAL model is implemented with the `ideal` command in Simon Jackman's (2015) package `pscl`. The package includes extremely useful auxiliary functions for downloading and manipulating roll call data, defining a `rollcall` object that is also used by the Bayesian and classical implementations of NOMINATE in packages `anominat` and `wnominat` (Poole et al., 2011; Carroll et al., 2017).

Martin and Quinn implement the basic IDEAL model in their package `MCMCpack` with `MCMCirt1d` and `MCMCirtKd`, and the dynamic version with the command `MCMCdynamicIRT1d` (Martin et al., 2011). Thus, for the standard Bayesian ideal point model, using either quadratic or Gaussian utility functions, the researcher can simply estimate on their own roll call matrices.

Extending these approaches or including these estimates in a larger structural model (as an independent or dependent variable) is more difficult. `MCMCpack` includes a command that estimates simultaneously the ideal points and a regression model that predicts them with the command `MCMCirtHier1d`. One could also implement a two-stage approach, decoupling the idea-point model from the regression model but still allowing the uncertainty of the ideal point estimates to propagate through the regression (see, e.g., Treier and Jackman, 2008: 215).

More complicated extensions can be implemented in one of several general-purpose Bayesian software packages: `WinBUGS`, `OpenBUGS`, `JAGS`, and `Stan` (Lunn et al., 2000, 2009; Plummer, 2003; Stan Development Team, 2018). These programs expanded the reach of Bayesian methods in general, and their existence has made many of the extensions discussed in this chapter possible. Instead of the command-line-based systems with which many social scientists are most familiar (with `Stata` being the most prominent example), the BUGS approach involves a general model description, where the research simply specifies the likelihood, prior distributions, and conditional (hierarchical) distributions, with parameters and fixed data being either explicitly identified or inferred from the model specification. The original, `WinBUGS`, could piece together the models and step through criteria to apply sampling schemes from the most specific to the most general. It recognizes the canonical conjugate models and samples directly from the posterior and identifies instances where the conditional distribution of parameters

(univariate or multivariate) are conjugate and amendable to Gibbs sampling. For more challenging cases, the BUGS environments break down the posterior into one-dimensional blocks, from which *some* sampling scheme is always possible. The major advance is the implementation of derivative-free adaptive rejection sampling (dfARS) for Gibbs sampling.<sup>7</sup> Applicable to conditional distributions that are log-concave – a common condition in many of our specifications (which often utilize exponential families) – it samples from piecewise linear functions, the number of segments increasing with each rejection until it forms an extremely close approximation to the target Gibbs sampling distributions. For non-log-concave distributions, either slice sampling or one-dimensional random-walk Metropolis–Hastings sampling from a normal distribution occurs.<sup>8</sup>

These general approaches have made numerous innovations possible, e.g., the estimation of the agenda constrained models would be daunting without general modeling framework software. The utility functions do not reduce to the linear forms in standard IRT models, and the status quo and proposal parameters are complicated functions of the other parameters. Other models that adapt the standard model by altering the fundamental specification of utilities for the *Yea/Nay* positions and add non-spatial predictors or hierarchical elements are also easily specified.

An obvious trade-off that typically accompanies such generality is speed. These implementations can be extremely slow so are best reserved for moderate-sized data. The agenda-constrained models are a perfect example, where the focus has typically been on a small subset of roll calls on a specific policy or one with few legislators. A single legislature (or chamber) in a specific sitting is perfectly manageable but inappropriate for much larger or sparse data.

The user-packaged functions available in R implement the specific models

directly (e.g., Gibbs sampling for the IRT model with quadratic utility and roll call voting). The package `pscl` implements C code directly through the `.C` interface in R, while `MCMCpack` created their own statistical C++ library `Scythe` (Pemstein et al., 2011). The availability of the `Rcpp` package has greatly simplified working with C++ functions (Eddelbuettel and Francois, 2011; Eddelbuettel, 2013). Most usefully, one can write and use isolated functions to which R objects can be passed (rather than implementing full C++ programs with `main()` functions). `Rcpp` also links to the powerful scientific computing library `Armadillo` through the `RcppArmadillo` package (Eddelbuettel and Sanderson, 2014; Sanderson and Curtin, 2016).

A compromise between these options involves the `Stan` program, implemented in standalone versions as well as within programs such as R and `Stata` or languages such as Python. `Stan`'s interface follows the BUGS model of general model specification, with the syntax sharing recognizable similarities. `Stan`, however, compiles the model though in C++, providing substantial improvements in speed. Furthermore, it implements the model using a completely different MCMC algorithm. Instead of breaking the posterior into one-dimensional pieces, `Stan` implements a Metropolis–Hastings multivariate sampler, generating proposals using Hamiltonian Monte Carlo (HMC) methods instead of random-walk steps. It calculates a gradient of the posterior and an auxiliary distribution using numerical differentiation, from which a trajectory is generated through a discretized process ('leapfrog integrator'), which must be parameterized through program options or adapted automatically. A standard Metropolis step is taken to reject or accept the draws. The more complex a posterior distribution, the greater the number of steps that may be taken, and both the tuning of the algorithm and the convergence of the approximation must be monitored.

## AGENDA

The ideal point model is estimated based on the reduced-form vote parameters. Working directly with the status quo and proposal parameters  $p_j$  and  $q_j$  is elusive: in one dimension, these parameters are barely identified; in two or more dimensions, the parameters are completely unidentified, and all that is recoverable are the cutlines or cut planes. With NOMINATE, these positions are identified but only due to functional form.

In many applications, researchers are not concerned with these parameters; the specification is a means to the end goal of recovering ideal points. Even if one is only interested in the estimate of legislator ideal points, recovery of the vote positions may define a more accurate and interpretable ideological space. Clinton and Meiowitz (2001) prove that ideal point estimates that exclude such relational restrictions will underestimate the dimensionality. Pope and Treier, e.g., do not analyze the positions directly but still incorporate information on the agenda at the Constitutional Convention in order to determine the dimensionality of delegation positions (Pope and Treier, 2012) and provide a better understanding of the ideological dimensions (Pope and Treier, 2015).

Recall, the utilities in terms of the proposal and status quo policies are

$$\begin{aligned} U(\mathbf{q}_j) &= \sum_{k=1}^K -(x_{ki} - q_{kj})^2 + \varepsilon_{i,\mathbf{q}_j} \\ &= -\|\mathbf{x}_i - \mathbf{q}_j\|^2 + \varepsilon_{i,\mathbf{q}_j} \\ U(\mathbf{p}_j) &= \sum_{k=1}^K -(x_{ki} - p_{kj})^2 + \varepsilon_{i,\mathbf{p}_j} \\ &= -\|\mathbf{x}_i - \mathbf{p}_j\|^2 + \varepsilon_{i,\mathbf{p}_j} \end{aligned}$$

Clinton and Meiowitz (2001, 2003, 2004) estimate the status quo policies by incorporating changes (or lack of changes) in the

status quo directly into the estimation. First, if proposal  $\mathbf{p}_j$  passes, then the new status quo point equals this proposal:  $\mathbf{q}_{j+1} = \mathbf{p}_j$ . If  $\mathbf{p}_j$  fails, then the status quo remains unchanged:  $\mathbf{q}_{j+1} = \mathbf{q}_j$ . Second, if a proposal does not concern dimension  $d$ , then even if  $\mathbf{p}_j$  passes, the coordinates of the status quo for irrelevant dimensions will remain unchanged:  $q_{j+1,d} = q_{j,d}$ . By imposing these constraints, one can directly estimate the status quo points and proposals and effectively see the unfolding agenda. The parameters  $\mathbf{q}_j$  and  $\mathbf{p}_j$  are represented as  $\boldsymbol{\theta}$ , indexed by  $y(j)$  and  $n(j)$ , where  $\mathbf{q}_j = \boldsymbol{\theta}_{n(j)}$  and  $\mathbf{p}_j = \boldsymbol{\theta}_{y(j)}$ .

The estimated model, in utility differences, is

$$y_{ij}^* = \|\mathbf{x}_i - \boldsymbol{\theta}_{n(j)}\|^2 - \|\mathbf{x}_i - \boldsymbol{\theta}_{y(j)}\|^2 + \varepsilon_{ij} \quad (8)$$

which does not simplify into a reduced-form structural model; consequently, a hybrid Gibbs approach is implemented, where each conditional step follows a Metropolis–Hastings approach, with a normal distribution centered at the previous estimate and a variance that has been tuned.

Clinton and Meiowitz (2004) apply this approach to the question of the famous ‘Dinner Party Bargain’, where reportedly the details of a log-roll were worked out at a party hosted by Thomas Jefferson, regarding the location of the national capital and the federal assumption of state debt during the first Congress. Before the dinner-party agreement, every vote loads only on one dimension, and afterwards, votes load on both dimensions. The model is identified by fixing the status quo position to (0, 0), fixing proposals on each issue and Madison to constants, as well as identical proposals before and after the supposed bargaining. Clinton and Meiowitz (2004) hypothesize that once a deal had been perceived, one would expect the unconstrained passage of the bill on capital location to represent a shift both towards a Southern capital and greater assumption of debt; their estimates, however, illustrate the critical vote clearly only involves the capital.

Further, continual movements of the status quo after this vote contradict the possibility of a settled deal. Their analysis depicts perfectly the greatest strengths of the Bayesian approach. First, extending the basic model to include agenda information – which leaves the utility in its most general structural form but includes enhanced indexing of the data – is simply estimated by an application of the general Metropolis–Hastings algorithm. Second, summarizing uncertainty – graphically, with two-dimensional 95% posterior intervals, and by calculating the probability that any two positions will be distinct (e.g., assumption dimension of proposal 2 is greater than proposal 1 and the location dimension is lesser) – is a simple calculation from the MCMC samples.

Similar to Clinton and Meirowitz, Pope and Treier (2011) map the agenda for the Constitutional Convention (a quasi-legislative body), defining the initial status quo as the Articles of Confederation, with changes being considered section by section in the Virginia plan and subsequent action through the Great Compromise. In contrast to the Dinner Bargain example though, the 92 votes are defined on more general dimensions (i.e., representation and scope of Government) rather than specific issues. Model checks are conducted through test statistics created by sampling from the posterior predictive distribution (Sinharay et al., 2006). Another advantage of the Bayesian approach is the straightforward extensibility of the model. Complicating the roll call analysis are split-state delegations resulting in divided votes. The model is extended by treating division as the result of ‘thick indifference’, when the cutpoint or cutline is close to the ideal point, incorporating the model of Sanders (1998, 2001).

These historical examples are perfect for this approach, as proposed changes in the agenda are focused and easily categorized. In modern legislatures, however, one is faced with juggling many agendas with bills thousands of pages long and complex amendments

and agenda maneuvers. In contemporary legislatures, this approach works best in single-policy areas, focusing on the most important proposals in the agenda. Clinton (2012) follows this approach with the Fair Labor Standards Act, tracing proposed and successful changes to the legislation from 1971 to 2000 (primarily in the 1970s and 1990s) for 114 votes with 112 vote parameters. Members are assumed to have constant preferences (unless they switched party or chamber), with 16 members (nine representatives, seven senators) appearing throughout the period.

Gyung-Ho Jeong has applied this model to a myriad of modern applications: considering the creation of the Federal Reserve System (Jeong, 2008; Jeong et al., 2009); joining the League of Nations (Jeong, 2017); civil-rights legislation (Jeong et al., 2009); immigration (Jeong et al., 2011; Jeong, 2012); and energy policy (Jeong et al., 2014). Each of these applications consider very specific policy issues that have a small set of important votes with a discernible agenda structure. In these applications, Jeong relies on modern procedure (e.g., rules specifying ‘perfecting’ amendments that change a dimension at a time followed by substitution amendments typically combining earlier proposals on each dimension) to substantially reduce the number of parameters considered. Most impressive of the analyses and illustrative of the simplicity of Bayesian analysis, are the complex calculations (with uncertainty bounds) of the uncovered set, using the estimated ideal points with the algorithm of Bianco et al. (2004), overlaid with the status quo and proposal locations.

## HIERARCHICAL ESTIMATION

One of the most obvious advantages to the Bayesian framework is its natural inclusion of hierarchical structures, especially those concentrated around multilevel data. The basic ideal point model can be extended by

incorporating covariates as predictors for  $x_i$  in a hierarchical specification:

$$x_i = \mathbf{z}\boldsymbol{\gamma} + v_i$$

Without covariates  $\mathbf{z}$ , the model reduces to the standard ideal point model with  $x_i = v_i$ , in one dimension. Rotational invariance of the model can be imposed by restricting the coefficient of a well selected covariate to be positive; the location and scale of the model can be identified by standardizing the ideal points and applying the appropriate transformations to the other parameters of the model. Bafumi et al. (2005) illustrate this setup with the ideal points of Supreme Court justices, partly determined by the party of the nominating president. Zucco Jr. and Lauderdale (2011) institute a unique design by estimating a two-dimensional IDEAL model where each dimension is centered on a party mean, with the first (ideological) dimension informed through legislator responses to a survey (where they place themselves and other parties) and a standard normal hyperprior for the party mean on the second dimension. They recover two dimensions that clearly reflect an ideological motivation for legislators and a government–opposition dynamic on the second dimension. Imai et al. (2016) formulate a very general hierarchical model specification, then estimate a special case of the model: an IDEAL model over time with the same polynomial specification as DW-NOMINATE.

More generally, the IRT model can be extended through regression relations for the discrimination and difficulty parameters: Bafumi et al. (2005: 178) formulate this as

$$x_i \sim N(\mathbf{Z}\boldsymbol{\gamma}, \sigma_x) \quad \forall i = 1, \dots, n$$

$$\beta_j \sim N(\mathbf{W}\boldsymbol{\alpha}, \sigma_\beta^2) \quad \forall j = 1, \dots, J$$

$$\kappa_j \sim N(\mathbf{K}\boldsymbol{\lambda}, \sigma_\alpha^2) \quad \forall j = 1, \dots, J$$

with respect to the specification  $y^* = \beta_j(x_i - \kappa_j) + \varepsilon_{ij}$ . There are no general considerations of this model but there are considerations in educational testing and

psychometric literatures (in particular, see Fox, 2010: ch. 6). The primary concern for ideal point estimation is how the hierarchical or multilevel model relate to the utility model. A change in the cutpoint parameterization will alter the interpretation of the item parameters, and a hierarchical prior on the discrimination parameter will alter the considerations between status quo and proposal.

While hierarchical models for the ideal points are more common, there are a few examples of other hierarchical structures. Bailey and Maltzman (2008) specify a standard ideal point model for case votes on the Supreme Court but include non-spatial factors that directly impact the probability of voting for the plaintiff:  $U(y_{ij} = \text{'Yea'}) = -(x_i - p_j)^2 + \delta_i \text{Law}_j + \eta_{ij}$ . Note, the additional factors define an intercept shift not related to the spatial model (the three case variables and judge-specific coefficients relate to issues of precedent, deference, and speech).

## OTHER DATA

Ideal points have been estimated almost exclusively from roll call data. One obvious reason is the widespread availability of such data, the high participation of and separation between legislators, and the clear position taking motivations for legislators. But roll calls can present some difficult limitations. Strict agenda control can result in the roll calls taken only reflecting parts of the basic-issue space, as other measures are simply blocked (Clinton 2007). A further complication is the presence in many legislatures of strict party discipline. Systems where partisans never or almost never deviate from the position of their party leadership produce roll call records that are unable to differentiate between members of the same party (even when there are clear perceptions of different ideological positions) and merely separate out the governing party from the opposition (Spirling and McLean 2007; Dewan and Spirling 2011).

Consequently, alternative sources of data have been considered as supplements or replacements to roll calls. An alternative is cosponsorships. These are an obvious form of public position taking on measures that have little to no party whip activity, and while many bills will be excluded from the agenda, legislators are still able to take positions on proposals which likely span the entire basic-issue space.

While cosponsoring indicates clear support, the effect of not cosponsoring is ambiguous. Some researchers do treat the cosponsorship data exactly like roll call data, with non-sponsorship being equivalent to opposition and standard implementations of W-NOMINATE or IDEAL being applied. Asmussen and Jo (2016) more conservatively only consider cosponsorship, while treating non-sponsorship as missing. Kellerman (2012), using early day motions to estimate positions from the UK House of Commons, treats the data as regular roll call data but adds an additional legislator intercept to the utility of signing an EDM to reflect the professional *cost* of stating a public position. Desposato et al. (2011) derive a variant of the roll call model that treats the utility of the implied proposal or status quo the same but incorporates the probability of whether or not one *considers* the cosponsoring.

Desposato et al. (2011) consider different specifications, including a constant factor and a ‘neighbor’ model that reflects the distance between a legislator and the original sponsor of the bill. Ultimately, though, they only test their implementation on simulated data (cosponsoring data is much larger than roll call data).

Cosponsorship defined as a complete social network is common, but the ideological element has not factored substantially. Fowler (2006), using a massive dataset covering 1973 to 2004, relies on network summary statistics, while Alemán et al.’s (2009) calculation of cosponsorship ideal points relies exclusively on the decomposition of a network agreement matrix.

An actual ideological model is applied to a different sort of network data: Twitter followers. Barberá (2015) specifies a *latent space* network model, following Hoff et al. (2002). This model considers Twitter user  $i$  deciding whether or not to follow politician  $j$  as

$$\text{logit}\left(\Pr(y_{ij} = 1 \mid \theta)\right) = \alpha_j + \beta_i - \gamma \|x_i - x_j\|^2$$

i.e., it is determined by the spatial distance between potential follower and elite (note that this also holds for elites following other elites), the general popularity of elite  $j$  and level of political engagement for user  $i$ . Six countries are analyzed (separately), with the general Twitter sample of users ranging from approximately 50,000 to 300,000 and between 118 and 318 political actors. The estimation is in two stages: a `Stan` model of the political actors alone (i.e., the basic-issue space will be first defined by their mutual affinity) and then a second model for general Twitter users’ decisions to follow elites. The estimates for members of Congress match DW-NOMINATE scores well, with the placements of other prominent elites appearing as expected. The latent-space model also seems quite appropriate for the application, where similarity (or *homophily*) is the overwhelming motivation for following.<sup>9</sup>

## IDEAL POINTS AND TEXT DATA

As we saw in the previous section, the ideal point model has been intrinsically linked with roll call voting, but a variety of behaviors could be related to ideology. A natural candidate, infeasible until very recently, is incorporating speech data. Differentiation through speech patterns allows greater intraparty distinction, even in the presence of strict party discipline (Schwarz et al., 2017). Furthermore, since the language of policy remains similar



through the period of analysis, this text-data approach can connect legislatures over time. Supplementing observed votes with speeches that occur in assembly is one of the latest methodological frontiers.

The first notable latent-trait model using text data was the non-Bayesian WORDFISH (Slapin and Proksch, 2008). This approach, applied to party manifestos, was a direct response to the supervised Wordscores approach (Laver et al., 2003). Wordscores provided estimates of location through comparison to two reference documents. The ideological positions of the reference documents were fixed, with all other party manifestos assigned a position based on their similarity to the references documents. The resulting score was a convex combination of the reference positions, based on the similarity between documents. Of course, a major limitation of this approach was that these estimates were guaranteed to take more moderate positions than the reference documents, and all attempts to alleviate the problem were merely unjustified rescalings. Wordfish, on the other hand, is based on a latent-trait approach for the counts of each word  $j$  for unit  $i$ ,  $y_{ij}^*$ , which is distributed Poisson with mean  $y_{ij}^*$ , the individual's latent word 'emphasis'. The parameterization is inspired by the IRT model,

$$y_{ij}^* = \gamma_i + \alpha_j + \beta_j x_i$$

with  $x_i$ , the ideological position and intercepts for unit and word, related to verbosity and aptness. This reduced form is borrowed directly from statistical and psychometric literature, without a structural representation of the ideal point model.

Kim et al. (2018) provide such a derivation from the spatial model for word choice by legislators in speeches and connect to the RUM over voting. This work is the most promising approach to combining votes and speeches into comparable measures of ideology. The roll call voting model is standard

Euclidean, with differential weights  $a_d$  for dimension  $d$ . These weights also appear in the word-choice model, where they represent the utility for emphasis  $y_{iw}^*$  for legislator  $i$  over word  $w$  (distinguishing from vote  $j$ ) varying proportionally with spatial and non-spatial terms, along with a penalty for word overuse and random error:

$$U_{iw}(y_{iw}^*) = y_{iw}^* \left( s_w + v_i - \frac{1}{2} y_{iw}^* - \frac{1}{2} \sum_{d=1}^D a_d (x_{id} - g_{wd})^2 + \varepsilon_{iw} \right)$$

Optimizing this utility in terms of  $y_{iw}^*$  results in a choice of

$$\begin{aligned} y_{iw}^* &= v_i + s_w \\ &- \frac{1}{2} \sum_{d=1}^D a_d (x_{id}^2 - 2x_{id}g_{wd} + g_{wd}^2) + \varepsilon_{iw} \\ &= \left( v_i - \frac{1}{2} \sum_{d=1}^D a_d x_{id}^2 \right) + \left( s_w - \frac{1}{2} \sum_{d=1}^D a_d g_{wd}^2 \right) \\ &+ \sum_{d=1}^D a_d g_{wd} x_{id} + \varepsilon_{iw} \\ &= c_i + b_w + \sum_{d=1}^D a_d g_{wd} x_{id} + \varepsilon_{iw} \end{aligned}$$

which corresponds to the Wordfish representation.<sup>10</sup> Instead of assuming a Poisson distribution, the word counts are connected to the latent emphases through an ordered probit model, where

$$\Pr(y_{iw} = k) = \Phi(\tau_{k-1} < y_{iw}^* < \tau_k)$$

with the thresholds being constant across words and individuals and estimated semi-parametrically from the empirical cumulative distributions of the words. The approach accounts for the large number of zero counts in the data, guaranteeing that the probability of no usage matches the observed frequencies.

Kim et al. (2018) also include the standard roll call model on votes, with one major extension: while the random errors in the utilities remain normal distributions with equal variances, the mean of each error differs. The specification is linear additive, with  $E(\varepsilon_{ij}^y) = \pi_i^y + \psi_j^y$  and  $E(\varepsilon_{ij}^n) = \pi_i^n + \psi_j^n$ , for individual random effects  $\pi_i = \pi_i^y - \pi_i^n$  and bill random effects  $\varphi = \varphi_j^y - \varphi_j^n$ . The translation from structural to reduced form is<sup>11</sup>

$$\begin{aligned} & U_{ij}(\text{'Yea'}) - U_{ij}(\text{'Nay'}) \\ &= -\frac{1}{2} \sum_{d=1}^D a_d (x_{id} - p_{jd})^2 \\ &\quad + \frac{1}{2} \sum_{d=1}^D a_d (x_{id} - q_{jd})^2 + (\varepsilon_{ij}^y - \varepsilon_{ij}^n) \\ &= (\pi_i^y - \pi_i^n) \\ &\quad + \left( (\varphi_j^y - \varphi_j^n) + \frac{1}{2} \sum_{d=1}^D a_d (q_{jd}^2 - p_{jd}^2) \right) \\ &\quad + \sum_{d=1}^D a_d (p_j - q_j) x_{id} + \tilde{\varepsilon}_{ij} \\ &= c_i^y + b_j^y + \sum_{d=1}^D a_d g_j^y x_{id} + \tilde{\varepsilon}_{ij} \end{aligned}$$

Again, there is a slight disconnect, with  $a_d$  also appearing in the intercepts, and the motivation for legislator random effects is partially suspect;<sup>12</sup> but Kim et al. (2018) present an impressive model that combines both individual speeches and roll call votes. All of the parameters, except for the term emphasis, cutpoints and the dimension weights are Gibbs sampled, with the semi-parametric approach to the cutpoints implemented through Hamiltonian methods. The conditional distribution of the dimensional weights  $a_d$  are recovered through a hierarchical Gibbs sampler for the Bayesian Lasso (Park and Casella, 2008); the shrinkage in these coefficients towards zero determines

the dimension from the model. The number of dimensions is determined within the model and is highly sensitive to whether text or votes dominate, so a weighting parameter  $\alpha$  between the two components is specified *a priori*.<sup>13</sup> This arbitrariness of dimensionality might suggest more substantive constraints on dimensions, where the core dimensions are defined *a priori*, associated with particular votes, with residual dimensions sorted out in the analysis.

Another promising aspect is that it can be applied to legislatures with strong party discipline. These are cases where standard approaches using roll call votes fail, since party members will have identical voting patterns, outside of absences and the occasional free vote. In their model, one might use only roll call data from the party leadership<sup>14</sup> and the speech text of all individuals, which dominates in the estimation.

A few alternative approaches utilize text in a different manner. Several focus on *bill* text instead of speeches; in these approaches, the bill text determines the substantive content of the vote. Gerrish and Blei (2011, 2012) apply a topic model to bill text, in order to determine the proportion of each issue considered on each vote. They estimate a ‘general’ or ‘baseline’ ideology on one dimension, with issue ‘offsets’, assuming that legislators’ positions on the issues may deviate from the baseline but are still anchored to their general ideological leanings. For Lauderdale and Clark (2014), the text data of opinions on the court determine the proportion of each issue on each case, similar to Gerrish and Blei (2011, 2012). Unlike Gerrish and Biel though, Lauderdale and Clark (2014) estimate multiple dimensions, with the number being determined by the results of the topic model; they settle on 24, according to predictive behavior. Instead of treating votes as inducing a series of offsets from the same dimension, they recover positions on each of the identified dimensions.

The use of bill text in this way supplements the standard ideal point model, with Gerrish and Blei (2011, 2012) providing an intriguing connection between general ideological spaces and specific issues and Lauderdale and Clark (2014) effectively automating the inclusion of substantive information that is typically mechanical. An approach currently unexplored, though, is treating the bill text as *speech* data, informing the position of the proposal in the agenda-structure framework. That adds substantial complications to the model of Kim et al. (2018), but it may help to define the dimensions while also relaxing the requirement of identifying the entire agenda structure to recover proposal/status quo positions.

A final concern, general to all ideal point estimation, is the interpretation of the latent space. This concern may be even more pressing with text data, as the use of particular words may have clearer non-ideological intentions. Grimmer and Stewart (2013) illustrate the potential for ideal point approaches to recover non-ideological divisions that contradict validity checks (such as a reasonable separation of ideological distinct partisans). They present two examples, one where an ideological scaling makes sense (regarding German party platforms), and another that simply recovers language style from Senate press releases (Grimmer and Stewart, 2013: 292–4). Lauderdale and Herzog (2016) advance a solution (‘Wordshoal’) based on the Wordfish latent-variable model, but it is applied to each debate separately before aggregation into general ideological scores. The objective of debate-based estimates is to limit the effects of variation based on language considerations, stylistic choices, and topic relevance, in order to avoid the problem of conflation, as highlighted by Grimmer and Stewart (2013). Their approach of focusing on the debate as the unit of analysis is conceptually similar to the vote-based approaches of Gerrish and Blei (2011, 2012) and others, with the relationship of issue-specific positions ordered hierarchically to a general ideological position of commonality.

## CONCLUSION

Bayesian ideal point estimation facilitates the mundane and the spectacular. With the development of the Bayesian approach, one could run models on a desktop in only a few days (eventually a few hours), regularly provide visual accounts of the uncertainty in ideal points, test hypotheses using auxiliary quantities whose frequentist distribution would have been impossible to contemplate, conceive of extensions to the model, and be able to code up quickly and run them in BUGS/JAGS/Stan. The basics of the model were detailed in the first half of this chapter.

These mundane aspects of estimation have been critical to the application of ideal points, but I am particularly excited by the potential for spectacular advances within this framework, with some really exciting work in development. In this chapter, some of those changes have been detailed. There have been advances in exploring new sources of data and combining with traditional sources, greater attention to comparability, and advances in computation, with faster procedures and bigger data. The developments in the last section particularly highlight these advances and challenges. It seems to be exciting new territory, promising greater understanding of legislatures while simultaneously struggling with frustrating perennial issues. But underneath, the basic framework of CJR is always visible and the geometry of Poole and Rosenthal in the mind, while the common grammar of Bayesian estimation and inference are ever familiar.

## Notes

- 1 That it is acceptable for the underlying motivations of these belief systems to be vague is important. There is no overwhelming demand, for the most part, that the placements uncovered by the measurement models distinguish between, say, personal preferences, constituent demands, or partisan pressures, just that the resulting measures reveal their constraints and predict their behavior.

- 2 Of course, one disadvantage of this approach is characterizing uncertainty around the posterior modes. To optimize an intractable function, variational approximation estimates the function with the product of independent distributions. While this provides excellent coverage of the mode, it misses the tails of the distribution and thus underestimates the posterior variance.
- 3 Alternatively, Clinton (2007) describes how strong partisan agenda control could also bias estimates.
- 4 These deficiencies make frequent appearances in the literature on Congress and are discussed by Asmussen and Jo (2016).
- 5 Bailey also summarizes this research in a *Mischiefs of Faction* post, 'Just how liberal (or conservative) is the Supreme Court?', 22 February 2016, <https://www.vox.com/mischiefs-of-faction/2016/2/22/11094172/supreme-court-conservative-liberal> (accessed 24 November 2019).
- 6 While not used to identify members' position changes over time, announced positions do provide an improved basis for comparison by comparing judges, presidents, and members of Congress taking positions on the same cases over time
- 7 For an extremely intuitive summary of the approach, see Bolstad (2010: 35–42).
- 8 The variance for the proposal distribution is *adapted*; the researcher simply must recognize that the first 4,000 samples in `WINBUGS` could not be utilized. Similarly, there was an adaptive period of 500 samples for slice sampling.
- 9 In other contexts though, the model might face criticism over misspecification by Minhas et al. (2019), who argue that the LSM is unable to distinguish third-order relations; that might be important if one was trying to separate out ideological affinity from party affiliation.
- 10 Note, there is a disconnect between the structural and reduced forms, since the dimension weights  $a_{id}$  and ideal points  $x_{id}$  are part of the intercepts, yet their appearance in these terms is not accounted for in the estimation. The practical consequences are perhaps minimal, other than complicating the interpretation of these parameters and rendering the cutlines in the voting model unidentifiable; but future researchers might consider whether or not this possibly biases the estimates of the weights and/or ideal points.
- 11 Derived in the supplemental materials of Kim et al. (2018).
- 12 In the word-choice model, this is interpreted as talkativeness. For the voting model, this term reflects a propensity to vote 'Yea'. Outside of a government/opposition framework, it is difficult to justify (within a government/opposition

framework, one votes 'Yea' because they are part of the governing party, who determines what votes take place and the measure relates to party coherence).

- 13 In the likelihood, term choice and vote are independent, and the combined likelihood is separated in a 'melded' format, weights  $\frac{W+J}{2J}$  and  $\frac{W+J}{2W}$ , which rebalances the likelihood contributions of words and votes to 50/50. Kim et al. (2018) add a user-fixed term  $\alpha$  which allows a different weighting between 0 and 1 and suggestions for determining  $\alpha$ .
- 14 Presumably, 'party leadership' will be the government, opposition, and any disciplined crossbench parties, as well as individual independent legislators and members of ill-disciplined parties.

## REFERENCES

- Albert, James H. 1992. 'Bayesian Estimation of Normal Ogive Item Response Curves Using Gibbs Sampling'. *Journal of Educational Statistics* 17(3): 251–269.
- Albert, James H. and Siddhartha Chib. 1993. 'Bayesian Analysis of Binary and Polychotomous Response Data'. *Journal of the American Statistical Association* 88(422): 669–679.
- Albert, Jim and Siddhartha Chib. 1995. 'Bayesian Residual Analysis for Binary Response Regression Models'. *Biometrika* 82(4): 747–759.
- Alemán, Eduardo, Ernesto Calvo, Mark P. Jones and Noah Kaplan. 2009. 'Comparing Cosponsorship and Roll-Call Ideal Points'. *Legislative Studies Quarterly* 34(1): 87–116.
- Asmussen, Nicole and Jinhee Jo. 2016. 'Anchors Away: A New Approach for Estimating Ideal Points Comparable across Time and Chambers'. *Political Analysis* 24(2): 172–188.
- Bafumi, Joseph and Michael C. Herron. 2010. 'Leapfrog Representation and Extremism: A Study of American Voters and Their Members in Congress'. *American Political Science Review* 104(3): 519–542.
- Bafumi, Joseph, Andrew Gelman, David K. Park and Noah Kaplan. 2005. 'Practical Issues in Implementing and Understanding Bayesian Ideal Point Estimation'. *Political Analysis* 13(2): 171–187.

- Bailey, Michael A. 2007. 'Comparable Preference Estimates across Time and Institutions for the Court, Congress, and Presidency'. *American Journal of Political Science* 51(3):433–448.
- Bailey, Michael A. and Kelly H. Chang. 2001. 'Comparing Presidents, Senators, and Justices: Interinstitutional Preference Estimation'. *Journal of Law, Economics, and Organization* 17(2): 477–506.
- Bailey, Michael A. and Forrest Maltzman. 2008. 'Does Legal Doctrine Matter? Unpacking Law and Policy Preferences on the U.S. Supreme Court'. *American Political Science Review* 102(3): 369–384.
- Bailey, Michael A. 2013. 'Is Today's Court the Most Conservative in Sixty Years? Challenges and Opportunities in Measuring Judicial Preferences'. *Journal of Politics* 75(3): 821–834.
- Bailey, Michael A. and Erik Voeten. 2018. 'A Two-Dimensional Analysis of Seventy Years of United Nations Voting'. *Public Choice* 176(1–2): 33–55.
- Bailey, Michael A., Anton Strezhnev and Erik Voeten. 2017. 'Estimating Dynamic State Preferences from United Nations Voting Data'. *Journal of Conflict Resolution* 61(2):430–456.
- Barberá, Pablo. 2015. 'Birds of the Same Feather Tweet Together: Bayesian Ideal Point Estimation Using Twitter Data'. *Political Analysis* 23(1): 76–91.
- Bateman, David A., Joshua D. Clinton and John S. Lapinski. 2017. 'A House Divided? Roll Calls, Polarization, and Policy Differences in the U.S. House, 1877–2011'. *American Journal of Political Science* 61(3): 698–714.
- Bertelli, Anthony M. and Christian R. Grose. 2011. 'The Lengthened Shadow of Another Institution? Ideal Point Estimates for the Executive Branch and Congress'. *American Journal of Political Science* 55(4): 767–781.
- Bianco, William T., Ivan Jeliaskov and Itai Sened. 2004. 'The Uncovered Set and the Limits of Legislative Action'. *Political Analysis* 12(3): 256–276.
- Bolstad, William M. 2010. *Understanding Computational Bayesian Statistics*. Hoboken, NJ: John Wiley & Sons.
- Bonica, Adam. 2013. 'Ideology and Interests in the Political Marketplace'. *American Journal of Political Science* 57(2): 294–311.
- Bonica, Adam. 2014. 'Mapping the Ideological Marketplace'. *American Journal of Political Science* 58(2): 367–386.
- Bonica, Adam. 2018. 'Inferring Roll-Call Scores from Campaign Contributions Using Supervised Learning'. *American Journal of Political Science* 62(4): 830–848.
- Bonica, Adam and Maya Sen. 2017. 'A Common-Space Scaling of the American Judiciary and Legal Profession'. *Political Analysis* 25(1): 114–121.
- Carroll, Royce, Jeffrey B. Lewis, James Lo, Keith T. Poole and Howard Rosenthal. 2009. 'Comparing NOMINATE and IDEAL: Points of Difference and Monte Carlo Tests'. *Legislative Studies Quarterly* 34(4): 555–591.
- Carroll, Royce, Jeffrey B. Lewis, James Lo, Keith T. Poole and Howard Rosenthal. 2013. 'The Structure of Utility in Spatial Models of Voting'. *American Journal of Political Science* 57(4): 1008–1028.
- Carroll, Royce, Christopher Hare, Jeffrey B. Lewis, James Lo, Keith T. Poole and Howard Rosenthal. 2017. *Alpha-NOMINATE: Ideal Point Estimator*. R package version 0.6.
- Clinton, Joshua D. 2007. 'Lawmaking and Roll Calls'. *Journal of Politics* 69(2): 457–469.
- Clinton, Joshua D. 2012. 'Congress, Lawmaking, and the Fair Labor Standards Act, 1971–2000'. *American Journal of Political Science* 56(2): 355–372.
- Clinton, Joshua D. and Adam Meirowitz. 2001. 'Agenda Constrained Legislator Ideal Points and the Spatial Voting Model'. *Political Analysis* 9(3): 242–259.
- Clinton, Joshua D. and Adam Meirowitz. 2003. 'Integrating Voting Theory and Roll Call Analysis: A Framework'. *Political Analysis* 11(4): 381–396.
- Clinton, Joshua D. and Adam Meirowitz. 2004. 'Testing Explanations of Strategic Voting in Legislatures: A Reexamination of the Compromise of 1790'. *American Journal of Political Science* 48(4): 675–89.
- Clinton, Joshua D. and Simon Jackman. 2009. 'To Simulate or NOMINATE?' *Legislative Studies Quarterly* 34(4): 593–621.
- Clinton, Joshua D., Simon Jackman and Douglas Rivers. 2004. 'The Statistical Analysis of Legislative Roll Call Data'. *American Political Science Review* 98(2): 355–370.

- de Jong, Edwin D., Marco A. Wiering and Mădălina M. Drugan. 2003. *Post-Processing for MCMC*. Technical Report UU-CS-2003-021, Utrecht University: Institute of Information and Computing Sciences.
- Desposato, Scott W., Matthew C. Kearney and Brian F. Crisp. 2011. 'Using Cosponsorship to Estimate Ideal Points'. *Legislative Studies Quarterly* 36(4): 531–565.
- Dewan, Torun and Arthur Spirling. 2011. 'Strategic Opposition and Government Cohesion in Westminster Democracies'. *American Political Science Review* 105(2): 337–358.
- Eddelbuettel, Dirk. 2013. *Seamless R and C++ Integration with Rcpp*. New York: Springer.
- Eddelbuettel, Dirk and Conrad Sanderson. 2014. 'RcppArmadillo: Accelerating R with High-Performance C++ linear algebra'. *Computational Statistics and Data Analysis* 71: 1054–1063.
- Eddelbuettel, Dirk and Romain Francois. 2011. 'Rcpp: Seamless R and C++ Integration'. *Journal of Statistical Software* 40(8):1–18.
- Epstein, Lee, Andrew D. Martin, Jeffrey A. Segal and Chad Westerland. 2007. 'The Judicial Common Space'. *Journal of Law, Economics, and Organization* 23(2): 303–325.
- Fowler, James H. 2006. 'Connecting the Congress: A Study of Cosponsorship Networks'. *Political Analysis* 14(4): 456–487.
- Fox, Jean-Paul. 2010. *Bayesian Item Response Modeling*. New York: Springer.
- Frühwirth-Schnatter, Sylvia. 1994. 'Data Augmentation and Dynamic Linear Models'. *Journal of Time Series Analysis* 15(2):183–202.
- Gelman, Andrew, Xiao-Li Meng and Hal Stern. 1996. 'Posterior Predictive Assessment of Model Fitness via Realized Discrepancies'. *Statistica Sinica* 6(4): 733–807.
- Gerrish, Sean and David M. Blei. 2011. Predicting Legislative Roll Calls from Text. In *Proceedings of the 28<sup>th</sup> International Conference on Machine Learning*, Bellevue, Washington, USA – 28 June to 2 July 2011. Eds L. Getoor and T. Scheffer, pp. 489–496. <http://www.icml-2011.org> (accessed 24 November 2019).
- Gerrish, Sean and David M. Blei. 2012. How They Vote: Issue-Adjusted Models of Legislative Behavior. In *Advances in Neural Information Processing Systems*, eds F. Pereira, C.J.C. Burges, L. Bottou and K.Q. Weinberger, vol. 25. Red Hook, NY: Curran Associates, Inc. pp. 2753–2761.
- Grimmer, Justin. 2011. 'An Introduction to Bayesian Inference via Variational Approximations'. *Political Analysis* 19(1): 32–47.
- Grimmer, Justin and Brandon M. Stewart. 2013. 'Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts'. *Political Analysis* 21(3): 267–297.
- Gruber, Lutz and Mike West. 2016. 'GPU-Accelerated Bayesian Learning and Forecasting in Simultaneous Graphical Dynamic Linear Models'. *Bayesian Analysis* 11(1): 125–149.
- Hoff, Peter D., Adrian E. Raftery and Mark S. Handcock. 2002. 'Latent Space Approaches to Social Network Analysis'. *Journal of the American Statistical Association* 97(460): 1090–1098.
- Imai, Kosuke, James Lo and Jonathan Olmsted. 2016. 'Fast Estimation of Ideal Points with Massive Data'. *American Political Science Review* 110(4):631–656.
- Jackman, Simon. 2009. *Bayesian Analysis for the Social Sciences*. Hoboken, NJ: John Wiley & Sons.
- Jackman, Simon. 2015. 'pscl: Classes and Methods for R Developed in the Political Science Computational Laboratory, Stanford University'. R package version 1.4.9. Stanford: Department of Political Science, Stanford University.
- Jeong, Gyung-Ho. 2008. 'Testing the Predictions of the Multidimensional Spatial Voting Model with Roll Call Data'. *Political Analysis* 16(2):179–196.
- Jeong, Gyung-Ho. 2012. 'Congressional Politics of U.S. Immigration Reforms: Legislative Outcomes Under Multidimensional Negotiations'. *Political Research Quarterly* 66(3):600–614.
- Jeong, Gyung-Ho. 2017. 'The Supermajority Core of the U.S. Senate and the Failure to Join the League of Nations'. *Public Choice* 173(3–4): 325–343.
- Jeong, Gyung-Ho, Gary J. Miller and Andrew Sobel. 2009. 'Political Compromise and Bureaucratic Structure: The Political Origins of the Federal Reserve System'. *Journal of Law, Economics, and Organization* 25(2): 472–498.
- Jeong, Gyung-Ho, Gary J. Miller, Camilla Schofield and Itai Sened. 2011. 'Cracks in the

- Opposition: Immigration as a Wedge Issue for the Reagan Coalition'. *American Journal of Political Science* 55(3): 511–525.
- Jeong, Gyung-Ho, Gary J. Miller and Itai Sened. 2009. 'Closing the Deal: Negotiating Civil Rights Legislation'. *American Political Science Review* 103(4): 588–606.
- Jeong, Gyung-Ho, William R. Lowry, Gary J. Miller and Itai Sened. 2014. 'How Preferences Change Institutions: The 1978 Energy Act'. *Journal of Politics* 76(2): 430–445.
- Jessee, Stephen A. 2009. 'Spatial Voting in the 2004 Presidential Election'. *American Political Science Review* 103(1): 59–81.
- Jessee, Stephen. 2016. '(How) Can We Estimate the Ideology of Citizens and Political Elites on the Same Scale?' *American Journal of Political Science* 60(4): 1108–1124.
- Kellerman, Michael. 2012. 'Estimating Ideal Points in the British House of Commons Using Early Day Motions'. *American Journal of Political Science* 56(3): 757–771.
- Kim, In Song, John Londregan and Marc Ratkovic. 2018. 'Estimating Spatial Preferences from Votes and Text'. *Political Analysis* 26(2): 210–229.
- Lauderdale, Benjamin E. and Alexander Herzog. 2016. 'Measuring Political Positions from Legislative Speech'. *Political Analysis* 24(3): 374–394.
- Lauderdale, Benjamin E. and Tom S. Clark. 2014. 'Scaling Politically Meaningful Dimensions Using Texts and Votes'. *American Journal of Political Science* 58(3): 754–771.
- Laver, Michael, Kenneth Benoit and John Garry. 2003. 'Extracting Policy Positions from Political Texts Using Words as Data'. *American Political Science Review* 97(2): 311–331.
- Lewis, Jeffrey B. and Chris Tausanovitch. 2018. *gpuideal: Fast Fully Bayesian Estimation of Ideal Points*. <https://github.com/JeffreyBLewis/gpuideal/> (accessed 24 November 2019).
- Lunn, David J., Andrew Thomas, Nicky Best and David Spiegelhalter. 2000. 'WinBUGS – A Bayesian Modelling Framework: Concepts, Structure, and Extensibility'. *Statistics and Computing* 10: 325–337.
- Lunn, David, David Spiegelhalter, Andrew Thomas and Nicky Best. 2009. 'The BUGS Project: Evolution, Critique and Future Directions'. *Statistics in Medicine* 28(25): 3049–3067.
- Martin, Andrew D. and Kevin M. Quinn. 2002. 'Dynamic Ideal Point Estimation via Markov Chain Monte Carlo for the U.S. Supreme Court, 1953–1999'. *Political Analysis* 10(2):134–153.
- Martin, Andrew D., Kevin M. Quinn and Jong Hee Park. 2011. 'MCMCpack: Markov Chain Monte Carlo in R'. *Journal of Statistical Software* 42(9): 1–21. <https://www.jstatsoft.org/v042/i09> (accessed 24 November 2019).
- Meng, Xiao-Li. 1994. 'Posterior Predictive p-Values'. *The Annals of Statistics* 22(3):1142–1160.
- Minhas, Shahryar, Peter D. Hoff and Michael D. Ward. 2019 'Inferential Approaches for Network Analysis: AMEN for Latent Factor Models'. *Political Analysis* 27(2): 208–222.
- Mislevy, Robert J. 2016. Missing Responses in Item Response Modeling. In *Handbook of Item Response Theory, Volume 2: Statistical Tools*, ed W. J. van der Linden. Boca Raton, FL: CRC Press pp. 171–194.
- Ormerod, J.T. and M.P. Wand. 2010. 'Explaining Variational Approximations'. *The American Statistician* 64(2): 140–153.
- Park, Trevor and George Casella. 2008. 'The Bayesian Lasso'. *Journal of the American Statistical Association* 103(482): 681–686.
- Patz, Richard J. and Brian W. Junker. 1999a. 'Applications and Extensions of MCMC in IRT: Multiple Item Types, Missing Data, and Rated Responses'. *Journal of Educational and Behavioral Statistics* 24(4): 342–366.
- Patz, Richard J. and Brian W. Junker. 1999b. 'A Straightforward Approach to Markov Chain Monte Carlo Methods for Item Response Models'. *Journal of Educational and Behavioral Statistics* 24(2): 146–178.
- Pemstein, Daniel, Kevin M. Quinn and Andrew D. Martin. 2011. 'The Scythe Statistical Library: An Open Source C++ Library for Statistical Computation'. *Journal of Statistical Software* 42(12): 1–26. <https://www.jstatsoft.org/v042/i12> (accessed 24 November 2019).
- Plummer, Martyn. 2003. JAGS: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling. In *Proceedings of the 3<sup>rd</sup> International Workshop on Distributed Statistical Computing*, vol. 124, no. 10. March

- 20–22, 2003, Technische Universität Wien, Vienna, Austria. Eds. Kurt Hornik, Friedrich Leisch and Achim Zeileis. <https://www.r-project.org/conferences/DSC-2003/Proceedings/> (accessed 24 November 2019).
- Poole, Keith T. 1998. 'Recovering a Basic Space from a Set of Issue Scales'. *American Journal of Political Science* 42(3): 954–993.
- Poole, Keith T. 2005. *Spatial Models of Parliamentary Voting*. New York: Cambridge University Press.
- Poole, Keith T. and Howard Rosenthal. 1985. 'A Spatial Model for Legislative Roll Call Analysis'. *American Journal of Political Science* 29(2): 357–384.
- Poole, Keith T. and Howard Rosenthal. 1991. 'Patterns of Congressional Voting'. *American Journal of Political Science* 35(1): 228–278.
- Poole, Keith T. and Howard Rosenthal. 1997. *Congress: A Political-Economic History of Roll Call Voting*. New York: Oxford University Press.
- Poole, Keith, Jeffrey Lewis, James Lo and Royce Carroll. 2011. 'Scaling Roll Call Votes with wnominate in R'. *Journal of Statistical Software* 42(14):1–21.
- Pope, Jeremy C. and Shawn A. Treier. 2011. 'Reconsidering the Great Compromise at the Federal Convention of 1787: Deliberation and Agenda Effects on the Senate and Slavery'. *The American Journal of Political Science*. 55(2): 289–306.
- Pope, Jeremy C. and Shawn A. Treier. 2012. 'Mapping Dimensions of Conflict at the Federal Convention of 1787'. *Legislative Studies Quarterly* 37(2): 145–174.
- Pope, Jeremy C. and Shawn A. Treier. 2015. 'Voting for a Founding: Testing the Effect of Economic Interests at the Federal Convention of 1787'. *Journal of Politics* 77(3): 519–534.
- R Core Team. 2016. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. <https://www.R-project.org/> (accessed 24 November 2019).
- Rivers, Douglas. 2003. Identification of Multidimensional Spatial Voting Models. Unpublished Manuscript.
- Rosas, Guillermo, Yael Shomer and Stephen R. Haptonstahl. 2015. 'No News Is News: Nonignorable Nonresponse in Roll-Call Data Analysis'. *American Journal of Political Science* 59(2): 511–528.
- Sanders, Mitchell S. 1998. 'Unified Models of Turnout and Vote Choice for Two-Candidate and Three-Candidate Elections'. *Political Analysis* 7: 89–115.
- Sanders, Mitchell S. 2001. 'Uncertainty and Turnout'. *Political Analysis* 9(1):45–57.
- Sanderson, Conrad and Ryan Curtin. 2016. 'Armadillo: A Template-Based C++ Library for Linear Algebra'. *Journal of Open Source Software* 1(2): 26.
- Schwarz, Daniel, Denise Traber and Kenneth Benoit. 2017. 'Estimating Intra-Party Preferences: Comparing Speeches to Votes'. *Political Science Research and Methods* 5(2): 379–396.
- Shor, Boris, Christopher Berry and Nolan McCarty. 2010. 'A Bridge to Somewhere: Mapping State and Congressional Ideology on a Cross-institutional Common Space'. *Legislative Studies Quarterly* 35(3): 417–448.
- Shor, Boris and Nolan McCarty. 2011. 'The Ideological Mapping of American Legislatures'. *American Political Science Review* 105(3): 530–551.
- Sinharay, Sandip. 2016. Bayesian Model Fit and Model Comparison. In *Handbook of Item Response Theory, Volume 2: Statistical Tools*, ed W. J. van der Linden. Boca Raton, FL: CRC Press. pp. 379–394.
- Sinharay, Sandip, Matthew S. Johnson and Hal S. Stern. 2006. 'Posterior Predictive Assessment of Item Response Theory Models'. *Applied Psychological Measurement* 30(4):298–321.
- Slapin, Jonathan B. and Sven-Oliver Proksch. 2008. 'A Scaling Model for Estimating Time-Series Party Positions from Texts'. *American Journal of Political Science* 52(3):705–722.
- Spiegelhalter, David J., Nicola G. Best, Bradley P. Carlin and Angelika Van Der Linde. 2002. 'Bayesian Measures of Model Complexity and Fit'. *Royal Statistical Society, Series B, Statistical Methodology* 64(4): 583–639.
- Spirling, Arthur and Iain McLean. 2007. 'UK OC OK? Interpreting Optimal Classification Scores for the U.K. House of Commons'. *Political Analysis* 15(1): 85–96.
- Stan Development Team. 2018. *Stan Modeling Language Users Guide and Reference Manual, Version 2.18.0*. <http://mc-stan.org> (accessed 24 November 2019).



- Treier, Shawn. 2010. 'Where Does the President Stand? Measuring Presidential Ideology'. *Political Analysis* 18(1): 124–136.
- Treier, Shawn. 2011. 'Comparing Ideal Points Across Institutions and Time'. *American Politics Research* 39(5): 804–831.
- Treier, Shawn and Simon Jackman. 2008. 'Democracy as a Latent Variable'. *American Journal of Political Science* 52(1): 201–207.
- van der Linden, Wim J., ed. 2016. *Handbook of Item Response Theory, Volume 2: Statistical Tools*. Boca Raton, FL: CRC Press.
- Voeten, Erik. 2004. 'Resisting the Lonely Superpower: Responses of States in the United Nations to U.S. Dominance'. *Journal of Politics* 66(3): 729–754.
- West, Mike and Jeff Harrison. 1997. *Bayesian Forecasting and Dynamic Models*. 2nd ed. New York: Springer.
- Zucco Jr., Cesar and Benjamin E. Lauderdale. 2011. 'Distinguishing Between Influences on Brazilian Legislative Behavior'. *Legislative Studies Quarterly* 36(3): 363–396.

# Bayesian Model Selection, Model Comparison, and Model Averaging\*

Florian M. Hollenbach and Jacob M. Montgomery

## INTRODUCTION

Applied researchers are often interested in testing competing theories against each other. Most often, the goal is to determine whether and how a limited number of variables are related to a single outcome. The question in these cases is, ‘which of these theoretical accounts is most consistent with the data?’; or, more ambitiously, ‘which of these theories is most consonant with the true data generating process (DGP)?’ Despite the ubiquity and importance of this research task, many scholars are still uncertain as to how to proceed in these situations. The purpose of this chapter is to explain how this analytical objective can be accomplished effectively using Bayesian model comparison, selection, and averaging, while also highlighting the key assumptions and limitations of these methods. This chapter’s overall purpose is to provide readers with a larger set of tools for tackling this task and to discourage the kinds of haphazard (and often incorrect) practices

for comparing theories, often seen in the literature.

There are two interlocking problems in comparing and contrasting alternative theories via standard statistical methods. First, in many cases, the alternative theories are not ‘nested’ in a way that allows them to be tested simultaneously in a single-regression model. When this is true – and it often is – the common practice of placing all of the variables from all of the theories into a single regression is inappropriate and can lead researchers to incorrect conclusions. Yet, there appears to be no widely accepted framework in the methods literature that allows scholars to compare non-nested models for the purposes of theory testing.<sup>1</sup>

At the same time, researchers testing any theory need to ‘control for’ additional covariates in order to rule out potential confounding factors, shrink the standard errors for key coefficients, or improve model fit. They must also choose from among many potential modeling options, including functional

forms, link functions, and more. Yet, in many cases, theory offers limited guidance as to which or how many potential confounders should be included or which exact modeling strategy is most appropriate. This leaves scholars facing the challenge of having to choose among many different models and yet having little guidance as to how to arbitrate between them. In response, researchers often engage in a haphazard search through a large implied model space and report only a handful of results to readers for evaluation. Even worse, scholars may either intentionally or unintentionally try out alternative model specifications only until they find a result that confirms their research hypotheses (see Montgomery and Nyhan, 2010, for additional discussion).

What criteria should scholars use when choosing between competing models or when considering alternative modeling strategies or model configurations? In this chapter, we present a number of tools from Bayesian statistics that allow scholars to approach these challenges in a more principled manner. The Bayesian framework significantly facilitates this task, since the model configuration itself can be viewed as an unknown quantity to which Bayesian reasoning can be applied. We can use the tools of Bayesian statistics to compare the relative evidence in favor of various models to select the ‘best’ model, an approach that can be loosely labeled *model selection*. We can also examine posterior estimates to assess the degree to which a candidate model has adequately captured the true data generating process, which we label *model evaluation*. Finally, we can take a more agnostic approach and incorporate the uncertainty about the appropriate model configurations directly into final estimates, i.e., *model averaging*.

Below, we provide a broad overview of the tools available for model selection, evaluation, and averaging, with a special emphasis on theory testing. First, we briefly discuss ‘traditional’ approaches to model selection via Bayes factors and model-fit comparisons.

This latter category of tools includes approximations of Bayes factors, as well as criteria based on out-of-sample prediction. We then discuss the idea that no ‘correct’ model exists, and therefore researchers should incorporate the uncertainty about model configuration directly into their statistical approach. Specifically, we cover Bayesian mixture models, Bayesian model averaging, and the recently developed Bayesian stacking. Throughout the chapter, although we do provide some details of the mathematics, our focus is on providing a general intuition about these various methods along with their relative strengths and weaknesses. Readers interested in more thorough treatments of this subject are directed to the works cited at the end of the chapter. All code to reproduce the models and model-selection examples that we describe in this chapter will be available online.<sup>2</sup>

## MOTIVATING EXAMPLE: TESTING THEORIES OF CONGRESS

As our working example throughout the chapter, we will rely on Richman (2011): an article that appeared in the *American Political Science Review* that seeks to test competing models of policymaking in the US Congress. Specifically, Richman (2011) tests competing models that make predictions about which status quo policies are likely to be enacted as the composition of the House, Senate, and presidency shift (Brady and Volden, 1998; Krehbiel, 1998; Cox and McCubbins, 2005). Using novel estimates of status quo locations in different policy areas, policy changes in those areas, and the ideal points of pivotal actors, Richman (2011) empirically tests four competing theories.<sup>3</sup>

Richman (2011) estimates predictions about where the status quo in 42 policy areas *should be* according to each theory for the 103rd through to the 110th Congress.

Richman (2011) then estimates a simple regression for the status quo policy in issue area  $i$  at time period  $t$  using the following formula:

$$y_{it} = \beta_1 \text{prediction}_{it} + \beta_2 \text{inflation}_{it} + \epsilon_{it}$$

The main difference across models is, therefore, how the *prediction* variable is calculated. In all cases, the theory is that the  $\beta_1$  coefficient should be equal to 1, although the main criteria is determining whether the coefficient is positively related to the outcome as expected. Richman (2011) generates predictions for the status quo to be located at the position of the median voter of the house (*Model 1*), as predicted by the pivotal-politics theory (*Model 2*), by the party cartel model (Cox and McCubbins, 2005) with only negative agenda control (*Model 3*), and by a hybrid cartel theory that assumes some degree of positive agenda control by party leaders (*Model 4*). The only control variable considered in Richman (2011) is a measure of inflation to reflect the natural change in status quo positions in some policy areas, which results from inflation rates. Richman (2011) calculates two versions of the inflation measure: one for the first and second models and one for the third and fourth models. The two inflation measures differ based on the relevant policy area according to the relevant theory. Richman (2011) then evaluates the different models according to their ability to predict the status quo based on a linear model. Specifically, the models are ranked based on their individual  $R^2$  values.

We replicate the four models presented in Table 3 in Richman (2011) using the *brms* package in R (Bürkner, 2017; 2018).<sup>4</sup> The *brms* package provides users with a large number of pre-specified Bayesian models that are then estimated in Stan using C++ (Carpenter et al., 2017; Stan Development Team, 2017). Stan is a relatively young probabilistic programming language, similar in spirit to WinBugs. In fact, writing model code in Stan is quite similar to doing so in WinBug.

At this point, most common Bayesian models can be fit in Stan, and a fast growing number of R packages provide users with pre-programmed routines for an extensive number of models. Stan allows users to estimate models using fully Bayesian sampling via the Hamiltonian Monte Carlo (HMC) methods or approximating posterior means and uncertainty using variational inference. The HMC approach to Markov chain Monte Carlo (MCMC) methods is particularly attractive because of its high scalability and ability to succeed in highly dimensional spaces.<sup>5</sup>

For each of the four models described above, we estimate a standard Gaussian linear model, where  $y = X\beta + \epsilon$ , and  $\epsilon \sim N(0, \sigma)$ . We specify Gaussian priors with mean zero and a standard deviation of five for the regression coefficients. For the residual standard deviation ( $\sigma$ ), we keep the default half student t prior with three degrees of freedom and scale parameter 10. In addition to the four models presented in Richman (2011), we add two additional models. First, we estimate a model that includes all six possible covariates and a lagged dependent variable. Second, we estimate the model with all covariates and the lag DV but also add random intercepts for each congress and issue area. For the standard deviation of the random effects, we use the same half student t prior as above.

The median estimates and 95% credible intervals for the estimates in all six models are shown in Table 49.1. There are two aspects of these results that are notable. First, Richman (2011) correctly identifies that these competing theories cannot be tested within a single model and does not attempt to do so. The result, however, is that we end up with four non-nested models that must be compared against each other. To arbitrate between them, Richman (2011) makes interpretive claims based on the overall model fit (as assessed by  $R^2$  values). For instance, Richman (2011: 161) states that Model 2 ‘dramatically improves upon the predictions that can be made’ relative to

**Table 49.1 Evaluating theories of Congress: median estimates and 95% credible intervals for Table 3 in Richman (2011) (Models 1–4) and two garbage-can models**

Variable	Model 1	Model 2	Model 3	Model 4	Full model	Full model and RE
lag DV					0.59 (0.38,0.8)	0.39 (0.09,0.71)
Median only	0.5 (0.07,0.93)				0.02 (-0.25,0.29)	-0.13 (-0.61,0.24)
Pivotal politics		0.92 (0.66,1.17)			-0.72 (-1.22,-0.23)	-0.61 (-1.19,-0.08)
Party-cartel open rule			0.89 (0.72,1.06)		1.1 (0.23,1.94)	0.81 (-0.16,1.84)
Party-cartel closed rule				0.82 (0.65,0.99)	-0.4 (-1.1,0.34)	-0.11 (-0.96,0.71)
Inflation (median/pivot)	0.07 (-0.1,0.25)	0.05 (-0.1,0.19)			-0.01 (-0.11,0.09)	0 (-0.1,0.1)
Inflation (party)			0.06 (0.02,0.1)	0.07 (0.03,0.11)	0.07 (0.04,0.11)	0.1 (0.05,0.15)
Sigma	4.52 (3.99,5.17)	3.83 (3.38,4.41)	3.09 (2.74,3.54)	3.15 (2.77,3.62)	2.49 (2.19,2.86)	2.21 (1.85,2.69)
N	117	117	117	117	117	117
Random effects	No	No	No	No	No	Yes

Model 1. Likewise, in comparing Models 3 and 4, he concludes that ‘the differences in fit between the models is modest enough that no definitive conclusion can be drawn’ (Richman, 2011: 161). While the model fits superficially suggest that these conclusions are true, without a formalized approach to non-nested model comparison these competing claims cannot be formally tested. Fit statistics such as  $R^2$  are simply not designed to allow us to say clearly that one model is better than another in a statistical sense; i.e., there is no threshold we can establish for when  $R^2$  values are ‘different enough’ to show that one is statistically superior to another.

Second, as is nearly always the case, the models reported in Richman (2011: 160) are not the only ones that were considered:

I have also analyzed the data using a wide range of assumptions, including ordinary least squares, fixed effects by issue, random effects with and without AR(1) errors, panel heteroskedastic errors

with an AR(1) process, and dynamic GMM models without analytic weights. All analytic approaches produced statistically significant effects in the expected direction (except for the median model), and all produced the same relative ranking.

From this description, it seems likely that these alternative specifications were tried in response to or in anticipation of reviewer questions and serve as robustness checks for the main model. Nonetheless, they do indicate that there are other potential model configurations that were considered and could have been evaluated relative to the reported results, if appropriate criteria were available.

**HOW NOT TO TEST COMPETING THEORIES**

How can we arbitrate among these competing theories? Before answering, we discuss

one approach *not* to take: tossing everything into a single model and examining which coefficients are significant. As Achen (2005) shows, even in trivial models with only two covariates, this approach can fail to appropriately test between theories but may actively mislead researchers by, for example, switching the sign of key coefficients once we have conditioned on competing (and likely correlated) concepts. In short, simply combining variables from competing models in the hope that the results will somehow point to the ‘right’ theory is a deeply flawed approach to science. While under strict conditions it *can* be correct,<sup>6</sup> in a more general setting it is ill advised and conclusions based on this approach should not be trusted.

As an example, recall that Columns 5 and 6 in Table 49.1 show what happens if we simply drop all of the various predictions into a single model. Column 5 is the same standard Gaussian model with all six predictors and a lag DV included, and Column 6 also adds Congress and issue specific random effects. The results in this case are dramatic and revealing about the non-sensical nature of the approach. For instance, all four of the critical variables are hypothesized to be positive and significant. However, in the full model, several have credible intervals that now include zero. The *pivotal-politics* variable actually flips to become negative.<sup>7</sup> The party cartel (with open rule) has a 95% credible interval that excludes zero on Column 5 but includes zero once we include random effects (Column 6). In general, we get a mishmash of results that in some cases are hard to interpret and in others do not correspond with the theory at all.

Of course, the problem here is that these coefficients have no interpretable meaning, since the key explanatory variables are deeply inter-dependent. The coefficient for the *pivotal-politics* variable does not reflect the *independent* effect of this prediction, since a change in one variable almost necessarily implies changed values in the

other variables. After all, the variables are all functions of shared precursors (e.g., the median ideological position of the Senate or the position of president). In a causal setting, we might consider them to be post-treatment (Acharya et al., 2016; Montgomery et al., 2018), but even in the current setting there is no justification for including them all in the same model and no way to interpret the coefficients directly when we do.

## MODEL SELECTION AND BAYES FACTORS

The simplest approach to Bayesian model selection is via the construction of Bayes factors. This approach is attractive due to its simplicity, and perhaps for this same reason was among the earliest proposed Bayesian methods for model selection. For these same reasons, we present this approach first. However, we note up front that Bayes factors have been extensively criticized by some authors, leading to the alternative methods discussed below.

A (seemingly) simple way of evaluating models from a Bayesian perspective is nothing more than applying Bayes’ rule to model probabilities (Gill, 2009). Bayes’ rule is a formalization of basic human intuition as to how we can take evidence to inform our beliefs about the ‘true’ state of the world:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} \quad (1)$$

Here,  $P(B)$  is our prior beliefs about  $B$  before any data is collected.  $P(A|B)$  is the conditional distribution of observing  $A$  given that we have observed  $B$ , which is simply another way of expressing the likelihood of a statistical model. Finally,  $P(A)$  is the marginal probability of observing  $A$ .

Bayes factors attempt to apply this same logic to the problem of using observed data

to inform our beliefs about the probability in favor of a specific model. Let  $\pi(\mathcal{M}_k)$  be the prior probability that model  $k$  is ‘true’, and let  $p(\mathbf{y}|\mathcal{M}_k)$  be the probability of observing the data  $\mathbf{y}$  under the assumption that  $k$  is true.<sup>8</sup> Further, assume that we are considering  $k \in 1, 2, \dots, K$  alternative model configurations. With a finite set of potential models, we can then calculate the marginal distribution of the data as

$$p(\mathbf{y}) = \sum_{k=1}^K \pi(\mathcal{M}_k) p(\mathbf{y} | \mathcal{M}_k). \quad (2)$$

By simply applying Bayes’ rule, we can express the posterior probability of any particular model  $k$  as

$$\begin{aligned} p(\mathcal{M}_k | \mathbf{y}) &= \frac{\pi(\mathcal{M}_k) p(\mathbf{y} | \mathcal{M}_k)}{\sum_{k=1}^K \pi(\mathcal{M}_k) p(\mathbf{y} | \mathcal{M}_k)} \\ &= \frac{\pi(\mathcal{M}_k) p(\mathbf{y} | \mathcal{M}_k)}{p(\mathbf{y})}. \end{aligned} \quad (3)$$

If we then want to compare two models (a vs b), we can construct a ratio between the two models’ posteriors. This has the advantage that the denominators will simply cancel out:

$$\begin{aligned} \frac{p(\mathcal{M}_a | \mathbf{y})}{p(\mathcal{M}_b | \mathbf{y})} &= \frac{\pi(\mathcal{M}_a) p(\mathbf{y} | \mathcal{M}_a)}{\pi(\mathcal{M}_b) p(\mathbf{y} | \mathcal{M}_b)} \\ &= \text{Prior odds}(\mathcal{M}_a; \mathcal{M}_b) \\ &\quad \times \text{Bayes factor}(\mathcal{M}_a; \mathcal{M}_b) \end{aligned} \quad (4)$$

Since the Bayes factor contains all of the ‘objective’ information about the models (i.e., information that is separate from the model priors), this has been the traditional quantity of interest. Higher values for this calculation represent evidence in favor of  $M_a$  and smaller values represent evidence in favor of  $M_b$ . Jeffreys (1961) made an early attempt to set thresholds for Bayes factors, where this evidence could be considered ‘conclusive’ or merely ‘suggestive’. A widely accepted threshold is that a Bayes factor of three is ‘substantial’ evidence in favor of  $M_a$  and values above 10 are considered ‘strong’.<sup>9</sup>

The advantage of the Bayesian approach to model evaluation is that estimating a posterior probability for each model allows us to talk about model selection in an intuitive way. Unlike alternatives such as likelihood-ratio tests that rely on confusing p-values and null hypothesis testing, we can talk directly about model probabilities. Statements like ‘there is a 90% chance that this is the best model’ have some possibility of making sense; i.e., we can directly assess various models and determine which one is most supported by the data and with what degree of certainty. Moreover, the models do not have to be nested to be comparable.

Despite these superficial advantages, however, the simplified description above obscures several complexities that make the problem of model comparison and selection difficult. To begin with, in almost all cases we are not just interested in choosing the right model but rather in estimating some set of model parameters  $\theta$  conditioned on our model choice and data; i.e., our actual learning target is often the posterior distribution,  $p(\theta_k | \mathbf{y}, \mathcal{M}_k)$ . Further, the presentation above makes implicit assumptions about prior structures for  $\theta$  that will not always hold. In more realistic settings, evaluating model fit based on Bayes factors comes with several additional difficulties that have, in combination, worked against their widespread adoption: computational intractability, prior sensitivity, incomplete accounting for uncertainty, and open model spaces.

### **Marginal Likelihoods**

In order to construct Bayes factors, we need to be able to calculate the probability of the data given the model after marginalizing out the model parameters. More concretely, let  $\theta \in \Theta$  be some set of model parameters of interest, let  $L(\theta_k) = p(\mathbf{y} | \theta_k, \mathcal{M}_k)$  represent the standard likelihood function for the data from model  $k$ , and let  $\pi(\theta_k | \mathcal{M}_k)$  be the prior

distribution for  $\theta_k$ .<sup>10</sup> In order to calculate model probabilities or Bayes factors, we need to marginalize out  $\theta_k$ :

$$p(\mathbf{y} | \mathcal{M}_k) = \int_{\Theta} p(\mathbf{y} | \theta_k, \mathcal{M}_k) \pi(\theta_k | \mathcal{M}_k) d\theta. \quad (5)$$

Unfortunately, the integral in Equation 5 is often not analytically tractable. Moreover, even where it can be approximated with fidelity for one model, the set of models being considered may be sufficiently large such that Equation 2 cannot be calculated in a reasonable amount of time. Note that in order to find the marginal distribution in Equation 2, we would need to complete these calculations  $K$  times, and without  $p(\mathbf{y})$ , individual model probabilities cannot be directly calculated.

### Approximations and BIC

Statisticians have developed several methods designed to approximate Equation 3 quickly. For example, one can use Laplace’s method to approximate  $p(\mathbf{y} | \mathcal{M}_k)$  using the following formula (Ando, 2010, 115):

$$p(\mathbf{y} | \mathcal{M}_k) \approx p(\mathbf{y} | \hat{\theta}_k, \mathcal{M}_k) \pi(\hat{\theta}_k | \mathcal{M}_k) \times \frac{(2\pi)^{\frac{p}{2}}}{n^{\frac{p}{2}} |J(\hat{\theta})|^{\frac{1}{2}}}, \quad (6)$$

where  $\hat{\theta}$  is the maximum likelihood estimator (MLE),  $p$  is the number of parameters in the model, and  $J(\theta) = -\frac{1}{n} \frac{\partial^2 \log L(\theta)}{\partial \theta \partial \theta^T}$  is a function of the Hessian of the log-likelihood such that the second term is related to the asymptotic covariance of the MLE.

This can be simplified further in instances where a large  $n$  and appropriate prior structure allow us to ignore the prior. In such cases, Schwarz (1978) proposes a simpler approximation of  $p(\mathbf{y} | \mathcal{M}_k)$ . Disregarding

portions of Equation 6 and that are constant in large- $n$  settings and taking the log, we get the Bayesian information criterion (BIC):<sup>12</sup>

$$\text{BIC} = -2 \log p(\mathbf{y} | \hat{\theta}_k) + p \log n,$$

where we let  $p(\mathbf{y} | \hat{\theta}_k, \mathcal{M}_k) = p(\mathbf{y} | \hat{\theta}_k)$  to simplify the notation. Likewise, we get

$$\log BF[\mathcal{M}_a; \mathcal{M}_b] \approx (\text{BIC}_b - \text{BIC}_a) / 2. \quad (7)$$

The advantage of using BIC is its simplicity. We can evaluate and compare models using straightforward calculations from the likelihood based on the MLE. The obvious drawback is that BIC will give inaccurate and even misleading approximations in small sample settings or when prior structures cannot be ignored (e.g., improper priors). Subsequent work has also raised questions about whether BIC can actually be considered an approximation of any valid quantity (Gelman and Rubin, 1995).<sup>11</sup> Indeed, Berger et al. (2003) shows that even in fairly simple models, BIC can lead to incorrect conclusions, even as  $n \rightarrow \infty$ .<sup>13</sup> For these reasons, we advise that researchers be cautious in interpreting BIC as a proper approximation of a Bayes factor or, even better, avoid the use of BIC altogether.

### Approximation via Simulation and Bridge Sampling

Several other approaches focus on approximating marginal likelihoods, using methods that take advantage of the simulation methods (e.g., MCMC) typically used to estimate posterior distributions of  $\theta_k$ . What these methods have in common is that they attempt to avoid the potentially high-dimensional integration problem in Equation 5 using Monte Carlo-like approximations.

One of the simplest approaches for simulation again rests on the Laplace method. Assume that we have  $s \in [1, 2, \dots, S]$  samples of  $\theta_k$  from model  $k$ , denoted  $\theta_k^{(s)}$ . We can then



approximate the posterior mode as (Ando, 2010: 170)

$$\begin{aligned} \hat{\theta}_k &\approx \max_s \{p(\theta_k^{(s)} | \mathbf{y})\} \\ &= \max_s \{p(\mathbf{y} | \theta_k^{(s)}, \mathcal{M}_k)\pi(\theta_k^{(s)} | \mathcal{M}_k)\}. \end{aligned}$$

The posterior covariance can be approximated as

$$\hat{V} \approx \frac{1}{S} \sum_{s=1}^S \{(\theta_k^{(s)} - \bar{\theta}_k)^T (\theta_k^{(s)} - \bar{\theta}_k)\}$$

where  $\bar{\theta}_k$  is the posterior mean. We can then get

$$p(\mathbf{y} | \mathcal{M}_k) \approx p(\mathbf{y} | \hat{\theta}_k)\pi(\hat{\theta}_k) \times (2\pi)^{\frac{p}{2}} |\hat{V}|.$$

More advanced examples of numerical approximations in the literature include reversible jump MCMC (Green, 1995),

Chib’s (1995) method, path sampling (Gelman and Meng, 1998), the harmonic mean estimator (Gelfand and Dey, 1994), and more. Each of these approaches has its relative advantages and disadvantages, but all can be technically difficult to implement. Further, estimators that rely on evaluations of the likelihood (e.g., the harmonic mean estimator) can be numerically unstable, since these are technically unbounded.

Perhaps the most generally useful approach within this family is bridge sampling (Meng and Wong, 1996; Meng and Schilling, 2002), which has a fairly ‘black box’ implementation available for applied researchers (Gronau et al., 2017a; Gronau et al., 2017b). The basic idea is to introduce a ‘bridge function’,  $h(\theta)$ , and a proposal distribution,  $\psi(\theta)$ . We can then set up the identity

$$\begin{aligned} 1 &= \frac{\int_{\Theta} h(\theta_k) p(\theta_k | \mathbf{y}, \mathcal{M}_k) \psi(\theta_k) d\theta_k}{\int_{\Theta} h(\theta_k) \psi(\theta_k) p(\theta_k | \mathbf{y}, \mathcal{M}_k) d\theta_k} = \frac{\frac{1}{p(\mathbf{y} | \mathcal{M}_k)} \int_{\Theta} h(\theta_k) p(\mathbf{y} | \theta_k, \mathcal{M}_k) \pi(\theta_k | \mathcal{M}_k) \psi(\theta_k) d\theta}{\int_{\Theta} h(\theta_k) \psi(\theta_k) p(\theta_k | \mathbf{y}, \mathcal{M}_k) d\theta_k} \quad (8) \\ p(\mathbf{y} | \mathcal{M}_k) &= \frac{\int_{\Theta} h(\theta_k) p(\mathbf{y} | \theta_k, \mathcal{M}_k) \pi(\theta_k | \mathcal{M}_k) \psi(\theta) d\theta}{\int_{\Theta} h(\theta_k) \psi(\theta_k) p(\theta_k | \mathbf{y}, \mathcal{M}_k) d\theta_k} \end{aligned}$$

For the denominator in Equation 8, we can use draws of  $\theta_k$  to approximate the integral numerically. Likewise, we can take draws from the

proposal density  $\psi(\theta_k)$  to accomplish the same task in the numerator. Assuming we take  $L$  draws from  $\psi(\cdot)$ , we can estimate Equation 5 as

$$p(\mathbf{y} | \mathcal{M}_k) = \frac{\frac{1}{L} \sum_{l=1}^L h(\theta_k^{(l)}) p(\mathbf{y} | \theta_k^{(l)}, \mathcal{M}_k) \pi(\theta_k^{(l)} | \mathcal{M}_k)}{\frac{1}{S} \sum_{s=1}^S h(\theta_k^{(s)}) \psi(\theta_k^{(s)})}.$$

Obviously, in order to implement this model, we must choose  $h(\cdot)$  and  $\psi(\cdot)$ . Meng and Wong (1996) provide an optimal choice for  $h(\cdot)$  in terms of the mean-squared error of the estimator. For efficiency, the proposal density will ideally be as close as possible to the posterior distribution. The bridge-sampling package relies on either a multivariate normal distribution, where the mean vector and covariance matrix are calculated from the full posterior for  $\theta$  and a

‘warped’ posterior (Meng and Schilling, 2002).

Once we have estimated  $p(\mathbf{y} | \mathcal{M}_k)$  for various models, we can then compare specific models by constructing Bayes factors. Once again, when comparing models  $a$  and  $b$ , we can get

$$BF[\mathcal{M}_a, \mathcal{M}_b] = \frac{p(\mathbf{y} | \mathcal{M}_a)}{p(\mathbf{y} | \mathcal{M}_b)}.$$

Alternatively, to maintain comparability with the BIC approach, we can calculate

$$\begin{aligned} \log BF[\mathcal{M}_a, \mathcal{M}_b] \\ = \log \{p(y | \mathcal{M}_a)\} - \log \{p(y | \mathcal{M}_b)\}. \end{aligned}$$

**Prior Sensitivity,  $\mathcal{M}$ -Closed Assumption, and Uncertainty**

Perhaps the greatest limitation of Bayes factors is that the results can be very sensitive to priors. The approach to model selection discussed above rests on the  $\mathcal{M}$ -closed assumption, meaning we are assuming that one of the  $\mathcal{M}_k \in \mathcal{M}$  is the *true* model, even if the researcher does not know which it is. Especially when this is untrue, Bayes factors will be driven by the choice of prior distributions chosen for  $\theta$ . Yao et al. (2018: 919) provide the following powerful example:

[C]onsider a problem where a parameter has been assigned a normal prior distribution with center 0 and scale 10, and where its estimate is likely to be in the range  $(-1, 1)$ . The chosen prior is then essentially flat, as would also be the case if the scale were increased to 100 or 1000. But such a change would divide the posterior probability of the model by roughly a factor of 10 or 100.

In essence, if we were to compare two models that are exactly the same except for one having a prior over  $\beta$  with standard deviation 10 and the other with standard deviation 100, the Bayes factor would strongly favor the first, despite the fact that the posterior estimates of  $\theta$  and even predictions for each observation would be essentially the same across models. Further, placing (most) improper vague priors on  $\theta$  – arguably the most agnostic approach to

model building – will lead to the Bayes factor not existing at all. This leads to the awkward result that the ultimate decision about model quality depends on choices we make about the prior structures on  $\theta$  parameters – choices that may be only incidental to the scientific question at hand.

A final concern is that some of these approaches to model comparisons are not ‘truly Bayesian’, in the sense that they rely on a single estimate of  $\hat{\theta}$  rather than reflecting the entire posterior. Even the approaches that leverage the full posterior over  $\theta$  do so only to marginalize the quantities away.

**Example Application**

Keeping these limitations in mind, we turn back to our running example. Based on the non-sensical results when including all covariates, we from now on only compare the four original models presented by Richman (2011). First, we calculate BIC for each of the four models, as presented in Table 49.2. As we can see, Model 3 has the smallest BIC value, which would indicate the highest model fit. This is in line with the evaluation by Richman (2011) based on the  $R^2$  values. Similarly, as with Richman’s evaluation, the difference between the BIC values of Models 3 and 4 is very small. Nevertheless, the BIC for Model 2 is only 50 points larger than that of Model 3. Thus, again, one would have to conclude that Model 3 seems to be slightly better than the other two models. In the bottom part of Table 49.2, we show the approximate Bayes factor

**Table 49.2 BIC and approximation to Bayes factor**

<i>BIC scores</i>			
<i>Model 1</i>	<i>Model 2</i>	<i>Model 3</i>	<i>Model 4</i>
696.3	657.2	607.2	611.5
Log Bayes factor approximation for Model 3			
$M_3: M_1$	$M_3: M_2$	–	$M_3: M_4$
44.4	25.0		2.2

**Table 49.3 Log marginal likelihood and Bayes factor using bridge sampling**

<i>Log marginal likelihood</i>			
<i>Model 1</i>	<i>Model 2</i>	<i>Model 3</i>	<i>Model 4</i>
-351.0	-332.5	-309.4	-311.5
Log Bayes factor via bridge sampling for Model 3			
$M_3: M_1$	$M_3: M_2$	-	$M_3: M_4$
41.7	23.2		2.2

(on the log scale), calculated as in Equation 7, for Model 3 compared to the three other models. The results lend additional support to the idea that Model 3 is clearly better than Models 1 and 2 while only being a slight improvement when compared to Model 4. On the original scale, the Bayes factor between Model 3 and Model 4 is nine; i.e., in the language discussed above, this could be considered ‘substantial’ but not quite ‘strong’ evidence in favor of Model 3 over Model 4.

Next we use bridge sampling, with a *bridge function* as in Equation 8, to first estimate the log marginal likelihood for the four models. We also generate estimations of the log Bayes factor comparing Model 3 to the others. The *brms* package again makes this very easy for applied users, as the estimations are integrated into the package. The top half of Table 49.3 shows the log marginal likelihood for each of the four models as estimated via bridge sampling. Similar to previous results, the differences are largest with respect to Models 1 and 2 compared to Models 3 and 4. The lower half of Table 49.3 presents the Bayes factor of  $\mathcal{M}_3 : \mathcal{M}_k$  on the log scale. For this exercise, the approximate Bayes factor using BIC and the estimation via bridge sampling are very similar. We note, however, that these different approximations are likely to diverge in more complicated settings.

**PREDICTIVE MODEL EVALUATION**

While model selection via Bayes factors and marginal model probabilities seems intuitive, as the above discussion indicates it is not always so straightforward in practice. This is

particularly true in instances where the model parameters  $\theta$  take on continuous values, requiring both informative priors and marginalization to complete the calculations. In addition to the difficulty of calculating the marginal likelihoods discussed above, Bayes factors in general are sensitive to priors on elements of  $\theta$  in ways that can be undesirable.

As an alternative, the literature contains a number of approaches intended to help scholars evaluate the quality of any given model based on their predictive capacity. Here, we follow the presentation in Gelman et al. (2013) and Gelman et al. (2014). The most common approaches are based on information theory, with the goal of minimizing the Kullback–Leibler (KL) divergence between the *true* (but unknown) data generating function and the predictive distribution implied by model  $\mathcal{M}_k$ .

Let  $\hat{\theta}$  be our current estimate for the model parameters,  $p(\tilde{y} | \hat{\theta})$  be predictions from model  $k$  for some new observation  $\tilde{y}$  that were not used to fit the model, and  $g(\cdot)$  be the *true* data generating function. Note that for the rest of this section we suppress the implied conditioning on the model for clarity. Rather than evaluating model fit based on *in-sample* performance, the idea is that we would like to choose the model that best fits *all* of the data, not just that which we have collected.

One way to summarize the predictive fit of a model is the log predictive density (lpd),  $\log\{p(\tilde{y})\}$ . This quantity has a nice feature in that, in the limit, the model with the lowest KL information also has the highest lpd (Gelman et al., 2014). For a single point, we can define the lpd as the point prediction from that point after marginalizing out  $\theta$ ,

$$\begin{aligned} \log\{p(\tilde{y}_i)\} &= \log E[p(\tilde{y}_i | \theta)] \\ &= \log \int_{\Theta} p(\tilde{y}_i | \theta) p(\theta | \mathbf{y}) d\theta. \end{aligned}$$

Note that the expectation is taken over the posterior of  $\theta$ , where the posterior is estimated based on the in-sample data  $\mathbf{y}$ .

We have to go further here, because the future data ( $\tilde{y}$ ) is itself unknown. However, we can again use Bayesian reasoning to calculate the expected log predictive density (elpd) as:

$$\begin{aligned} E[\log\{p(\tilde{y}_i)\}] \\ = \int g(\tilde{y}_i) \log\{p(\tilde{y}_i)\} d\tilde{y} = \text{elpd}. \end{aligned}$$

In this case, the expectation is taken in terms of the unknown function  $g(\cdot)$ .

For more than one data point, we can simply sum this value to create the expected log pointwise predictive density (elpdp),

$$\text{elpdp} = \sum_{i=1}^n E[\log\{p(\tilde{y}_i)\}]. \tag{9}$$

The model that scores highest on this value can be considered the ‘best’ model in terms of its predictive accuracy and, with large samples, optimal in terms of KL information. However, since this quantity cannot be calculated directly, we must again turn to approximations below.

Before moving onto specific approximations, however, it is helpful to define two additional quantities. First, if we assume some specific point estimate  $\hat{\theta}$ , we can calculate the elpd as  $E[\log\{p(\tilde{y}_i | \hat{\theta})\}]$ . Further, in this case and given standard *iid* assumptions, we can simplify the notation to get

$$p(\tilde{y} | \hat{\theta}) = \prod_{i=1}^n p(\tilde{y}_i | \hat{\theta}).$$

**Information Criteria**

In the particular case where we use the MLE, the elpd can be approximated accurately using the Akaike information criterion (AIC) (Ando, 2010).

$$AIC = -2 \log L(\hat{\theta}_{MLE}) + 2p,$$

where  $\text{elpd} \approx -\frac{1}{2} AIC$ . However, it is important to note that this assumes (i) we have not included informative priors on  $\theta$ , (ii) the posterior distribution for  $\theta$  can be represented as a multivariate normal, and (iii) the model is correct (the true data generating process corresponds to some unknown member of the specified parametric distributions) (Ando, 2010: 199). Thus, while it is simple to calculate, it is probably not applicable in many situations.

Note that AIC has two additive components, which is a feature of all of these similar criteria. The first represents the degree to which the model fits well given the data already collected; i.e., the in-sample fit. The problem, of course, is that models that better explain the data we have will *not* always be a superior representation of the underlying DGP. Instead, more complex models – models that include more variables, interactions, non-linearities, and the like – may simply capture random noise in the dataset, mistaking it for true information. More formally, improved in-sample model fit may actually decrease the ability of the model to explain (or predict) new observations generated by the same process.<sup>14</sup>

Following this logic, the second term in the AIC formula is a penalty term that punishes for complexity to work against selecting models that over-fit the data. Under the assumptions stated above, the penalty term in the AIC is exact. However, when we move beyond a world of flat priors and linear models, this penalty term will no longer be correct (Gelman et al., 2013).

The deviance information criterion (DIC) overcomes these issues by approaching the problem from a more strict Bayesian perspective. Spiegelhalter et al. (2002) proposed the following criteria:

$$DIC = -2 \log\{p(\mathbf{y} | \hat{\theta}_{EAP})\} + P_D$$

where  $P_D$  is a Bayesian measure of model complexity that is based on the posterior

mean  $\hat{\theta}_{EAP}$  and posterior covariance  $\text{Var}(\theta)$ . The exact penalty is

$$P_D = 2 \left( \begin{array}{l} \log \{ p(\mathbf{y} | \hat{\theta}_{EAP}) \} \\ - \int_{\Theta} \log \{ p(\mathbf{y} | \theta) p(\theta | \mathbf{y}) \} d\theta \end{array} \right).$$

Since the second term is simply the expected value for the log predictive distribution where the expectation is taken over the posterior of  $\theta$ , we can use Monte Carlo integration using the  $s = 1, \dots, S$  draws from the posterior,

$$P_{D(1)} \approx 2 \left( \begin{array}{l} \log \{ p(\mathbf{y} | \hat{\theta}_{EAP}) \} \\ - \frac{1}{S} \sum_{s=1}^S \log \{ p(\mathbf{y} | \theta^{(s)}) \} \end{array} \right).$$

An alternative approximation,

$$P_{D(2)} \approx 2 \text{Var}[\log \{ p(\mathbf{y} | \theta) \}],$$

has the advantage of always providing a positive value (Gelman et al., 2013).

**Model Evaluation Based on Pointwise Predictive Densities**

While both AIC and DIC aim to approximate (or proxy for) the out-of-sample fit, they are subject to two criticisms. First, the models use the same training data both to fit the model and evaluate the appropriate complexity penalty. This can bias estimates towards models that are too complex, leading researchers to the wrong decision. Further, both criteria use only a single point estimate (either the MLE or the EAP) to evaluate model fit. This means that we ignore the full posterior of the model parameters, making the methods not ‘fully Bayesian’, which can lead to problems, including negative estimates for the effective number of parameters (Vehtari et al., 2017: 1414).

From a Bayesian perspective, several other criteria are generally preferred for model evaluation. These allow us to make use of the full posterior distribution but also avoid the problems of prior sensitivity that plague Bayes factors discussed above.

These methods are also, to differing degrees, closely related to out-of-sample performance, which is more consonant with recent trends in model evaluations and further protects against over-fitting. For these criteria, we move towards evaluating fit based on how well predictive *densities* approximate the true data generating process for individual data points.

The goal is to try to evaluate each model based on its predictive accuracy, where accuracy is evaluated based on the predictive distribution rather than the point estimate. Let  $\tilde{y}$  be some set of data points we are trying to predict (either new data or a data ‘held out’ during fitting) and  $\mathbf{y}^{obs}$  be the data we are currently using to fit the model. We can then write the posterior predictive distribution as

$$p(\tilde{y}) = \int_{\Theta} p(\tilde{y} | \theta_k) p(\theta | \mathbf{y}^{obs}) d\theta = E[p(\tilde{y} | \theta_k)].$$

For similar reasons as those noted above, we want to evaluate the model based on some function of the logged value,  $\log \{ p(\tilde{y}) \}$ . Using  $s = \{ 1, \dots, S \}$  draws of  $\theta$  from a sampler, we can approximate using Monte Carlo integration. Summing over all observations in the held-out dataset, we then get the ‘computed log pointwise predictive density [clppd]’ (Gelman et al., 2013: 169):

$$\text{clppd} = \sum_{i=1}^n \log \left\{ \frac{1}{S} \sum_{s=1}^S p(\tilde{y}_i | \theta^{(s)}) \right\}.$$

For a fully Bayesian treatment, we would then *like* to calculate the elppd shown in Equation 9. The general problem is that since we cannot marginalize over the unknown function  $g(\cdot)$ , we must again settle for approximations based on clppd. When clppd is calculated within the sample, we will overestimate the elppd, which we then need to adjust (similar to the penalties for AIC and DIC discussed above). Alternatively, we might rely on true out-of-sample forecasts for calculating clppd. However, this comes

with the problem of either introducing error in the clppd as an estimate of elppd (because each  $\theta$  is estimated on a subset of the data and therefore our out-of-sample forecasts may be noisier) or imposing considerable computational requirements.

### WAIC

The widely available information criteria (WAIC, or alternatively the Watanabe information criterion) is intended to provide a computationally friendly way of evaluating the performance of models based on predictive distributions (Watanabe, 2010, 2013). To generate the WAIC, we compute the clppd within the training sample and then apply a penalty term for complexity:

$$WAIC = \frac{1}{n} \sum_{i=1}^n \log \{ p(\tilde{y}_i) \} - P_w.$$

As with DIC, the trick is to find the correct penalty term  $P_w$ . One approach is to let  $P_w = V/n$ , where  $V$  is the functional variance (Piironen and Vehtari, 2017b):

$$V = \sum_{i=1}^n \{ E[p(\tilde{y}_i | \theta)^2] - E[p(\tilde{y}_i | \theta)]^2 \}.$$

Gelman et al. (2013, Equation 7.12), however, recommend an alternative penalty equivalent to the summed variance of the log predictive density of each data point:

$$P_w = \sum_{i=1}^n \text{Var}_s \left[ \log \{ p(y_i | \theta^{(s)}) \} \right],$$

where  $\text{Var}_s[\cdot]$  is the sample variance function, and we calculate the variance across the  $S$  draws from the posterior.<sup>15</sup>

### Cross validation

Rather than trying to arrive at a correct penalty for complexity, another approach is cross-validation to reduce the bias from overfitting.<sup>16</sup> First, the data is split into  $A$  subsets (e.g.,  $A = 10$ ) of approximately equal size  $\left( \frac{A-1}{A} \times n \right)$ . The model is then estimated on

each of the subsets and predictions are made for the observations that were held out. Once completed  $A$  times, we now have out-of-sample predictions for each observation in the dataset. Based on this procedure, we can define a ‘cross validated predictive density’ (Ando, 2010):

$$\text{cvpd}_k = \prod_{\alpha=1}^A \int p(y_\alpha | \theta^{(\alpha)}) p(\theta^{(\alpha)} | y_{-\alpha}).$$

In this case, we ‘hold out’ observations in partition  $\alpha$  and calculate the posterior on  $\theta$ . Based on this distribution, we can then calculate the clppd for model evaluation.<sup>17</sup>

While relying on out-of-sample predictions does reduce the bias in our estimate of elppd, reducing the sample size during model fitting can decrease the accuracy of the overall model itself. Generally,  $A$  between 8 and 16 has been recommended as reasonable to trade off error and computational cost (Vehtari and Lampinen, 2002). Of course, estimating even a single Bayesian model, especially with large  $N$  or many parameters, can be computationally intensive and time consuming. Cross-validation requires repeatedly ( $A$ -times) estimating each of the models that one is interested in comparing. When we have many models to compare, cross-validating each can involve fitting thousands of total models. Since this can be parallelized easily, in simple cases cross-validation may be a good option. But even with modern computing, full cross-validation is only practical when comparing relatively few models and when each of the models is relatively computationally inexpensive. Further, depending on the type of dataset and estimated models (e.g., consider hierarchical data or spatial models), scholars have to give serious consideration as to how to partition the data for analysis (and for some there may be no clean approach to doing so).

### LOO-CV

The extreme case of cross-validation methods is leave-one-out cross-validation

(loo-cv). Here, the model is estimated  $n$  times, each time leaving out one observation in the estimation and predicting that particular left-out observation. As in cross-validation, the predictive densities for all separately left out and predicted observations are then evaluated using some criteria. This approach has been shown to have a number of good properties and, when feasible, is perhaps the best way to evaluate alternative models. Watanabe (2010) shows that WAIC is asymptotically equivalent to the Bayesian leave-one-out cross-validation (loo-cv). Some argue that WAIC and loo-cv 'give a nearly unbiased estimate of the predictive ability of a given model' (Piironen and Vehtari, 2017a: 712). Of course, since all models need to be estimated  $n$  times, loo-cv is likely to be too computationally intensive for many applied researchers.

To make loo-cv computationally tractable, Gelfand et al. (1992) and Gelfand (1996) suggested importance sampling leave-one-out cross-validation. Effectively, we want to avoid estimating each model  $n$  times and sampling a new  $\theta_i$  to create a predictive distribution for each left-out observation  $i$ . To do so, we approximate  $\theta_i$  using the posterior draws for  $\theta$ , taking into account the degree to which data point  $i$  affects the estimate. In the original approach, the importance ratio used to approximate the model with the left-out observation  $i$  would be  $r_i^s = \frac{1}{p(y_i | \theta^s)}$  (Vehtari and Lampinen, 2002; Vehtari et al., 2017). Thus, instead of having to estimate each model  $n$ -times, we only generate one estimate of  $\theta$  based on the whole data and then approximate the posterior for each of the data subsets by re-weighting with the importance ratios.

This strategy, however, can be problematic under several circumstances, such as for data with highly influential cases or high-dimensional models.<sup>18</sup> More recently, Vehtari et al. (2017) have suggested an improvement to importance sampling for loo-cv by

smoothing the importance ratios, such that extreme values are not too influential or problematic. This is done using the Pareto distribution with its heavy tails, i.e., 'Pareto smoothed importance sampling (PSIS)' (Vehtari et al., 2017: 1413).

Specifically, psis-loo-cv smooths the 20% largest, and therefore potentially problematic, importance ratios. To do so, first a generalized Pareto distribution is fit to the largest importance ratios. These potentially problematic ratios are then replaced with 'the expected values of the order statistics of the fitted generalized Pareto distribution' (Vehtari et al., 2017: 1415). Not only should the resulting smoothed weights perform better, the shape parameter of the fitted Pareto distribution can be used to check the reliability of the new importance weights. A large estimated shape parameter of the fitted Pareto distribution can indicate problems with the underlying distribution of the original importance samples. In that case, the estimates from the Pareto-smoothed importance sampling may also be problematic. One immediate advantage is that one can then identify those problematic observations. This allows scholars to then estimate the full leave-one-out posterior for those observations that were identified as problematic. We can then directly sample from this actual leave-one-out posterior  $p(\theta | y_{-i})$  for those observations. The full model evaluation would then be based on both psis-loo and full-loo estimates. Claassen (2019) has recently used psis-loo-cv for model evaluation in political science.

As Vehtari et al. (2017) argue, psis-loo-cv has considerable *computational* advantages over exact leave-one-out cross-validation. It also performs better than WAIC, traditional importance sampling, and truncated importance sampling loo-cv on a variety of models. For hierarchical models with few data points per group and high variation in the parameters between groups, however, the performance of WAIC and psis-loo decreases and exact loo-cv becomes more valuable.<sup>19</sup>

**Table 49.4** Information criteria for evaluating theories of Congress

	<i>Model 1</i>	<i>Model 2</i>	<i>Model 3</i>	<i>Model 4</i>
psis-loo-ic	687.6	650.1	600.7	605.3
WAIC	687.6	650.0	600.7	605.2
k-fold-ic	688.7	650.4	600.1	603.8

### Application

We now return to our example application from above and compare the different models in terms of WAIC, 10-fold cross-validation, and psis-loo. The *loo* package in *R* allows scholars to easily generate these model evaluation scores for models estimated in *Stan* (Vehtari et al., 2019). Moreover, the *loo* package is again integrated into the *brms* package and the information-criteria scores are readily available for our estimated models.

The *loo* package produces the expected log predictive density (elpd) as well as the information criteria on the deviance scale (i.e.,  $-2 \times \text{elpd}$ ) for all three information criteria. In Table 49.4, we present the different information-criteria scores on the deviance scale for the four models. First, in our simple example with 117 observations, two parameters per model, and an estimated Gaussian linear model, the values of the different information criteria for each model are quite similar to each other. For example, for Model 3, the WAIC score is 600.65, the psis-loo-ic score is 600.74, and the k-fold-ic score is 604. The same is true for the other three models, where the WAIC, psis-loo-ic, and k-fold-ic closely correspond with each other. In line with the results of the previous information criteria presented, Model 3 performs best, i.e., has the lowest information criteria scores.

Readers might wonder how to decide whether the evidence in favor of a particular model is strong enough to make a claim about it being the *best* model. As mentioned above, for the Bayes factor, a value of 10 or larger would, according to some, be

considered *strong* evidence in favor of the better model. As a first step in comparing two models in terms of the information criteria, we can calculate the difference in their scores. If we want to evaluate whether Model 3 should be strictly preferred to Model 1, we calculate  $\text{psis-loo-ic}_1 - \text{psis-loo-ic}_3 = 86.9$ . In Table 49.5, we present the psis-loo-ic scores of the individual models in the top four rows, but in the bottom part of the table, we present the calculated difference in psis-loo-ic scores between the four different models. Recall that positive values indicate a better score for the second model in the difference. As we can see, the second model is preferred for all comparisons except when comparing Models 3 and 4. The difference in psis-loo-ic scores again indicates that Model 3 performs better than Model 4. But is this difference large enough to strictly prefer Model 3?

Fortunately, the model-evaluation criteria discussed here allow us to calculate uncertainty estimates, which can be used to judge whether differences between models are large enough to draw conclusions. Each of these methods is estimated using functions applied to the individual observations in the data to create the information criteria; i.e., the total information criteria are combinations of  $n$  scores. Using the standard deviation of those  $n$  components, one can estimate an approximate standard error for the information criteria. For example, in the case of psis-loo-cv, we can estimate a standard error based on the standard deviation of the  $n$  individual components of the elppd:  $\widehat{\text{elpd}}_{i,loo}$  (Vehtari et al., 2017: 1426). Similarly, the individual components can be used to calculate an approximate standard error for the difference in information criteria scores.



**Table 49.5** Loo information criteria for evaluating theories of Congress with standard errors

Model	<i>psis-loo-ic</i>	<i>SE</i>
Model 1	687.6	17.5
Model 2	650.1	24.0
Model 3	600.7	24.1
Model 4	605.3	24.7
Differences and SEs		
$M_1 - M_2$	37.6	12.2
$M_1 - M_3$	86.9	16.4
$M_1 - M_4$	82.3	16.8
$M_2 - M_3$	49.3	9.5
$M_2 - M_4$	44.7	11.0
$M_3 - M_4$	-4.6	4.4

The top half of Table 49.5 shows the estimated  $\widehat{elpd}_{loo}$  score on the deviance scale for each of the four models as well as the estimated standard error. In the bottom half of the table, we present the difference in *psis-loo-ic* scores and the standard error for each difference. There is no clear and hard rule about how large the difference compared to its standard error would have to be to conclude that a model is strictly superior. One might take the general rule of thumb that we would like to see a difference that is *at least* twice the size of the standard error. It has been suggested, however, that these standard error estimates are an optimistic approximation and, especially for smaller sample sizes, might not be appropriate (Vehtari et al., 2017). When model differences are sufficiently small, scholars may prefer the less complex model that provides comparable fit. Again, both differences in information criteria and their standard errors are easily available in the *loo* package and integrated into the *brms* package.

Based on the results presented in Table 49.5, we would conclude that Model 3 is significantly better than Models 1 and 2. On the other hand, comparing Model 3 and Model 4 suggests that both do a similar job at explaining the variation in the data; i.e., there is no strong reason to prefer one to the other.

## FINITE MIXTURE MODELS, BAYESIAN MODEL AVERAGING, AND STACKING

In this section, we turn to a somewhat different approach for handling multiple potential models. In particular, we consider statistical approaches where each of the competing models is considered a component of an overarching model; i.e., we eschew the task of selecting or even comparing models and instead consider how much each *component* model contributes to the model combination. In the end, we may heuristically prefer the component that contributes the most, making this distinction seem somewhat arbitrary. However, from a statistical standpoint, the approaches we cover next can be seen as quite distinct from those discussed above.

Specifically, we return to the finite model-space approach discussed in the earlier section on model selection and Bayes factors.<sup>20</sup> We can then engage in model selection either by choosing the model that receives the most weight in this mixture or skipping the task of model selection entirely and attempting to make inferences that actually reflect our uncertainty about the true DGP.

### Mixture Models

One family of models attempts to assign each observation to one of the potential candidate models and estimate  $\theta_k$  based on this assignment (McLachlan and Peel, 2000; Imai and Tingley, 2012). Let  $p(y_i | \theta_k, M_k)$  represent the predictive distribution of observations  $i$  from model  $k$  and let  $\tau = [\tau_1, \tau_2, \dots, \tau_k]$  index which model actually generated each observation such that  $\tau_i \in [1, 2, \dots, K] \forall i \in [1, 2, \dots, n]$ . We can construct our mixture model as

$$p(y_i | \tau, \theta_k) \sim \sum_{k=1}^K p(y | \theta_k) \mathcal{J}(\tau_i = k),$$

where  $\mathcal{J}(\cdot)$  is the standard indicator function. We can then complete the model by placing

appropriate priors over  $\theta_k$  and a hierarchical prior structure on  $\tau$ ,

$$\pi(\tau) \sim \text{Multinomial}(\omega)$$

$$\pi(\omega) \sim \text{Dirichlet}(\alpha).$$

In this case, the  $\omega_i$  parameter represents the probability a generic observation is assigned to each model, and we can interpret the posterior estimate in a fashion similar to (but distinct from) the model probabilities shown in Equation 3. Likewise, we can look at the posterior distributions on the  $\tau$  vector to get an estimate of how many observations are assigned to each model.

One important factor to note is that the model parameters for each component  $\theta_k$  are estimated for observations ‘assigned’ to that component using standard Bayesian methods; i.e., we are imagining that all of the models are operating at the same time but that different units belong to each (Imai and Tingley, 2012).

In many cases, this can be desirable, although it can increase uncertainty for components that are assigned few observations. Scholars must also be careful in how they interpret individual parameters, as the parameters do not correspond to estimates for the entire population. Further, simultaneous estimation of model weights and model parameters can also lead to model degeneracy and identification problems during estimation. Standard regression models can be estimated in a fully Bayesian fashion using the `BayesMix` package in R or with alternative estimation routines (via the EM algorithm) using the `FlexMix` package. However, with attention paid to issues of identification, they can also be fit using `brms`.

### EBMA

Ensemble Bayesian model averaging (EBMA), largely deriving from the literature on forecasting, is different in that we fit component models separately using the entire

dataset and then combine them into a weighted ensemble (Raftery et al., 2005; Montgomery et al., 2012). In this case, we can divide our dataset into three partitions: a training set used to fit each individual model ( $\mathbf{y}^{\text{train}}$ ); a calibration used to determine the model weights ( $\mathbf{y}^{\text{cal}}$ ); and a true test set that we are hoping to accurately predict ( $\mathbf{y}^{\text{test}}$ ). We assume that each of the component models is fit to the training data (although the component models need not be statistical models at all). The calibration set represents observations that were predicted out-of-sample by each component model and allows us to appropriately weight them without having to develop penalties for complexity. The goal is then to combine the forecasts in order to make accurate predictions of the test observations.

Let  $w_k = p(\mathcal{M}_k | \mathbf{y}^{\text{cal}})$ , and  $p(\mathbf{y}^{\text{test}} | \mathcal{M}_k)$  represent the predictive pdf for the test set from model  $k$ . Our goal is then to generate an ensemble prediction,

$$p(\mathbf{y}^{\text{test}}) = \sum_{k=1}^K w_k p(\mathbf{y}^{\text{test}} | \mathcal{M}_k).$$

To complete the model, therefore, we need to estimate the model weights for each component. Formally, we need to find the values of  $\mathbf{w}$  that will maximize the log-likelihood function,

$$\mathcal{L}(\mathbf{w}, \Theta) = \sum_{i=1}^{n^{\text{cal}}} \log \left\{ \sum_{k=1}^K w_k p(\mathbf{y}^{\text{cal}} | \theta_k) \right\}$$

subject to the constraint that  $\sum \omega_k = 1$ , which can be calculated efficiently using an EM algorithm or gibbs sampler in the `EBMAforecast` package in R.

### BMA

As the name suggests, EBMA is closely related to Bayesian model averaging (Madigan and Raftery, 1994; Raftery, 1995; Bartels, 1997; Gill, 2004; Montgomery and Nyhan, 2010; Cranmer et al., 2017; Plümper

and Traummüller, 2018). In the traditional approach to BMA,<sup>21</sup> each model is again fit to the entire dataset. Model weights are calculated according to Equation 3.

Of course, this again leaves us with the problem of needing to estimate marginal likelihoods. One option is to use one of the several proxies discussed above, such as AIC or BIC. Another option is to select priors that allow for closed form solutions. For instance, Zellner’s (1986) *g*-prior is

$$\pi(\beta | \sigma^2) \sim \text{Normal}(\mathbf{0}, g\sigma^2(\mathbf{X}'\mathbf{X})^{-1}),$$

where  $\mathbf{X}$  is some set of covariates and  $\sigma^2$  is the variance for the residuals in a standard regression model.<sup>22</sup> This yields a marginal model likelihood of

$$p(\mathbf{y} | \mathbf{X}, \mathcal{M}_k) = \frac{\Gamma(n/2)}{\pi^{n/2}} (1 + g)^{-p} S^{-1},$$

where  $S$  is a function only of the data and the prior mean for  $\beta$ .<sup>23</sup>

The result is that for any quantity of interest, we can simply construct a weighted ensemble from the full posterior. For instance, to estimate the posterior distribution of a specific regression coefficient, we need only calculate

$$\begin{aligned} p(\beta | \mathbf{y}) &= \sum_{k=1}^K p(\mathcal{M}_k | \mathbf{y}) p(\beta_k | \mathbf{y}, \mathcal{M}_k) \\ &= \sum_{k=1}^K w_k p(\beta_k | \mathbf{y}, \mathcal{M}_k). \end{aligned}$$

For models where this coefficient is excluded, we will then have a point mass at zero. The rest will create an ensemble posterior that reflects our uncertainty based on the set of covariates. It also allows us to focus on two separate quantities of interest that are commonly confused in interpreting regression analysis. First, we might be interested in the posterior probability that some particular variable should be in the model. This will be the sum of the model weights where that variable is included. Second, we might be interested in the distribution of  $\beta$  conditioned on the fact that it is included in the model,  $p(\beta | \mathbf{y}, \beta \neq 0)$ .

### Stacking

BMA is based on the marginal likelihood of each model under an  $\mathcal{M}$ -closed assumption. Thus, there are several problems with the BMA approach that largely correspond to the issues with Bayes factors discussed above. First, model weights can be sensitive to prior specifications. Second, it is not always clear how to place priors on the probability of specific models, since seemingly innocuous assumptions can also affect weights in unintended ways. For instance, placing an agnostic prior that each coefficient has a 50% prior probability for inclusion biases the ensemble towards models that include 50% of the covariates. Finally, an underlying assumption of BMA is that the true model is included within the model space (i.e., the  $\mathcal{M}$ -closed case). BMA will place all of its posterior weight on this one model asymptotically.

Estimating model weights in stacking is done in a different two-step process. First, the candidate models are estimated based on the data available. Second, model weights are calculated for each of the estimated candidate models. In the original stacking, model weights are generated by minimizing the leave-one-out mean-squared error for each observation based on each model’s point prediction.

Yao et al. (2018) built on several recent developments to further develop stacking to use the full leave-one-out predictive distribution instead of the point estimates. Adjusting the notation in Yao et al. (2018), let

$$\hat{p}_{ki}(\tilde{y}_i) = \int_{\Theta} p(\tilde{y}_i | \theta_k, \mathcal{M}_k) p(\theta_k | \mathbf{y}, \mathcal{M}_k) d\theta.$$

Applying a logarithmic scoring rule,<sup>24</sup> we can then find the stacking weights for each model by solving the optimization problem:

$$\max_w \frac{1}{n} \sum_{i=1}^n \log \sum_{k=1}^K w_k \hat{p}_{ki}(\tilde{y}_i).$$

If we adopt the psis-loo-ic approximations above, we can simplify this further to give us

weights based on estimates of the elpd, providing stacking weights that can be constructed from a single posterior but which reflect each models' out-of-sample properties.<sup>25</sup>

Stacking has the advantage in that it does not assume  $\mathcal{M}$ -closed, but rather allows that the true DGP may not be well represented by any of the candidate models. Additionally, because weights are calculated based on the unit-specific elpd for each model, stacking takes into account when models have different strengths in predicting certain observations. One of the advantages of stacking is, therefore, that it is able to combine weights from models that are very similar to the better model, instead of splitting the weight between very similar models, as often occurs in BMA.<sup>26</sup>

### Application

As a last example based on our application, we generate stacking model weights based on the *psis-loo-ic* scores presented above. Stacking can be easily done using the *loo* package (Vehtari et al., 2019) and is integrated into *brms* (Bürkner, 2017). Once the *psis-loo-ic* scores are calculated, stacking weights are readily available. Additionally, a warning would be given if the Pareto-smoothed importance sampling is questionable for some observations and full leave-one-out resampling is then suggested for those observations.

As we can see in Table 49.6, stacking weights are highly concentrated on Model 3, even though – based on the information criteria – Model 3 and 4 performed quite similarly (see Table 49.5 above). This suggests that while BMA may have split the weight between Models 3 and 4 due to their similar performance, stacking combines the weights to the slightly better model. While Model 3 may be only slightly better than Model 4, the two models seem to be quite similar in the cases they predicted well; thus, adding

**Table 49.6 Model weights based on stacking**

<i>Model 1</i>	<i>Model 2</i>	<i>Model 3</i>	<i>Model 4</i>
0.04	0.00	0.96	0.00

weight to Model 4 beyond Model 3 must not improve predictive accuracy. The remaining weight (0.04) is assigned to Model 1. In this case, this model is weighted poorly because it is not as accurate as the others. However, it is still weighted more highly than Models 2 and 4 because it gives sufficiently different predictions.

### CONCLUSION

The current state of the literature when it comes to testing competing models is sometimes unsatisfactory. In many cases, the conclusions that researchers draw from their analysis are sensitive to modeling choices. Traditional null-hypothesis-testing strategies in particular lead to analysts focusing excessively or even exclusively on whether or not key variables are ‘significant’ in the models they consider. However, significance can change easily depending on parametric assumptions, the set of covariates included, and more. In total, in many cases, competing theories are not meaningfully tested against each other, and the conclusions we draw from our data are driven by modeling choices that are irrelevant or orthogonal to the scientific question at hand. Even worse, the a common practice we observe in the literature – simply tossing all variables from all theories into a single regression – is known to produce incorrect and misleading conclusions. In all, we are left with a picture that for a fundamental task for science – arbitrating between competing theories – standard practices are pushing researchers towards either building incorrect garbage-can models or haphazardly exploring the potential space of models with little guidance other than important p-values.

Throughout this chapter, we have provided researchers with a set of alternative strategies for arbitrating between theories and models. The advantage of the Bayesian framework in tackling this problem is that we can use the laws of probability to think clearly about model probabilities directly. There are no null hypotheses required and no misleading p-values. Instead, we can talk meaningfully about whether a given model is better either in terms of Bayes factors or KL divergence (an optimal criteria implied by Bayesian decision theory). While imperfect, we believe that this set of tools offers a superior approach to arbitrating among theories than do standard practices in the discipline.

With that said, it is important to keep the limitations of these methods in mind. In general, the appropriateness of these model-evaluation techniques depends on the specific model specification and settings in which they are used. Further, the use of information criteria or predictive accuracy for model selection, for example, should not be a substitute for theoretical considerations of which covariates are important to include, given the question asked. Similarly, these model-selection techniques are not able to distinguish between pre- and post-treatment variables. In fact, full garbage-can models may perform better on these scores, even though the associated results are not theoretically meaningful. Thus, as with all methodological tools, they should be used with the necessary understanding of the actual problem being considered and theories of how the data was generated.

Finally, given recent advances in this area in the field of statistics, we hope that this overview will renew attention in the field of political science to these questions. Surely, none of the methods above represent the last word on this topic. Nor does our presentation take into account many practical difficulties applied scholars face in practice such as clustering, spatial correlations, time-series, confounding requiring an identification strategy, and measurement error. Further research is

needed to evaluate the usefulness of these various methods in such circumstances.

## Notes

- \* Authors listed in alphabetical order. We thank Richard A. Nielsen, Bruce A. Desmarais, and Andrew Gelman for helpful comments on earlier drafts. All remaining errors are our own.
- 1 See Clarke (2001) for a review of some of the many methods in this area that have been proposed. Clarke (2007) provides further improvements to the Vuong test that allows for comparison of two non-nested models. Desmarais and Harden (2013) note that the Clarke test can be based on biased estimates and be inconsistent, but they suggest an alternative implementation using cross-validation. Our claim is not that there are no methods available, only that many applied scholars appear to be very uncertain about which, if any, method to use.
- 2 Code and data can be found on GitHub at <https://github.com/fhollenbach/BayesModelSelection>.
- 3 Our specific focus here is on Table 3 in the original paper. We deviate from Richman (2011) in not estimating panel corrected standard errors, as this sort of ‘correction’ does not easily translate into a Bayesian framework, and we want to present a very simple example. The *brms* package we use does allow users to estimate models with autoregressive terms or other solutions to serial correlation in time and space.
- 4 To have the same sample in all our models, we delete three observations that have missing data on the lagged dependent variable.
- 5 The intricacies behind Hamiltonian Monte Carlo go beyond the scope of this chapter; for a more thorough introduction to Hamiltonian Monte Carlo, see Neal (2011) and Betancourt (2017).
- 6 Imagine, for instance, that we had conducted a large experiment with 12 different treatment arms. In this case, simply tossing all of the variables into a single regression would indeed be an appropriate way to test the effectiveness of each treatment. However, social scientists are rarely, if ever, blessed with explanatory variables where the effects are strictly linear and co-linearity is sufficiently low to allow for this approach to work.
- 7 This is not a function of the lagged dependent variable being included. Without the lagged dependent variable but with all other covariates included, the estimate of the pivotal-politics coefficient is  $-1.37$  and the 95% credible interval ranges from  $-1.87$  to  $-0.86$ .

- 8 The assumption that the *true* model is in the model space is not necessary for methods of model selection discussed later.
- 9 If these two models have equal prior distributions, the models are nested, and we estimate them via maximum likelihood, then Equation 4 simply reduces to a likelihood ratio statistic.
- 10 The  $k$  subscript on  $\theta$  allows that each model under consideration may have a different set of parameters.
- 11 Despite the name, notice that BIC is primarily useful for evaluating the model fit from the MLE in an asymptotic setting where priors are irrelevant. Further, the motivation and derivation for BIC differs significantly from the other ‘information criteria’ covered below.
- 12 Note that some texts and software packages may define BIC as the negative of how we have defined it.
- 13 More advanced approaches for estimating marginal model probabilities include the generalized Bayesian information criterion (Konishi et al., 2004), which allows for more informative priors.
- 14 See Hastie et al. (2016) for a fuller discussion of the problems of in-sample and out-of-sample model fit. See Kung (2014) and Bagozzi and Marchetti (2017) for recent applications of AIC for model selection.
- 15 See Kim et al. (2018) for a recent application of WAIC in political science.
- 16 In most literatures, this approach is referred to as ‘k-fold’ cross-validation. However, to avoid notational confusion with the model space, we do not adopt this language here.
- 17 Note that each competing model should be fit based on the same partitioning of the data.
- 18 In particular, the variance of the importance weights may be too large or infinite, since the denominator is not bounded away from zero.
- 19 Similarly, for models with spatial or temporal dependence, psis-loo-cv is likely to be problematic. Ongoing work is considering possible approaches for these cases.
- 20 For reasons of space, we confine ourselves here to finite mixture models with a known number of potential components. However, we note that there is also a large literature on Bayesian models that can relax or eliminate this assumption.
- 21 For large model spaces, several scholars have developed stochastic samplers that simultaneously estimate the model probabilities and model parameters (George and McCulloch, 1993).
- 22 We also place an improper prior on  $\sigma^2$  of  $\frac{1}{\sigma^2}$ .
- 23 See chapter 5 in Ando (2010) for a complete proof.
- 24 We suppress discussion of alternative scoring rules for simplicity.

25 This last variant is referred to as pseudo-BMA.

26 Interested readers may want to consult the original article, introducing stacking via psis-loo (Yao et al., 2018) and the full discussion appended to the article.

## REFERENCES

- Acharya, Avidit, Matthew Blackwell and Maya Sen. 2016. ‘Explaining causal findings without bias: Detecting and assessing direct effects’. *American Political Science Review* 110(3):512–529.
- Achen, Christopher H. 2005. ‘Let’s put Garbage-can Regressions and Garbage-can Probabilities where they Belong’. *Conflict Management and Peace Science* 22(4):327–339.
- Ando, Tomohiro. 2010. *Bayesian Model Selection and Statistical Modeling*. New York: CRC Press.
- Bagozzi, Benjamin E. and Kathleen Marchetti. 2017. ‘Distinguishing Occasional Abstention from Routine Indifference in Models of Vote Choice’. *Political Science Research and Methods* 5(2):277–294.
- Bartels, Larry M. 1997. ‘Specification, Uncertainty, and Model Averaging’. *American Journal of Political Science* 41(2):641–674.
- Berger, James O., Jayanta K. Ghosh and Nitai Mukhopadhyay. 2003. ‘Approximations to the Bayes Factor in Model Selection Problems and Consistency Issues’. *Journal of Statistical Planning and Inference* 112:241–258.
- Betancourt, Michael. 2017. ‘A Conceptual Introduction to Hamiltonian Monte Carlo’. *CoRR*. <https://arxiv.org/pdf/1701.02434.pdf> (Accessed March 2019).
- Brady, David W and Craig Volden. 1998. *Revolving Gridlock: Politics and Policy from Carter to Clinton*. Boulder, CO: Westview Press.
- Bürkner, Paul-Christian. 2017. ‘brms: An R Package for Bayesian Multilevel Models using Stan’. *Journal of Statistical Software* 80(1):1–28.
- Bürkner, Paul-Christian. 2018. ‘Advanced Bayesian Multilevel Modeling With The R Package BRMS’. *The R Journal* 10(1):395–411.
- Carpenter, Bob, Andrew Gelman, Matthew Hoffman, Daniel Lee, Ben Goodrich, Michael

- Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li and Allen Riddell. 2017. 'Stan: A Probabilistic Programming Language'. *Journal of Statistical Software, Articles* 76(1):1–32.
- Chib, Siddhartha. 1995. 'Marginal Likelihood from the Gibbs Output'. *Journal of the American Statistical Association* 90(432):1313–1321.
- Claassen, Christopher. 2019. 'Estimating Smooth Country–Year Panels of Public Opinion'. *Political Analysis* 27(1): 1–20.
- Clarke, Kevin A. 2001. 'Testing Nonnested Models of International Relations: Reevaluating Realism'. *American Journal of Political Science* 45(3):724–744.
- Clarke, Kevin A. 2007. 'A Simple Distribution-Free Test for Nonnested Model Selection'. *Political Analysis* 15(3):347–363.
- Cox, Gary W. and Matthew D. McCubbins. 2005. *Setting the Agenda: Responsible Party Government in the US House of Representatives*. New York: Cambridge University Press.
- Cranmer, Skyler J, Philip Leifeld, Scott D McClurg and Meredith Rolfe. 2017. 'Navigating the Range of Statistical Tools for Inferential Network Analysis'. *American Journal of Political Science* 61(1):237–251.
- Desmarais, Bruce A. and Jeffrey J. Harden. 2013. 'An Unbiased Model Comparison Test Using Cross-Validation'. *Quality & Quantity* 48(4):2155–2173.
- Gelfand, Alan E. and Dipak K. Dey. 1994. 'Bayesian Model Choice: Asymptotics and Exact Calculations'. *Journal of the Royal Statistical Society. Series B (Methodological)* 56:510–514.
- Gelfand, Alan E. 1996. Model Determination Using Sampling-Based Methods. In *Markov Chain Monte Carlo In Practice*, eds. Walter R. Gilks, Sylvia Richardson and David J. Spiegelhalter. Boca Raton, FL: Chapman & Hall, chapter 9, pp. 145–162.
- Gelfand, Alan E., Dipak K. Dey and Hong Chang. 1992. Model Determination Using Predictive Distributions With Implementation Via Sampling-Based Methods. In *Bayesian Statistics*, eds. José M. Bernardo, James O. Berger, Dawid A. Philip and Adrian F.M. Smith, vol. 4 Oxford, UK: Clarendon Press, pp. 147–167.
- Gelman, Andrew and Donald B. Rubin. 1995. 'Avoiding Model Selection in Bayesian Social Research'. *Sociological Methodology* 25:165–173.
- Gelman, Andrew, Hal S. Stern, John B. Carlin, David B. Dunson, Aki Vehtari and Donald B. Rubin. 2013. *Bayesian Data Analysis*. Boca Raton, FL: Chapman and Hall/CRC.
- Gelman, Andrew, Jessica Hwang and Aki Vehtari. 2014. 'Understanding Predictive Information Criteria for Bayesian Models'. *Statistics and Computing* 24(6):997–1016.
- Gelman, Andrew and Xiao-Li Meng. 1998. 'Simulating Normalizing Constants: From Importance Sampling to Bridge Sampling to Path Sampling'. *Statistical Science* 13(2): 163–185.
- George, Edward I. and Robert E. McCulloch. 1993. 'Variable Selection via Gibbs Sampling'. *Journal of the American Statistical Association* 88(423):881–889.
- Gill, Jeff. 2004. 'Introduction to the Special Issue'. *Politi* 12(4):647–674.
- Gill, Jeff. 2009. *Bayesian methods: A Social and Behavioral Sciences Approach*. Boca Raton, FL: CRC Press.
- Green, Peter J. 1995. 'Reversible Jump Markov chain Monte Carlo Computation and Bayesian Model Determination'. *Biometrika* 82(4):711–732.
- Gronau, Quentin F., Alexandra Sarafoglou, Dora Matzke, Alexander Ly, Udo Boehm, Maarten Marsman, David S. Leslie, Jonathan J. Forster, Eric-Jan Wagenmakers and Helen Steingroever. 2017a. 'A Tutorial on Bridge Sampling'. *Journal of Mathematical Psychology* 81:80–97.
- Gronau, Quentin F., Henrik Singmann and Eric-Jan Wagenmakers. 2017b. 'Bridgesampling: An r Package for Estimating Normalizing Constants'. *arXiv preprint, arXiv:1710.08162*.
- Hastie, Trevor, Robert Tibshirani and Jerome Friedman. 2016. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Second ed. New York: Springer.
- Imai, Kosuke and Dustin H. Tingley. 2012. 'A Statistical Method for Empirical Testing of Competing Theories'. *American Journal of Political Science* 56(1):218–236.
- Jeffreys, Harold. 1961. *Theory of Probability*. Oxford: Oxford University Press.
- Kim, In Song, John Londregan and Marc Ratkovic. 2018. 'Estimating Spatial Preferences from Votes and Text'. *Political Analysis* 26(2):210–229.

- Konishi, Sadanori, Tomohiro Ando and Seiya Imoto. 2004. 'Bayesian Information Criteria and Smoothing Parameter Selection in Radial Basis Function Networks'. *Biometrika* 91:27–43.
- Krehbiel, Keith. 1998. *Pivotal Politics: A Theory of U.S. Lawmaking*. Chicago, IL: University of Chicago Press.
- Kung, James Kai-sing. 2014. 'The Emperor Strikes Back: Political Status, Career Incentives and Grain Procurement during China's Great Leap Famine'. *Political Science Research and Methods* 2(2):179–211.
- Madigan, David and Adrian E. Raftery. 1994. 'Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window'. *Journal of the American Statistical Association* 89(428):1535–1546.
- McLachlan, Geoffrey and David Peel. 2000. *Finite Mixture Models*. New York: John Wiley & Sons, Ltd.
- Meng, Xiao-Li and Stephen Schilling. 2002. 'Warp Bridge Sampling'. *Journal of Computational and Graphical Statistics* 11(3):552–586.
- Meng, Xiao-Li and Wing Hung Wong. 1996. 'Simulating Ratios of Normalizing Constants via a Simple Identity: A Theoretical Exploration'. *Statistica Sinica* 6(4): 831–860.
- Montgomery, Jacob M. and Brendan Nyhan. 2010. 'Bayesian Model Averaging: Theoretical Developments and Practical Applications'. *Political Analysis* 18(2):245–270.
- Montgomery, Jacob M., Brendan Nyhan and Michelle Torres. 2018. 'How Conditioning on Posttreatment Variables can Ruin your Experiment and What to do About it'. *American Journal of Political Science* 62(3):760–775.
- Montgomery, Jacob M., Florian M. Hollenbach and Michael D. Ward. 2012. 'Improving Predictions Using Ensemble Bayesian Model Averaging'. *Political Analysis* 20(3): 271–291.
- Neal, Rashford M. 2011. MCMC Using Hamiltonian Dynamics. In *Handbook of Markov chain Monte Carlo*, eds. Steve Brooks, Andrew Gelman, Galin L. Jones and Xio-Li Meng, Boca Raton, FL: CRC Press, chapter 5, pp. 113–162.
- Piironen, Juho and Aki Vehtari. 2017a. 'Comparison of Bayesian Predictive Methods for Model Selection'. *Statistics and Computing* 27(3):711–735.
- Piironen, Juho and Aki Vehtari. 2017b. On the Hyperprior Choice for the Global Shrinkage Parameter in the Horseshoe Prior. In *Proceedings of the 20<sup>th</sup> International Conference on Artificial Intelligence and Statistics (AISTATS)*, eds. Aarti Singh and Jerry Zhu, Fort Lauderdale, FL: Proceedings of Machine Learning Research 57: 905–913.
- Plümper, Thomas and Richard Traunmüller. 2018. 'The Sensitivity of Sensitivity Analysis'. *Political Science Research and Methods*. First View.
- Raftery, Adrian E. 1995. 'Bayesian Model Selection in Social Research'. *Sociological Methodology* 25(HUH):111–163.
- Raftery, Adrian E., Tillman Gneiting, Fadoua Balabdaoui and Michael Polakowski. 2005. 'Using Bayesian Model Averaging to Calibrate Forecast Ensembles'. *Monthly Weather Review* 133(5):1155–1174.
- Richman, Jesse. 2011. 'Parties, Pivots, and Policy: The Status Quo Test'. *American Political Science Review* 105(1):151–65.
- Schwarz, G. 1978. 'Estimating the Dimension of a Model'. *Annals of Statistics* 6(2): 461–464.
- Spiegelhalter, David J., Nicola G. Best, Bradley P. Carlin and Angelika Van Der Linde. 2002. 'Bayesian Measures of Model Complexity and Fit'. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64(4):583–639.
- Stan Development Team. 2017. 'Stan Modeling Language User's Guide and Reference Manual: Stan Version 2.17.0'. <https://github.com/stan-dev/stan/releases/download/v2.17.0/stan-reference-2.17.0.pdf>. (Accessed September 15, 2017).
- Vehtari, Aki, Andrew Gelman and Jonah Gabry. 2017. 'Practical Bayesian Model Evaluation Using Leave-One-Out Cross-Validation and WAIC'. *Statistics and Computing* 27(5): 1413–1432.
- Vehtari, Aki, Jonah Gabry, Yuling Yao and Andrew Gelman. 2019. *loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models*. R package version 2.1.0. <https://CRAN.R-project.org/package=loo> (Accessed March 9, 2019).
- Vehtari, Aki and Jouko Lampinen. 2002. 'Bayesian Model Assessment and Comparison Using Cross-Validation Predictive



- Densities'. *Neural Computation* 14(10): 2439–2468.
- Watanabe, Sumio. 2010. 'Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory'. *Journal of Machine Learning Research* 11(Dec):3571–3594.
- Watanabe, Sumio. 2013. 'A Widely Applicable Bayesian Information Criterion'. *Journal of Machine Learning Research* 14(Mar): 867–897.
- Yao, Yuling, Aki Vehtari, Daniel Simpson, Andrew Gelman. 2018. 'Using Stacking to Average Bayesian Predictive Distributions'. *Bayesian Analysis* 13(3):917–1007.
- Zellner, Arnold. 1986. 'On Assessing Prior Distributions and Bayesian Regression Analysis with g-prior Distributions'. In *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, eds. Prem K. Goel and Arnold Zellner. Amsterdam: Elsevier, pp. 233–243.

# Bayesian Modeling and Inference: A Postmodern Perspective

Jeff Gill and Simon Heuberger<sup>1</sup>

## INTRODUCTION

Bayesian methods and Bayesian inference are now ubiquitous in the social sciences, including political science and international relations. The reasons for this are numerous and include: a superior way to describe uncertainty, freedom from the deeply flawed Null Hypothesis Significance Testing paradigm, the ability to include previous information, more direct description of model features, and, most recently, statistical computing tools that make model computations easy. Yet this is not a static area in social science methodology, and new methodological developments are published at a rapid pace. The objective of this *Handbook* chapter is to describe basic Bayesian methods, chronicle the recent history of their application in political science and international relations, and then highlight important recent developments.

The key tenets of Bayesian inference are that all unknown quantities are assigned probability statements, these probability

statements are updated when new information (data) are observed, and the results are described distributionally. This does not sound like a radical departure from traditional statistical thinking, but it constitutes a different way of thinking about uncertainty that is strictly probability based, eschewing assumptions like an unending stream of independent and identically distributed (IID) data. Furthermore, this paradigm comes along with some historical and practical conventions:

- Overt and clear model assumptions.
- A rigorous way to make *probability* statements about the real quantities of theoretical interest.
- An ability to update these statements (i.e. learn) as new information is received.
- Systematic incorporation of *qualitative* knowledge on the subject of interest.
- Recognition that population quantities are changing over time rather than fixed immemorial.
- Straightforward assessment of both model quality and sensitivity to assumptions.
- Freedom from the flawed null hypothesis significance testing (NHST) paradigm.

These come mostly from a different intellectual path than more conventionally accepted procedures, including relentless criticism during a roughly 100-year period by giants of the field. For the purpose of this exposition we divide the history of Bayesian statistics, as just described, into three eras: classical, modern, and postmodern. Each of these is discussed with attention to issues relevant to social science research. Our objective is to provide a chapter that is both introductory and a summary of recent research trends.

## CLASSICAL BAYES

Bayes' Law was published in 1763, but the classical era of Bayesian statistical inference occupied the middle of the 20th century up until 1990. Much has been written about this time (Gelman and Shalizi, 2013, Strevens, 2006) but the defining hallmark was a sense of principled philosophical correctness combined with analytical frustration. That is, it was not particularly difficult to write logical and mathematical arguments that demonstrated the inferential, and general scientific, superiority of Bayesian inference over classical methods for real problems, but it was often hard to execute this process for real and pragmatic data problems. Practical work often took incremental and modest directions based on what was mathematically possible. So, one sees publications during this time that give 'recipes' for broad classes of models that one could adapt to their specific question and therefore get results. The key problem here was the common creation of Bayesian models where the final joint specification of the model could not be mathematically decomposed into constituent parts to create a standard regression table of results.

Our starting point is much earlier. Bayes' (1763) famous essay was published two years after his death. The paper's release was facilitated by his friend Richard Price rather than Bayes himself (both are now buried in Bunhill Cemetery in central London). A reverend and amateur mathematician, Bayes developed a probability theorem that relates the order of conditional probabilities. It turns out that Laplace (1774, 1781) was developing similar ideas on the continent but communication was not without problem at the time. Let  $A$  and  $B$  be two non-independent events. Basic probability laws tell us that the conditional probability of  $A$  given  $B$  is given by  $p(A|B) = \frac{p(A,B)}{p(B)}$ , with  $p(A, B)$  being the probability of  $A$  and  $B$  both occurring and  $p(B)$  the unconditional probability that the event  $B$  occurs. We thus obtain the probability of  $A$  conditioned on  $B$  happening from  $p(A|B)$ . Conditional probability can also be defined in the reversed order, with  $A$  happening before considering  $B$ , that is,  $p(B|A) = \frac{p(B,A)}{p(A)}$ .

The joint probability remains the same in both directions, meaning that  $p(A, B) = p(B, A)$ . Since  $p(A, B) = p(A|B)p(B)$ , we can thus surmise that  $p(A|B)p(B) = p(B|A)p(A)$  and by rearranging  $p(A|B) = \frac{p(A)}{p(B)} p(B|A)$ .

The statistical application of Bayes' Law begins with the definition of the likelihood function (Fisher, 1925a, Birnbaum, 1962), which is nothing more than the joint probability of the observed data. Define the collected data  $X$  that is observed once and therefore treated as a fixed quantity. We assume that we know the respective probability mass/density function for these data to accurately describe the data-generating process:  $p(X|\beta)$ , conditioned on some unknown parameter  $\beta$  to be estimated. This joint function is then defined by

$$L(\beta|X) = \prod_{i=1}^n p(X_i|\beta) = p(X_1|\beta) \times p(X_2|\beta) \times \dots \times p(X_n|\beta). \quad (1)$$

The switching of the order of conditionality in this statement is not an application of Bayes' Law, but is Fisher's (1925a) recognition that once the data are observed, they are fixed for inferential purposes, and it is the unknown parameter of interest that is the focus of inference at this point. This perspective differs dramatically from canonical Frequentist inference whereby there is an unending stream of independent and identically distributed data available to the researcher (Gill, 1999).

Fisher (1925b) and others provide the standard and convenient inferential treatment of the likelihood function, which is to find the multi dimension mode of this – usually log-concave to the  $x$ -axis – function and to treat curvature around this mode as a measure of uncertainty. These procedures are incredibly well-established and well-accepted and need not be reviewed here. See Gill and Torres (2019) for a recent detailed exposition.

The next required quantity for Bayesian inference is the *prior distribution*: the presumed distribution of the parameters of interest before the current data for analysis are collected and analyzed:  $p(\beta)$ . Notice that this is not conditioned on anything explicitly in its formulation. In truth it is conditioned on previous knowledge about the effect of interest that is symbolized by the parameter  $\beta$  in the regression context. So if the effect of interest is size of nations' militaries and the outcome of interest is the proclivity to wage militarized conflict, then we would typically notate the magnitude (regression coefficient) of the former as  $\beta$  and the vector outcomes as  $Y$ . The likelihood function,  $L(\beta|X)$ , is central to Bayesian methods because it enables the researcher to find a most probable value  $\hat{\beta}$  by testing differing values of  $\beta$ . The value  $\hat{\beta}$  is the most likely to have generated the data given  $H_0$  expressed through a specific parametric form relative to other possible values in the sample space of  $\beta$ . In this regard, the likelihood function resembles the inverse probability,  $p(H_0|X)$ , but as we will see, this clearly is not the socially desired

value that readers sometimes quest for but is not provided under this framework: the probability of the null hypothesis given some data. Crucially, however, the produced likelihood function from this setup now matters only in relation to other likelihood functions with differing values of  $\beta$  since it is no longer bounded by 0 and 1.

Bayesian inference combines data information in the likelihood function with researchers' prior information. This is done by conditioning the prior on the likelihood using Bayes' Law,

$$\pi(\beta|X) = \frac{p(\beta)}{p(X)} p(X|\beta), \quad (2)$$

which 'flips' the order of the conditional distribution by using the ratio of the prior to the unconditional distribution of the data. Now recall the Bayesian maxim that once a quantity is observed, it is treated as fixed. Here the data are assumed to be observed once and fixed in perpetuity for that given collection. Therefore, the probability of observing the observed data is simply one:  $p(X) = 1$ . This means that the denominator in (2) can be treated as simply a normalizing constant to ensure that  $\pi(\beta|X)$  sums or integrates to one and is therefore a normalized distribution. Since we can always recover this normalizing information later in the inferential process, it is customary and convenient to express (2) with proportionality:

$$\pi(\beta|X) \propto p(\beta)L(\beta|X). \quad (3)$$

Here we have also used the more intuitive form of the likelihood function, as described above. So it is common to say that the posterior is just proportional to the prior times the likelihood. So Bayesian inference is codified by a compromise between prior information and data information.

One source of discomfort for Bayesians during this era was the specification of prior distributions for unknown parameters. This is primarily because non-Bayesian scholars of the era criticized prior specification as

‘subjective’ since it came from sources outside of  $X$ , as determined by the researcher. There were many aspects of this view that were misguided. First, all modeling decisions, in all statistical settings, are subjective. It is subjective what data to use. It is subjective what likelihood function to specify (e.g. link function). It is subjective which estimation procedure to use. It is subjective what software to use. It is subjective how to present results, and so on. In addition, prior specifications are opportunities to include qualitative information, known earlier research on a question of interest, values to be updated over time, and more. Also, prior distributions can be constructed to have many different mathematical properties. Despite all of these facts, the specification of priors remains as a pointedly discussed issue well into the 21st century. We will return to this topic at several points over the course of this chapter.

The difficult classical period of Bayesian statistics produced different strains of prior forms for different purposes. Sometimes these were in competition and sometimes they overlapped. Importantly, the bulk of the 20th century was a period when Bayesians needed to be very circumspect in selecting and describing prior selection since non-Bayesians routinely used this as leverage for criticism. Zellner (1971: 18) divides prior distributions during this time into two types: *data-based* (DB) types that are ‘generated in a reasonably scientific manner and which are available for further analysis’, and *non-data-based* (NDB) types that ‘arise from introspection, casual observation, or theoretical considerations’. The critical problem faced by Bayesians of the classical era is that ‘one person’s NDB prior information can differ from that of another’s’ (Zellner, 1971: 19). However, this issue is not limited to NDB priors: ‘It is possible that two investigators working with the same model and DB prior information can arrive at different posterior beliefs if they base prior information on different bodies of past data’ (Zellner, 1971: 19). Thus, something as simple as a difference of

opinion on relevant prior information divided some Bayesians and left the door open for the classical ‘subjective’ criticism in prior selection.

In truth there was a defined set of principled prior approaches that emerged from this challenging era. Proper Bayes is the group most accurately described by Zellner above: those that constructed prior distributions from compiled evidence, such as earlier studies or published work, researcher intuition, or substantive experts. This is a rich literature that seeks to be build on existing scientific knowledge but emboldens the subjective criticism. Empirical Bayes uses the data not only in the likelihood function but also to estimate these hyperpriors values (parameters of prior distributions). This can be done with or without a specific distributional form at this level (parametric versus nonparametric empirical Bayes, see Morris, 1983). This approach offends some Bayesians because the data are used in both right-hand-side elements of (Equation 3). Lindley (in Copas, 1969: 421) accordingly observed that ‘there is no one less Bayesian than an empirical Bayesian’. Reference Bayes seeks priors that move the prior as little as possible from the likelihood function (Bernardo, 1979) – minimizing the distance between the chosen likelihood and the resulting posterior according to a criteria like the Kullback–Leibler distance, which comparatively measures distributions (this is also referred to as dominant likelihood prior). Note that the term ‘reference prior’ sometimes confusingly also refers to a prior that is used as a default (Box and Tiao, 1973: 23). Related to this approach, but different in purpose, are priors that seek to minimize any sense of subjective information: diffuse or uninformative priors such as parametric forms with large variance or uniform specification (bounded and unbounded). These were often specified because it reduced arguments with non-Bayesians and sometimes led to easily calculated results.

Continuous unbounded uniform priors were referred to as ‘improper’ since they did

not integrate to one over the real line, yet they generally yielded proper posteriors due to a cancellation in Bayes' Law. Consider an unbounded uniform prior for a regression parameter defined by  $p(\beta) = hf(\beta)$  over  $[-\infty:\infty]$ . This is essentially a rectangle of height  $hf(\beta)$  and width  $k = \infty$ . So, re-expressing Bayes' Law with this prior gives

$$\pi(\beta | X) = \frac{p(\beta)p(X | \beta)}{\int_{-\infty}^{\infty} p(\beta)p(X | \beta)d\beta}, \quad (4)$$

which is Equation (2) with the replacement

$$\begin{aligned} & \int_{-\infty}^{\infty} p(\beta)p(X | \beta)d\beta \\ &= \int_{-\infty}^{\infty} p(\beta, X)d\beta = p(X) \end{aligned} \quad (5)$$

by the definition of conditional probability and the process of integration for  $\beta$ , which is hypothetical to us in practical terms but exists in nature. Therefore, replacing the definition of the improper rectangular prior gives

$$\begin{aligned} \pi(\beta | X) &= \frac{(hf(\beta) \times k)p(X | \beta)}{\int_{-\infty}^{\infty} (hf(\beta) \times k)p(X | \beta)d\beta} \\ &= \frac{k}{k} \frac{hf(\beta)p(X | \beta)}{\int_{-\infty}^{\infty} hf(\beta)p(X | \beta)d\beta} \\ &\propto f(\beta)p(X | \beta), \end{aligned} \quad (6)$$

where the cancellation of  $k / k$  occurs because these are the exact same flavor of infinity, otherwise it is undefined. Thus, the posterior is proportional to the likelihood times some finite prior (the  $h$  values can also be canceled or left off due to proportionality). A similar prior for a variance component is  $p(\sigma) = 1/\sigma$  over  $[0: \infty]$ . Unfortunately, as mathematically tractable as improper priors are, they did not overwhelmingly convince Bayesian skeptics, some of whom considered it a form of arithmetic trickery.

This classification of classical priors presents the most common forms but is by no means exhaustive. Others include maximum entropy priors, mixtures priors, decision-theoretic priors, conjugate priors, and more (O'Hagan, 1994). Conjugate priors can be informed or diffuse and provide an especially nice choice since, for a given likelihood function, the parametric form of the prior flows down to the posterior giving simple calculations. In decision-theoretic Bayesian inference (Gill, 2014), results are presented in a full decision-theoretic framework where utility functions determine decision losses and risk, which are minimized according to different probabilistic criteria. These approaches were especially appealing in an age with limited computational tools.

### MODERN BAYES

An important hallmark of the classical era of Bayesian statistics was the relative ease with which a joint posterior distribution that could not be integrated into the constituent marginal distributions could be produced with actual social science data and a realistic model based on theoretical principles. Consider the following example that motivated the work in Gill and Waterman (2004) using the data collected by Mackenzie and Light (1987) on every US federal political appointee to full-time positions requiring Senate confirmation from November 1964 through to December 1984 (the collectors of the data needed to preserve anonymity so they embargoed some variables and randomly sampled 1,500 cases down to 512). This survey queries various aspects of the Senate confirmation process, acclamation to running an agency or program, and relationships with other functions of government. A key question is why these executives last only about two years on average after assuming the position, given the difficulty of the process. This work hypothesizes that running a federal agency (or sub agency) is

considerably more stressful than alternative positions for these executives. The outcome variable of interest is stress as a surrogate measure for self-perceived effectiveness and job satisfaction, measured as a five-point scale from ‘not stressful at all’ to ‘very stressful’. Explanatory variables specified for the model are: Government Experience, Ideology, Committee Relationship, Career.Exec-Compet, Career.Exec-Liaison/Bur, Career.Exec-Liaison/Cong, Career.Exec-Day2day, Career.Exec-Diff, Confirmation Preparation, Hours/Week, and President Orientation (Gill and Casella, 2009, for detailed descriptions). A Bayesian random effects specification for ordered survey outcomes is specified, so latent variable thresholds for  $Y$  are assumed to be on the ordering:

$$U_i : \theta_0 \underset{c=1}{\iff} \theta_1 \underset{c=2}{\iff} \theta_2 \underset{c=3}{\iff} \theta_3 \dots \theta_{C-1} \underset{c=C}{\iff} \theta_C.$$

The vector of (unseen) utilities across individuals in the sample,  $U$ , is determined by a linear additive specification of  $K$  explanatory variables:  $U = -X'\gamma + \epsilon$ , where  $\gamma = [\gamma_1, \gamma_2, \dots, \gamma_p]$  does not depend on the  $\theta_j$ , and  $\epsilon \sim F_\epsilon$ . This means that the probability that individual  $i$  in the sample is observed to be in category  $r$  or lower is

$$\begin{aligned} P(Y_i \leq r | X_i) &= P(U_i \leq \theta_r) \\ &= P(\epsilon \leq \theta_r + X_i'\gamma) \quad (7) \\ &= F_{\epsilon_i}(\theta_r + X_i'\gamma), \end{aligned}$$

which is differently signed than in R: ‘logit  $P(Y_i = k - x) = \text{zeta}_k - \text{eta}$ ’ from the help page (there is no fixed convention on the sign). Specifying a logistic distributional assumption on the errors and adding the random effect term produces this logistic cumulative specification for the whole sample:

$$\begin{aligned} F_\epsilon(\theta_r + X'\gamma) &= P(Y \leq r | X) \\ &= [1 + \exp(-\theta_r - X'\gamma)]^{-1}. \quad (8) \end{aligned}$$

Prior distributions are given to be either semi-informed or skeptical:

$$p(\gamma_k) \sim \mathcal{N}(\mu_{\gamma_k}, \sigma_{\gamma}^2), k = 1, \dots, k$$

for each of the  $K$  explanatory variables,

$$p(\theta_j) \sim \mathcal{N}(0, \sigma_{\theta}^2), j = 1, \dots, C - 1$$

for the four latent variable thresholds, with assigned hyperparameter values  $\mu_{\gamma_k}, \sigma_{\gamma}^2, \sigma_{\theta}^2$

All this produces a joint posterior distribution according to

$$\begin{aligned} \pi(\gamma, \theta | X, Y) &\propto L(\gamma, \theta | X, Y) p(\theta) p(\gamma) \\ &\propto \prod_{i=1}^n \prod_{j=1}^{C-1} \prod_{k=1}^p \left[ \Lambda(\theta_j - X_i'\gamma + b_i) - \Lambda(\theta_{j-1} - X_i'\gamma) \right]^{z_{ij}} \\ &\quad \times \exp \left( -\frac{(\gamma_k - \mu_{\gamma_k})^2}{2\sigma_{\gamma}^2} - \frac{\theta_j^2}{2\sigma_{\theta}^2} \right), \quad (9) \end{aligned}$$

which is kind of ugly (and hard to marginalize). While this form tells us everything we need to know about the *joint distribution* of the model parameters, we need to marginalize (integrate) it for every one of these parameters to create an informative regression table. That is, this is a joint probability statement like  $p(A, B, C)$  for arbitrary example random variables  $A, B$ , and  $C$ , and we need to create the marginal (solitary) probability statements for each. Continuing the contrived example for one of these, say  $A$ , we get  $P(A) = \iint_{CB} p(A, B, C) dBdC$  from elementary integral calculus. Using  $P(A)$ , we can get the mean and standard error to present as the first and second column of a regression table for  $A$  in the conventional way. The key point is that it was easy in the middle of the 20th century to produce a model such as this whereby it was prohibitively difficult or impossible to integrate-out each parameter for a series of marginal posterior distributions to describe inferentially. This led Evans

(1994) to retrospectively describe Bayesians of the time as ‘unmarried marriage guidance councilors’ because they could tell others how to do inference when they often could not do it themselves.

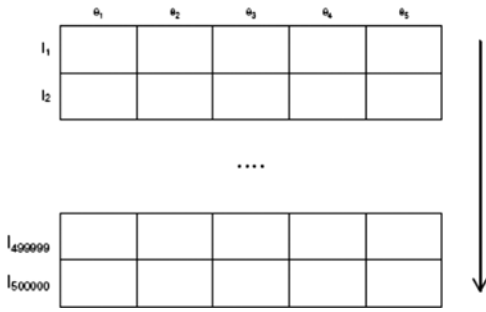
This world, and *the* world, changed in 1990 with a review essay by Gelfand and Smith in the *Journal of the American Statistical Association*. In what is without a doubt one of the ten most important articles published in statistics, they found that a tool, *Gibbs sampling*, hiding in engineering and image restoration (Geman and Geman, 1984), solved this very problem for the Bayesians. Gibbs sampling replaced analytical derivation of marginals from a joint distribution with work from the computer. In this algorithm, the full conditional distributions for each parameter to be estimated are expressed with the conditionality on the other parts of the model that are relevant for this parameter. Then the sampler draws consecutively at each step with the latest drawn versions of the conditioned parameters until the Markov chain reaches its stationary (converged) distribution and these conditional draws can then be treated as marginal samples. This means that the marginal distributions of interest from complicated model specifications can simply be described empirically from a large number of draws sampled by the computer. The date of 1990 is not just important for the publication of this article, but also because, by 1990, statisticians (and others) were free from centralized mainframe computing (with attendant punch cards, JCL, batch processing, and other encumbrances) via ubiquitous and reasonably powerful desktop machines. The popularity of bootstrapping in the 1980s presaged this development for the same computational reason.

The general name for the new tool is *Markov Chain Monte Carlo* (MCMC), which includes Gibbs sampling as well as the *Metropolis-Hastings algorithm* that was dormant in statistical physics but has existed since the 1953 paper by Metropolis et al. in

the *Journal of Chemical Physics* (as was the slightly sexist, somewhat quaint, custom of the time in some of the natural science fields, where their wives typed the paper and were added as coauthors). The principle behind MCMC is that a Markov chain can be setup to describe a high dimension posterior distribution by wandering around the state space visiting sub-regions in proportion to the density that needs to be summarized. The linkage between ‘MC’ and ‘MC’ is the *ergodic theorem* that says that if the Markov chain is created according to some specific technical criteria and is run long enough such that it converges to its stationary (limiting) distribution, the draws from the path of the chain can be treated as if they are IID from the respective marginal distributions. Thus, in the stationary distribution, each step is recorded as a multidimensional draw as if from a regular Monte Carlo process, even though these draws are actually produced from a serial, conditional Markovian algorithm. Robert and Casella (2004), Gill (2014), and Gamerman and Lopes (2006) provide the necessary theoretical background, and Robert and Casella (2011) give an entertaining history of this development in statistics during the early 1990s.

A convenient way to visualize how the MCMC process works practically is given in Figure 50.1, which shows the general data structure as produced by running the Markov chain for a model with five parameters to be estimated  $[\theta_1, \dots, \theta_5]$  over  $m = 1, \dots, M$  MCMC simulations after convergence is asserted. The arrow on the right indicates the direction of the production of simulated vectors. The serial Markovian process produces one row at a time that is conditional on previous rows. Thus, row  $I_1$  is followed by row  $I_2$  on up to row  $I_{499999}$  and row  $I_{500000}$  (the second half of the full chain run of 1 million iterations) as the chain explores the six-dimensional space (five parameters plus posterior density). Each row is a sample, and the ergodic theorem states that under specific circumstances (met in practice with standard





**Figure 50.1** How Markov Chain Monte Carlo works

algorithms) in the converged state, each of these rows can be treated like an independent multidimensional draw from the posterior of interest produced by the Bayesian regression model. Inference is now produced by analyzing down columns for each parameter. From the ergodic theorem we can take the column vector, for say  $\theta_1$ , and summaries for this parameter are performed by simply applying means, variances, quantiles, etc. on the draws down the first column in this case. Thus the ‘MC’ (Markov chain) that produces the rows is later ignored and the second ‘MC’ (Monte Carlo) is used to get inferences.

To say that MCMC revolutionized Bayesian inference and general Bayesian work is to say that commercial air flights slightly improved global travel. It was in fact the tool that freed the Bayesians; no longer were models too complex for the marginalization of joint posteriors to create regression tables and other intuitive summaries. The result was a flood of Bayesian applications that were pent-up in researchers’ file drawers giving a Bayesian perspective to a wide swath of model types and general statistical approaches. What got relatively left behind in this revolution was introspection about the choice of priors. Actually, prior choice was sublimated down to diffuse normals, uniforms, and other vague forms since the focus had turned to issues of estimation: the 1990s saw an explosion of customized MCMC procedures, calculations of quantities of

interest from posterior samples (Chib, 1995; Chib and Jeliazkov, 2001), and a deeper understanding of Markov chain theoretical principles (Gilks et al., 1996; Polson, 1996; Roberts and Tweedie, 1996; Brooks et al., 1997; Roberts and Rosenthal, 1998; Brooks and Roberts, 1999; Robert and Mengersen, 1999; Neal, 2003). In the social sciences, applied Bayesian researchers defaulted almost exclusively to these diffuse forms and thus became *Bayesians of convenience* in the theoretical sense since prior informative was routinely ignored.

Returning to the example of political executives in the US government, Gill and Waterman (2004) used a Gibbs sampler (implemented in the BUGS language) to marginalize the model described in (Equation 9), running the chain for 1,000,000 iterations and disposing of the first 500,000 to feel assured that the chain was exploring its stationary distribution during the second period (a standard set of empirical and graphical diagnostics were used as well). The results are summarized in Table 50.1.

Notice in Table 50.1 that the results are given in terms of posterior quantiles only. For a traditional view of these results, readers can look at the posterior median (0.5) and the 95% credible interval ([0.025:0.975]) for each parameter estimated, as shown in the lighter grey columns, but many Bayesians prefer the more detailed descriptive view of the more posterior distributions given here, which incorporates all columns along with associated general probability statements. So, for instance, we can say that with President Orientation there is a 97.5% probability that this effect is above 0.2462, even if a 98% credible interval ([0.01:0.99]) covers zero. Bayesians tend not to be fixated with arbitrary thresholds, however. It is not directly in the table but from the sorted MCMC values for this parameter, we can see that there is only a 1.2% probability that this effect is negative. Increased levels of Committee Relationship are reliably associated with increased stress from these

**Table 50.1** Posterior quantiles, ordered model for survey of political executives

	0.01	0.025	0.25	0.5	0.75	0.975	0.99
<b>Explanatory variables</b>							
Government Experience	-2.0644	-1.8422	-1.0640	-0.6558	-0.2481	0.5332	0.7552
Ideology	-1.2583	-1.1360	-0.7136	-0.4917	-0.2696	0.1544	0.2761
Committee Relationship	0.2345	0.3784	0.8809	1.1446	1.4089	1.9128	2.0546
Career.Exec-Compet	-0.2450	0.0668	1.1570	1.7304	2.3053	3.3996	3.7096
Career.Exec-Liaison/Bur	-4.1108	-3.9079	-3.1929	-2.8188	-2.4435	-1.7309	-1.5265
Career.Exec-Liaison/Cong	-0.1072	0.0708	0.7036	1.0362	1.3680	1.9991	2.1798
Career.Exec-Day2day	-1.6484	-1.5090	-1.0223	-0.7660	-0.5096	-0.0200	0.1182
Career.Exec-Diff	-0.4171	-0.3076	0.0770	0.2780	0.4791	0.8625	0.9725
Confirmation Preparation	-0.0389	0.1277	0.7154	1.0242	1.3333	1.9223	2.0903
Hours/Week	-1.7215	-1.5653	-1.0156	-0.7273	-0.4390	0.1095	0.2660
President Orientation	-0.0712	0.2462	1.3650	1.9504	2.5355	3.6539	3.9720
<b>Threshold intercepts</b>							
None-Little	-10.2633	-9.5795	-7.1895	-5.9355	-4.6826	-2.2947	-1.6125
Little-Some	-6.3966	-5.8141	-3.7648	-2.6912	-1.6194	0.4197	0.9985
Some-Significant	-2.3605	-1.8037	0.1625	1.1935	2.2229	4.1847	4.7451
Significant-Extreme	3.5151	4.2837	6.9931	8.4080	9.8269	12.5227	13.2905

results since all of the observed posterior quantiles are positive for this coefficient. This is a somewhat paradoxical finding but may be attributed to a closer relationship with the relevant oversight committee that is not voluntary but mandated by congressional concern over agency policies. We see another reliable but subtle positive relationship between *Career.Exec-Compet* and stress since the 95% credible interval is bounded away from zero. This may be because underlings challenge political executives' decisions or because, as career public servants, they know the agency and its mission or history better. There are other statistically reliable findings here as well. These statements are very Bayesian in that we are describing regions of the marginal posterior space for this variable in strictly probabilistic terms. There is no notion of confidence or p-values required. In fact, when people misinterpret standard statistical inference, it is often the Bayesian interpretation, such as we have done here, that is desired.

The modern Bayes era was typified with the ease of production of standard models

with standard assumptions that were beyond the abilities of Bayesians in the classical era. One after another, complex regression style models – including item response theory (IRT), multinomial probit, complex hierarchical forms, causal specifications, and more – were simply solved. Thus, the 1990s into the 2000s saw article after article in statistics as well as methodological social sciences that used MCMC tools to estimate increasingly intricate models from increasingly complex theories, all from a Bayesian perspective.

## POSTMODERN BAYES

It gradually became apparent that the MCMC revolution was about more than just estimating models that had frustrated Bayesians of the previous generation. It turns out that estimation with Bayesian stochastic simulations (MCMC) provides opportunities to extend modeling and to produce quantities of interest beyond regular posterior inference. So, if the modern Bayesian era is

characterized by freedom to estimate pent-up models from 100 years of frustration, then the postmodern Bayesian era is defined by a realization that Bayesian stochastic simulation does not just *allow* estimation (marginalization) of previously inestimable models, it also (almost inadvertently) gives additional inferential information that can be exploited for enhanced purposes. The underlying point is that empirical description, rather than just analytical math-stat description, of posterior parameters of interest gives distributional information that can be used for other purposes such as model checking, model comparison, and enhanced specifications. This section provides a sense of these new tools through several described examples.

**Poster Predictive Checks**

A very useful way to judge model quality is the posterior predictive check, an approach that compares fabricated data implied by the model fit to the actual outcome data. The general idea is that, if the predictions closely resemble the truth, then the model is fitting well. In addition, deviations from an ideal fit are often informative about where the model could fit better in terms of direction or category. In the most basic form, we are simply comparing (usually plotting) the  $y_i, \dots, y_n$  outcome values from the data with the  $\hat{y}_i, \dots, \hat{y}_n$  values produced from the model. In the case of a linear model, this is incredibly simple in the Bayesian or non-Bayesian context as  $\hat{y} = X\hat{\beta}$ , but with more complex specifications, it may require more involved calculations. In the non-Bayesian sense, or in the simplified Bayesian sense, we can often analytically calculate  $\hat{y}_i$  values and, importantly, measures of uncertainty for these values that allow us to measure or plot accuracy.

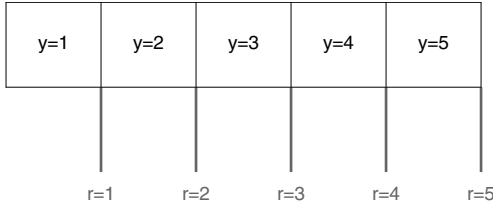
As it turns out, it is an ancillary benefit of the MCMC process that it is not only easy to calculate these posterior predictive values for the outcome variable, but it is also almost ‘free’ to get measures of uncertainty since

the empirical distributional descriptions of the estimated model parameters can be made to flow through to the predicted quantities. More specifically, for  $K$  explanatory variables (including a constant if appropriate) define  $\beta^{(s)} = \beta_{k=1, \dots, K, s=1, \dots, n_{sim}}$  for  $n_{sim}$  (a large) number of MCMC iterations collected after the assertion of convergence. In a fairly general sense, this lets us produce  $\hat{y}^{(s)} = f(X, y, \beta^{(s)})$  values from the Bayesian model specification (including prior distributions), giving  $n_{sim}$  values from the full distribution of the poster predicted values.

Returning to the public executives data from Gill and Waterman (2004), subsample  $n_{sample} = 10,000$  values from the total post convergence MCMC runs to create  $\bar{\beta}^{(s)}$  for the regression parameters and set  $\bar{\theta} = (\theta_{r=1}, \dots, \theta_{r=5})$  as the mean of the thresholds parameters across all of the 500,000 saved MCMC runs (although we can also make this component stochastic if desired). The first, and most elementary, summary uses the mean of the data vector  $\bar{X}$  to create an archetypal simplified data case (e.g. the mean taken down columns of the  $K$  explanatory variable matrix). Define first  $\bar{\mu} = \bar{X}\bar{\beta}^{(s)}$  temporarily averaging the simulated coefficient values, then for the ordered logit specification we create the cumulative and marginal outcome probabilities according to

$$\begin{aligned}
 P_{cumulative}(y \leq r) &= \left[ 1 + \exp(-\bar{\theta}_r - \bar{\mu}) \right]^{-1} \\
 r &= 1, \dots, 5 \\
 P_{marginal}(y = r) &= P_{cumulative}(y \leq r) - P_{cumulative}(y \leq r-1) \\
 &= P_{cumulative}(y \leq r), \quad r = 2, \dots, 4 \\
 P_{marginal}(y = 1) &= P_{cumulative}(y = 1) \\
 P_{marginal}(y = 5) &= 1 - P_{cumulative}(y \leq 4).
 \end{aligned}
 \tag{10}$$

This is illustrated in Figure 50.2 showing the outcome values and the threshold values. Thus, we have five ordered categorical probabilities averaged across cases and averaged across simulated estimations given by [0.0001, 0.0019, 0.0865, 0.9040, 0.0075]. This compares somewhat unfavorably to the



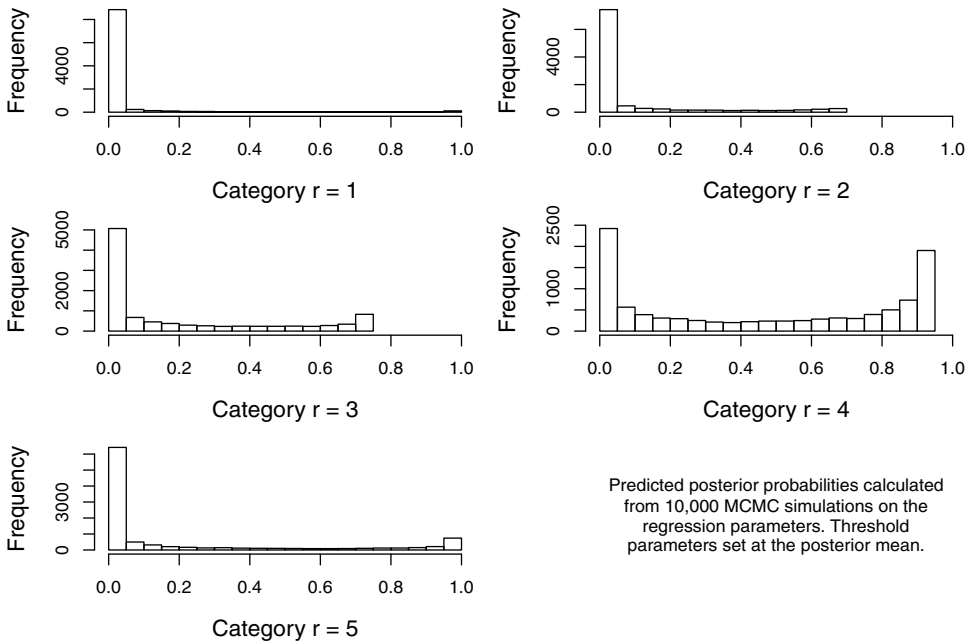
**Figure 50.2 Categories illustration**

distribution of actual outcomes in the  $n = 512$  size data: [51, 54, 96, 200, 131].

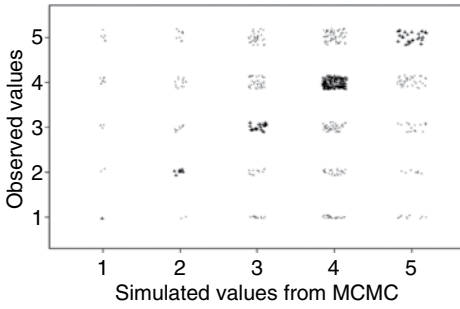
Now suppose we keep the mean vector of the data values but replace the mean of the MCMC parameter draws with the  $n_{sample} = 10,000$ . This is done exactly as just described, where (10) is done 10,000 separate times with 10,000 MCMC sub draws producing a  $(10,000 \times 5)$  matrix of probability vectors by row. From this we gain the measure of uncertainty on the parameter estimates that is provided by draws from the posterior distributions through the MCMC process. The result of this process is 10,000 draws

describing the unconditional probabilities for the complete sample of 512 individuals (averaged) for each of  $P_{marginal}(y) r = 1, \dots, 5$ . These are summarized in Figure 50.3. Notice that the marginal probabilities differ substantially across each category with interesting bunching at extremes in some categories reflecting strong covariate information that flows through these predictions.

This analysis provides useful information, but it is still an incomplete picture because variance across the 512 cases is still suppressed (averaged over). To unlock this added level of uncertainty, we make further use of the MCMC draws using the same procedure as done with the average person case but now expand the data structures to let the posterior variability in the estimation process flow down to individual level predictions that include individual level covariate differences. This is literally done in a loop in the code whereby each of the 512 cases is predicted instead of a mean case done at first above. Figure 50.4 shows a random selection of



**Figure 50.3 Posterior predictive probabilities by ordered outcomes**



**Figure 50.4 Posterior predictive probabilities by outcome categories**

10,000 of the 500,000 MCMC draws again (more can be done but it crowds the figure), where the resulting  $\hat{y}$  posterior predicted value is plotted against the  $y$  value for each of the 512 data cases, where each are jittered (the addition of random noise) to slightly separate cases visually for such categorical data. What we see here is a pretty good fit to the data whereby many of the cross-plotted points are on the main diagonal of the plot where  $y$  and  $\hat{y}$  take on the same values. What we also see is a slight underestimation for cases where the true observed values are in categories 4 and 5. This could not be shown without letting the full uncertainty flow down to the individual predictions, showing the usefulness of the distributional information from the MCMC simulations.

The point of this subsection is that the assessment of model quality through prediction is straightforward (easy actually) with MCMC output because the parameter estimation comes from a large number of draws from the posterior distribution of these parameters. Thus, each of these draws from the underlying Bayesian distributions of the model can be ‘flowed’ through to quantities of interest like outcome predictions with the uncertainty preserved through the empirical draws. Therefore, the MCMC process actually makes this process easier since complex mathematical-statistics analytical calculations are completely unneeded now, including

lengthy derivations, transformations, use of the delta method, and more.

### Hamiltonian Monte Carlo

Hamiltonian Monte Carlo (or analogously Hybrid Monte Carlo; henceforth HMC) is a modification of the standard Metropolis–Hastings algorithm that uses ‘physical system dynamics’ as a means of generating candidate values for a Metropolis decision to move to a new point in the state space. The Metropolis–Hastings algorithm (Metropolis et al., 1953) is a Markovian accept-reject procedure that moves through the multidimensional space of interest. The basic Metropolis–Hastings algorithm for a single selected parameter vector starts with a  $J$ -length parameter vector,  $\theta \in \Theta^J$ , to empirically describe target distribution of interest,  $\pi(\theta)$ , by ‘visiting’ substates proportionally to the density of interest. At the  $t^{th}$  step of the Markov chain ( $t$  stands for time), the chain is at the position indicated by the vector  $\theta^{[t]}$ . We then draw  $\theta'$  from a distribution over the same support, from a *proposal distribution* denoted  $q_t(\theta'|\theta)$ . This function must be ‘reversible’, meaning that  $\pi(\theta)p(\theta, \theta') = \pi(\theta')p(\theta', \theta) \forall \theta, \theta'$  in order to pick from a wide range of non-symmetric forms, where  $p(\cdot)$  is the *actual transaction function* – the probability of generating a candidate *and* accepting it – and  $\pi(\cdot)$  is the target density. We then decide to move with probability

$$\min \left[ \frac{\pi(\theta') q(\theta|\theta')}{\pi(\theta) q(\theta'|\theta)}, 1 \right]. \tag{11}$$

Importantly, rejecting  $\theta'$  means moving to  $\theta$  (the current position) as the new position in the time series of samples. An important feature of the algorithm is the flexibility of the choice of the proposal distribution. Many variations are based on strategic choices for this distribution. For example, the Hit and Run algorithm (Chen and Schmeiser, 1993)

separates the direction and distance decision in the proposal so that it can be optimized for highly constrained posterior structures and the algorithm does not reject a large number of candidate destinations.

The HMC algorithm exploits flexibility in the choice of candidate distribution by incorporating information about posterior topography to traverse the sample space more efficiently. Topography in this sense means the curvature of the target distribution, which is easy for humans to visualize in three dimensions and impossible for humans to see in the 8–12 dimensions that political science models often specify. The basic idea predates the modern advent of MCMC from Duane et al. (1987) and was developed in detail by Neal (1993, 2011).

Like the original Metropolis (1953) algorithm, Hamiltonian Monte Carlo comes from physics (Meyer, Hall and Offin 1992). Here we are concerned with an object’s trajectory within a specified multidimensional system as a way to describe joint posterior distributions. Now define  $\vartheta_t$  as a  $k$ -dimensional location vector and  $\mathbf{p}_t$  as a  $k$ -dimensional momentum (e.g. mass times velocity) vector, both recorded at time  $t$ . The Hamiltonian system at time  $t$  with  $2k$  dimensions is given by the joint Hamiltonian function

$$H(\vartheta_t, \mathbf{p}_t) = U(\vartheta_t) + K(\mathbf{p}_t), \quad (12)$$

where  $U(\vartheta_t)$  is the function describing the *potential energy* at the point  $\vartheta_t$ , and  $K(\mathbf{p}_t)$  is the function describing the *kinetic energy* for momentum  $\mathbf{p}_t$ . Neal (2011) gives the simple one-dimensional example

$$U(\vartheta_t) = \frac{\vartheta_t^2}{2} \quad K(\mathbf{p}_t) = \frac{p_t^2}{2}, \quad (13)$$

which is equivalent to a standard normal distribution for  $\vartheta$ . Commonly, the kinetic energy function is defined as

$$K(\mathbf{p}_t) = \mathbf{p}_t' \Sigma^{-1} \mathbf{p}_t, \quad (14)$$

where  $\Sigma$  is a symmetric and positive-definite matrix that can be as simple as an identity matrix times some scalar that can serve the role of a variance:  $\Sigma = \sigma^2 \mathbf{I}$ . This simple form is equivalent to the log PDF of the multivariate normal with mean vector zero and variance-covariance matrix  $\Sigma$ .

Hamiltonian dynamics describe the gradient-based way that potential energy changes to kinetic energy and kinetic energy changes to potential energy as the object moves over time throughout the system (multiple objects require equations for gravity, but that is fortunately not our concern here). The mechanics of this process are given by Hamilton’s equations, which are the set of simple differential equations

$$\begin{aligned} \frac{\partial \vartheta_{it}}{\partial t} &= \frac{\partial H}{\partial \mathbf{p}_{it}} = \frac{K(\partial \mathbf{p}_{it})}{\partial \mathbf{p}_{it}} \\ \frac{\partial \mathbf{p}_{it}}{\partial t} &= -\frac{\partial H}{\partial \vartheta_{it}} = -\frac{U(\partial \vartheta_{it})}{\partial \vartheta_{it}} \end{aligned} \quad (15)$$

for dimension  $i$  at time  $t$ . For continuously measured times these equations give a mapping from time  $t$  to time  $t + \tau$ , meaning that from some position  $\vartheta_t$  and momentum  $\mathbf{p}_t$  at time  $t$  we can predict  $\vartheta_\tau$  and  $\mathbf{p}_\tau$ . Returning to the one-dimensional standard normal case, these equations are simply  $d\vartheta_t/dt = p$  and  $dp/dt = -\vartheta$ .

There are three important properties of Hamiltonian dynamics that are actually *required* if we are going to use them to construct an MCMC algorithm (Neal, 2011). First, Hamiltonian dynamics is *reversible*, meaning that the mapping from  $(\vartheta_t, \mathbf{p}_t)$  to  $(\vartheta_{t+\tau}, \mathbf{p}_{t+\tau})$  is one-to-one and therefore also defines the reverse mapping from  $(\vartheta_{t+\tau}, \mathbf{p}_{t+\tau})$  to  $(\vartheta_t, \mathbf{p}_t)$ . Second, *total* energy is conserved over time  $t$  and dimension  $k$ , and the Hamiltonian is invariant, as shown by

$$\begin{aligned} \frac{\partial H}{\partial t} &= \sum_{i=1}^k \left[ \frac{\partial \vartheta_i}{\partial t} \frac{\partial H}{\partial \vartheta_i} + \frac{\partial \mathbf{p}_i}{\partial t} \frac{\partial H}{\partial \mathbf{p}_i} \right] \\ &= \sum_{i=1}^k \left[ \frac{\partial H}{\partial \mathbf{p}_i} \frac{\partial H}{\partial \vartheta_i} - \frac{\partial H}{\partial \vartheta_i} \frac{\partial H}{\partial \mathbf{p}_i} \right] = 0. \end{aligned} \quad (16)$$

This provides detailed balance (reversibility) for the MCMC algorithm. Second, Hamiltonian dynamics preserve volume in the  $2k$  dimensional space. In other words, elongating some region in a direction requires withdrawing another region as the process continues over time. This ensures that there is no change in the scale of Metropolis–Hastings acceptance probability. Finally, Hamiltonian dynamics provides a *symplectic mapping* in  $\mathcal{R}^{2k}$  space. Define first the smooth mapping  $\psi : \mathcal{R}^{2k} \rightarrow \mathcal{R}^{2k}$  with respect to some constant and invertible matrix  $\mathbf{J}$  with  $\mathbf{J}' = -\mathbf{J}$  and  $\det(\mathbf{J}) \neq 0$ , along with having Jacobian  $\psi(z)$  for some  $z \in \mathcal{R}^{2k}$ . The mapping  $\psi$  is symplectic if

$$\psi(z)' \mathbf{J}^{-1} \psi(z) = \mathbf{J}^{-1}. \tag{17}$$

Leimkuhler and Reich (2005: 53) give the following mapping in 2-dimensional space  $z = (\vartheta, p)$ :

$$\psi(\vartheta, p) = \begin{bmatrix} p \\ 1 + b\vartheta + ap^2 \end{bmatrix}, \tag{18}$$

with constants  $a, b \neq 0$ . The Jacobian of  $\psi(\vartheta, p)$  is calculated by

$$\frac{\partial}{\partial \vartheta} \frac{\partial}{\partial p} \psi(\vartheta, p) = \begin{bmatrix} 0 & 1 \\ b & 2ap \end{bmatrix}. \tag{19}$$

We check symplecticness by

$$\begin{aligned} & \left[ \frac{\partial}{\partial \vartheta} \frac{\partial}{\partial p} \psi(\vartheta, p) \right]' \mathbf{J}^{-1} \left[ \frac{\partial}{\partial \vartheta} \frac{\partial}{\partial p} \psi(\vartheta, p) \right] \\ &= \begin{bmatrix} 0 & 1 \\ b & 2ap \end{bmatrix}' \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ b & 2ap \end{bmatrix} \tag{20} \\ &= -b\mathbf{J}^{-1}. \end{aligned}$$

Thus we say that  $\psi(\vartheta, p)$  is symplectic for  $b = -1$  and any  $a \neq 0$ .

Everything discussed so far assumed continuous time, but obviously for a computer implementation in a Markov chain Monte

Carlo context, we need to discretize time. Thus, we will grid  $t + \tau$  time into intervals of size  $v$ :  $v, 2v, 3v, \dots, mv$ . We need a way to obtain this discretization while preserving volume, and so we use a tool called the *leapfrog methods*. The notation is more clear if we now move  $t$  from the subscript to functional notation:  $\vartheta(t)$  and  $p(t)$ , which is also a reminder that time is now discrete rather than continuous. To complete a single step starting at time  $t$ , first update each of the momentum dimensions by  $v/2$  with the following:

$$p_i \left( t + \frac{v}{2} \right) = p_i(t) - \frac{v}{2} \frac{\partial U(\vartheta_t)}{\partial \vartheta_i(t)}. \tag{21}$$

Now take a full  $v$ -length step to update each of the position dimensions to leapfrog over the momentum:

$$\vartheta_i(t+v) = \vartheta_i(t) + v \frac{\partial K(p_i)}{\partial p_i \left( t + \frac{v}{2} \right)}. \tag{22}$$

Then finish with the momentum catching up in time with the step:

$$p_i(t+v) = p_i \left( t + \frac{v}{2} \right) - \frac{v}{2} \frac{U(\vartheta_t)}{\partial \vartheta_i(t+v)}. \tag{23}$$

Notice that the leapfrog method is reversible since it is a one-to-one mapping from  $t$  to  $t+v$ . Obviously, running these steps  $M$  times completes the Hamiltonian dynamics for  $M \times v$  periods of total time. The determination of  $v$  is a key tuning parameter in the algorithm since smaller values give a closer estimation to continuous time but also add more steps to the algorithm.

A Metropolis–Hastings algorithm is configured such that the Hamiltonian function serves as the candidate-generating distribution. This requires connecting the regular posterior density function,  $\pi(\theta)$ , to a potential energy function,  $U(\vartheta_t)$ , where a kinetic energy function,  $K(p_i)$ , serves as a (multidimensional and necessary) auxiliary variable. This connection is done via the

canonical distribution commonly used in physics,

$$p(x) = \frac{1}{Z} \exp\left[-\frac{E(x)}{T}\right], \quad (24)$$

where  $E(x)$  is the energy function of some system at state  $x$ ,  $T$  is the temperature of the system (which can simply be set at 1), and  $Z$  is just a normalizing constant so that  $p(x)$  is a regular density function. In the Hamiltonian context (Equation 24) is

$$\begin{aligned} p(\vartheta, \mathbf{p}) &= \frac{1}{Z} \exp\left[-\frac{H(\vartheta, \mathbf{p})}{T}\right] \\ &= \frac{1}{Z} \exp\left[-\frac{U(\vartheta_t) + K(\mathbf{p}_t)}{T}\right] \\ &= \frac{1}{Z} \exp\left[-\frac{U(\vartheta_t)}{T}\right] \exp\left[-\frac{K(\mathbf{p}_t)}{T}\right], \end{aligned} \quad (25)$$

demonstrating that  $\vartheta$  and  $\mathbf{p}$  are independent. Finally, we connect the energy function metric with the regular posterior density metric with the function

$$E(\vartheta) = -\log(\pi(\vartheta)), \quad (26)$$

thus completing the connection. Notice that the  $\theta$  variables must all be continuous in the model, although Hamiltonian Monte Carlo can be combined with other MCMC strategies in a hybrid algorithm.

The Hamiltonian Monte Carlo algorithm uses two general steps at time  $t$ :

- Generate, independent of the current  $\vartheta_t$ , the momentum  $\mathbf{p}_t$  from the multivariate normal distribution implied by  $K(\mathbf{p}_t) = \mathbf{p}_t' \Sigma^{-1} \mathbf{p}_t$  with mean vector zero and variance-covariance matrix  $\sigma^2 I$  (or some other desired symmetric and positive-definite form).
- Run the leapfrog method  $M$  times with  $\nu$  steps to produce the candidate  $(\tilde{\vartheta}, \tilde{\mathbf{p}})$ .
- Accept this new location or accept the current location as the  $t + 1$  step with a standard Metropolis decision using the  $H$  function

$$\min\left[1, \exp(-H((\tilde{\vartheta}, \tilde{\mathbf{p}})) + H(\vartheta, \mathbf{p}))\right]. \quad (27)$$

While this process looks simple, there are several complications to consider. We must be able to take the partial derivatives of the log-posterior distribution, which might be hard. The chosen values of the leapfrog parameters,  $M$  and  $\nu$ , are also critical. If  $\nu$  is too small then exploration of the posterior density will be very gradual with small steps, and if  $\nu$  is too big then many candidates will be rejected. Choosing  $M$  is important because this parameter allows the Hamiltonian process to explore strategically with respect to gradients. Excessively large values of  $M$  increase compute time, but excessively small values of  $M$  lead to many rejected candidates. In both cases where the parameters are too small, we lose the advantages of the gradient calculations and produce an inefficient random walk. Finally,  $\sigma^2$  affects efficiency of the algorithm in the conventional sense of appropriating tuning the variance of the multivariate normal for the momentum. These can be difficult parameter decisions and Neal (2011) gives specific guidance on trial runs and analysis of the results. The Hamiltonian Monte Carlo dynamics are difficult to implement in complex multilevel generalized linear models that aim to apply full Bayesian inference. While these models can be carried out with BUGS or JAGS, this takes an enormous amount of time and computational resources. To circumvent this, a group of academics, among them Andrew Gelman and Bob Carpenter, developed STAN. STAN is written in C++ and, unlike BUGS and JAGS, employs reverse-mode algorithmic differentiation to implement HMC in a much faster way. It supports a range of functions (e.g. probability functions, log gamma, inverse logit) and integrates matrix operations on linear algebra. See <https://mc-stan.org/> for details and downloads.

### Bayes Factor Calculations

Another example where MCMC output makes mathematical calculations much easier is the calculation of the Bayes Factor



for non-linear regression models. This also follows the principle that MCMC simulation provides a natural distributional summary that can be used for multiple purposes besides the original purpose of simply producing marginal distributions from a complicated joint distribution from the Bayesian model specification. In the classical era, it was recognized that Bayes Factors were an extremely useful model assessment and comparison tool going back to Jeffreys (1983), but they were often very difficult to calculate for realistic regression-style models.

Bayes Factors start with observed data  $x$  for testing two models, with associated parameter vectors  $\theta_1$  and  $\theta_2$ :  $M_1: f_1(x|\theta_1)$   $M_2: f_2(x|\theta_2)$ . Here these parameter vectors can define nested or non-nested alternatives, unlike the more simple likelihood ratio test. With prior distributions,  $p_1(\theta_1)$  and  $p_2(\theta_2)$ , and prior probabilities on the two models,  $p(M_1)$   $p(M_2)$ , we can produce the odds ratio for Model 1 versus Model 2 by Bayes' Law:

$$\frac{\pi(M_1 | x)}{\pi(M_2 | x)} = \frac{p(M_1) / p(x)}{p(M_2) / p(x)} \times \frac{\int_{\theta_1} f_1(x | \theta_1) p_1(\theta_1) d\theta_1}{\int_{\theta_2} f_2(x | \theta_2) p_2(\theta_2) d\theta_2} \tag{28}$$

Bayes Factor

So, we are actually interested in the ratio of marginal likelihoods – (Equation 5) from the two models. By canceling and algebraically rearranging, we get the common form of the Bayes Factor:

$$BF_{(1,2)} = \frac{\pi(M_1 | x) / p(M_1)}{\pi(M_2 | x) / p(M_2)} \tag{29}$$

(Gill, 2014). As suggested by these forms, analytical calculation for reasonably realistic social science regression models can be challenging. Fortunately, this is direct and easy for most Bayesian generalized linear

models estimated with MCMC. Chib (1995) and Chib and Jeliazkov (2001), for instance, give a handy and generalizable recipe in the context of probit regression models. To begin, rearrange (Equation 4) and take logs for a single model (for the moment) using the log-likelihood:

$$\log p(x) = \ell(\theta' | x) + \log p(\theta') - \log \pi(\theta' | x). \tag{30}$$

Here  $\theta'$  is a completely arbitrary point in the appropriate sample space, such as a point in the high density region, for instance the posterior mean. To start, we use  $\pi(\theta' | x)$  from simulation for a generic MCMC estimation approach (details in Chapter 14 of Gill, 2014). Define the probability of the Metropolis–Hastings Markov chain as transitioning to an arbitrary point  $\theta'$  from a starting point  $\theta$  with the candidate-generating distribution that produces  $\theta'$  times the probability that it is accepted from above:

$$\alpha(\theta', \theta) = \min \left[ \frac{\pi(\theta') q_r(\theta | \theta')}{\pi(\theta) q_t(\theta' | \theta)}, 1 \right]. \tag{31}$$

This candidate-generating distribution produces  $\theta'$  times the probability that it is accepted from above:  $p(\theta, \theta') = q(\theta' | \theta) \alpha(\theta', \theta)$ , such that for any arbitrary point detailed balance is preserved,  $\pi(\theta) q(\theta' | \theta) \alpha(\theta', \theta) = \pi(\theta') q(\theta | \theta') \alpha(\theta, \theta')$ . Now take integrals of both sides with respect to  $\theta$ , realizing that  $\pi(\theta')$  is a function evaluation at an arbitrary point and can therefore be moved outside of the integration process, and, with some algebra, reach

$$\pi(\theta') = \frac{\int \Theta \pi(\theta) q(\theta' | \theta) \alpha(\theta', \theta) d\theta}{\int_{\Theta} q(\theta | \theta') \alpha(\theta, \theta') d\theta} = \frac{E_{\pi(\theta)} [q(\theta' | \theta) \alpha(\theta', \theta)]}{E_{q(\theta|\theta')} [\alpha(\theta, \theta')]} \tag{32}$$

which is simply the ratio of two expected value calculations: Chib and Jeliazkov (2001) observed that in the course of running an MCMC estimation process for marginal posterior distributions, we can get the marginal likelihood without extra trouble replacing (Equation 32) with its simulation version,

$$\pi_{sim}(\theta') = \frac{\frac{1}{M} \sum_{m=1}^M \alpha(\theta', \theta_m) q(\theta' | \theta_m)}{\frac{1}{N} \sum_{n=1}^N \alpha(\theta_N, \theta')}, \quad (33)$$

which uses known quantities readily at hand for some number of simulations  $M$ . Here  $\theta'$  is chosen arbitrarily but within a high-density region of the posterior distribution. Therefore, this process substitutes a challenging integration process with simulation of the posterior density at a single point by completing (Equation 30) with the simulated result

$$\log p_{sim}(\mathbf{x}) = \ell(\theta' | \mathbf{x}) + \log p(\theta') - \log \pi_{sim}(\theta' | \mathbf{x}), \quad (34)$$

where all of these quantities are easily available.

**Bayesian Nonparametrics**

We are concerned in this section with how *nonparametric priors* can enhance the increasing use of Bayesian models for the presence of unobserved heterogeneity, which is a common problem across the social sciences. Researchers commonly deal with the problem by specifying non-unique random effect terms  $\psi_j, j \in J < n$  to capture grouping or clustering information where the mapping from  $i = 1, \dots, n$  to  $j = 1, \dots, J$  is known.

Generally, the distribution of the  $\psi_i$  is unknown but safely assumed through custom or testing. The normal distribution is a useful default for both practical and asymptotic reasons. This convenience cannot be directly

tested through residuals analysis but affects overall model fit, which can be tested. In the Bayesian setting, a better and more flexible alternative exists as a so-called nonparametric Bayesian prior in which  $\psi_j$  is drawn from a vastly more flexible distributional setup, starting with

$$\begin{aligned} (Y_1, \dots, Y_n) &\sim f(y_1, \dots, y_n | \beta, \psi_1, \dots, \psi_n) \\ &= \prod_i f(y_i | \beta, \psi_i), \quad \psi_i \sim G, \quad i = 1, \dots, n, \end{aligned} \quad (35)$$

where  $f$  can be taken as normal and where a popular choice for  $G$  is the Dirichlet Process (DP),

$$\psi_i \sim G \sim \mathcal{DP}(\lambda, \phi_0), \quad i = 1, \dots, n, \quad (36)$$

with base measure  $\phi_0$  and precision parameter  $\lambda$ . In particular, the observations are modeled as

$$Y_i = X_i \beta + \psi_i + \epsilon_i, \quad (37)$$

where the  $\epsilon_i$  are treated as independent normal random variables and  $\psi_i$  indicates the random effects assignment for the  $i$ th case. Alternatively, specification of a link function turns this into a generalized linear model (GLM) in the classic sense,  $\hat{Y}_i = g^{-1}(X_i \beta + \psi_i)$ . Since the  $\psi_i$  are drawn from a DP distribution, they are not necessarily unique and thus can be represented by a  $K$ -vector,  $\eta$ , where  $K < n$ . Furthermore, the model can be written as

$$Y = X \beta + A \eta + \epsilon, \quad (38)$$

where  $\psi = A \eta$  and  $A$  is an  $n \times K$  matrix of zeros with a single one in each row which denotes the specific  $\eta_k$  assigned to  $\psi_i$  (Kyung et al., 2010).

Dirichlet process mixture models were originally formulated by Ferguson (1973), who defined the underlying process and derived the key properties. Blackwell and MacQueen (1973) then showed that the

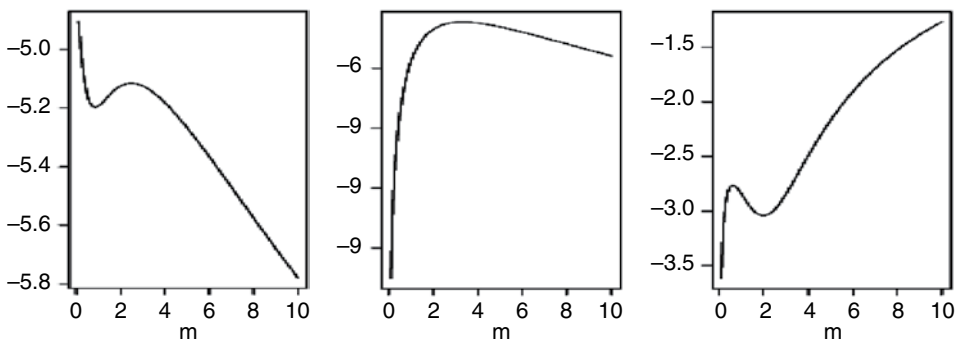
marginal distribution for the Dirichlet process can be treated as that of the  $n^{\text{th}}$  step of a Polya urn process. Other key theoretical work includes Korwar and Hollander (1973) and Sethuraman (1994). The contributions that have particular importance for the likelihood function development are that of Lo (1984), who derives the analytic form of a Bayesian density estimator, and Liu (1996), who derives an identity for the profile likelihood estimator of  $m$ . This is an interesting sociology of science story in that these works mostly predated the computational tools that made the models possible for real datasets. Follow-on work that changed this state are typified by Escobar and West (1995), MacEachern and Müller (1998), and Jain and Neal (2004).

The model specified in (Equation 35) is actually a classical semiparametric random effects model and, with further Bayesian modeling of the parameters, can be implemented with MCMC. Unfortunately, the presence of the Dirichlet term makes the use of the standard Gibbs sampler somewhat complicated in non-conjugate situations such as with is the model that was developed in Gill and Casella (2009). These authors find that this approach can model difficult data and produce results that existing alternative methods fail to discover. They then account for unobserved, important clustering structures with the non-parametric process that do not necessarily reflect intervening or confounding variables

but still provide information about agency environment that was not explicitly available.

Gill and Casella (2009) introduced a GLMDM with an ordered probit link to model political science data, specifically modeling the stress, from the Gill and Waterman (2004) data already described here. Their Dirichlet precision parameter was not an influential model parameter and was therefore fixed at a value that made the MCMC sampler more efficient. Kyung et al. (2010, 2012), looked at the maximum likelihood estimation of the precision parameter and found that the standard approach to finding the maximum likelihood estimate, given in Liu (1996), could yield a maximum, a minimum, or even a ridge. Figure 50.5, from Kyung et al., (2010), shows some observed shapes of this likelihood function for simulated circumstances. Since likelihood estimation is not reliable for this parameter, Kyung, et al., (2010) proved that introducing a prior distribution on the precision parameter guarantees an interior mode and so stabilizes the estimation procedure.

Models with Dirichlet process priors are treated as hierarchical models (Gill and Womack, 2013) in a Bayesian framework, and the implementation of these models through Bayesian computation and efficient algorithms has had substantial progress. Escobar and West (1995) produced a Gibbs sampling algorithm for the estimation of posterior distribution for all model parameters and the



**Figure 50.5** Log-likelihood functions for configurations of component likelihoods

direct evaluation of predictive distributions. MacEachern and Müller (1998) developed a Gibbs sampler with non-conjugate priors by using auxiliary parameters, and Neal (2000) provided an extended and more efficient Gibbs sampler to handle general Dirichlet process mixture models with non-conjugate priors by using a set of auxiliary parameters. Teh et al. (2006) extended the method of Escobar and West for posterior sampling of the precision parameter with a gamma distributed prior. Kyung et al.(2011) extended these results to a generalized Dirichlet process mixed model with a probit link function. They derived a Gibbs sampler for the model parameters and the important subclusters of the Dirichlet process using new parameterization of the hierarchical model to derive a Gibbs sampler that more fully uses the structure of the model.

Again,  $\mathbf{X}_i$  are the covariates associated with the  $i^{\text{th}}$  observation,  $\boldsymbol{\beta}$  be the coefficient vector, and  $\psi_i$  be the random effect accounting for subject-specific deviation from the underlying model. Assume that  $Y_i|\boldsymbol{\psi}$  are conditionally independent, each with a density from the exponential family, where  $\boldsymbol{\psi} = (\psi_1, \dots, \psi_J)$ ,  $J < n$ . Based on the notation on McCulloch et al. (2008), the Generalized Linear Mixed Dirichlet Model is expressed as

$$\begin{aligned}
 Y_i | \boldsymbol{\psi} &\overset{ind}{\sim} f_{Y_i|\boldsymbol{\psi}}(y_i | \boldsymbol{\psi}), \quad i = 1, \dots, n \\
 f_{Y_i|\boldsymbol{\psi}}(y_i | \boldsymbol{\psi}) &= \exp\left[\{y_i\gamma_i - b(\gamma_i)\} / \xi^2 - c(y_i, \xi)\right],
 \end{aligned}
 \tag{39}$$

where  $y_i$  is assumed discrete valued. It is assumed that  $[Y_i | \boldsymbol{\psi}] = \mu_i = \partial b(\gamma_i) / \partial \gamma_i$ . Using the arbitrary link function  $g(\cdot)$ , we can express the transformed mean of  $Y_i$ ,  $E[Y_i | \boldsymbol{\psi}]$ , as a linear function,  $g(\mu_i) = \mathbf{X}\boldsymbol{\beta} + \psi_i$ . For the Dirichlet process mixture models, we assume that

$$\begin{aligned}
 \psi_i &\sim G \\
 G &\sim \mathcal{DP}(mG_0),
 \end{aligned}
 \tag{40}$$

where  $\mathcal{DP}$  is the Dirichlet Process with base measure  $G_0$  and precision parameter  $m$ . Blackwell and MacQueen (1973) proved that for  $\psi_1, \dots, \psi_n$  iid from  $G \sim \mathcal{DP}$ , the joint distribution of  $\boldsymbol{\psi}$  is a product of successive conditional distributions of the mixture form

$$\begin{aligned}
 \psi_i | \psi_1, \dots, \psi_{i-1}, m &\sim \frac{m}{i-1+m} g_0(\psi_i) \\
 &+ \frac{1}{i-1+m} \sum_{l=1}^{i-1} \delta(\psi_l = \psi_i)
 \end{aligned}
 \tag{41}$$

where  $\delta(\cdot)$  denotes the Dirac delta function and  $g_0(\cdot)$  is the density function of base measure. The likelihood function from Liu (1996) and Lo (1984) is produced by integrating over the random effects,

$$\begin{aligned}
 L(\boldsymbol{\theta} | \mathbf{y}) &= \frac{\Gamma(m)}{\Gamma(m+n)} \sum_{k=1}^n m^k \\
 &\sum_{C:|C|=k} \prod_{j=1}^k \Gamma(n_j) \int f(\mathbf{y}_{(j)} | \boldsymbol{\theta}, \boldsymbol{\psi}_j) dG_0(\boldsymbol{\psi}_j),
 \end{aligned}
 \tag{42}$$

where  $C$  defines the subclusters (not actual clusters in the social science since there is no penalty here for over-fitting in the algorithm),  $\mathbf{y}_{(j)}$  is the vector of  $y_i$ s that are in subcluster  $j$ , and  $\boldsymbol{\psi}_j$  is the common parameter for that subcluster. There are  $S_{n,k}$  different subclusters  $C$ , the Stirling Number of the Second Kind (Abramowitz and Stegun, 1972: 824–825). Now we consider again the  $n \times k$  matrix  $\mathbf{A}$  defined by

$$\mathbf{A} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix},$$

where  $a_i$  is a  $1 \times k$  vector of all zeros except for a 1 in one position that indicates which group the observation is from. So, each column of matrix  $\mathbf{A}$  represents a partition of

the sample of size  $n$  into  $k$  groups. If the sub-cluster  $C$  is partitioned into groups  $\{S_1, \dots, S_k\}$ , then if  $i \in S_j$ ,  $\psi_i = \eta_j$  and the random effect can be rewritten as  $\psi = A\eta$ , where  $\eta = (\eta_1, \dots, \eta_k)$  and  $\eta_j \stackrel{iid}{\sim} G_0$  for  $j = 1, \dots, k$ .

The results of implementing the GLMDM model for the data in Gill and Waterman (2004) are given in Table 50.2 as posterior quantiles. Notice first that these results differ markedly from the previous analysis of these data with a conventional Bayesian ordered choice model as summarized in Table 50.1. The nonparametric specific is a fundamentally different approach that includes and leverages underlying heterogeneity by accounting for subclusters in the estimation process. For instance, the effect of the variable `Committee Relationship` and stress is reliably in the opposite direction: closer ties to the oversight committee are associated with lower stress levels, when accounting for group level latent heterogeneity. This actually makes sense when considering the wide range of relationship types, policy spaces, and administrative

histories that exist between congressional committees and administrative agencies. So, the Dirichlet process that accounts for such underlying grouping reveals a different type of relationship effect. Moreover, the previously seen, positive relationship between `Career.Exec-Compet` and stress is now overturned: there is a reliably negative finding here likely recognizing agency heterogeneity in senior staffing. The hours per week worked have a positive relationship with stress. This statistically reliable finding is shown in the tightly bounded and positive quantiles for `Hours/Week`. Interestingly, political executives who required preparation for the hearings on their Senate confirmation later provided lower stress scores. This reliable finding is likely related to the committee relationship variable.

**CONCLUSION**

The purpose of this chapter is to introduce the Bayesian inferential process with a focus

**Table 50.2 Posterior quantiles, GLMDM for survey of political executives**

	0.01	0.025	0.25	0.5	0.75	0.975	0.99
<b>Explanatory variables</b>							
Government Experience	-0.1071	-0.0861	-0.0117	0.0275	0.0665	0.1409	0.1623
Ideology	-0.0421	-0.0309	0.0077	0.0280	0.0483	0.0870	0.0980
Committee Relationship	-0.3146	-0.3021	-0.2581	-0.2350	-0.2119	-0.1679	-0.1554
Career.Exec-Compet	-0.3600	-0.3431	-0.2823	-0.2505	-0.2186	-0.1579	-0.1404
Career.Exec-Liaison/Bur	-0.0371	-0.0240	0.0226	0.0470	0.0714	0.1181	0.1313
Career.Exec-Liaison/Cong	-0.1438	-0.1299	-0.0811	-0.0556	-0.0299	0.0191	0.0330
Career.Exec-Day2day	-0.3195	-0.3041	-0.2499	-0.2215	-0.1931	-0.1391	-0.1236
Career.Exec-Diff	-0.0383	-0.0241	0.0262	0.0525	0.0787	0.1288	0.1431
Confirmation Preparation	-0.6267	-0.5978	-0.4955	-0.4419	-0.3883	-0.2859	-0.2568
Hours/Week	0.3411	0.3509	0.3858	0.4040	0.4222	0.4571	0.4669
President Orientation	-0.6502	-0.6210	-0.5188	-0.4653	-0.4116	-0.3090	-0.2798
<b>Threshold intercepts</b>							
None-Little	-1.9915	-1.9580	-1.8402	-1.7782	-1.7160	-1.5979	-1.5644
Little-Some	-1.4407	-1.4096	-1.3010	-1.2439	-1.1866	-1.0778	-1.0466
Some-Significant	-0.9007	-0.7847	-0.3788	-0.1660	0.0473	0.4541	0.5718
Significant-Extreme	0.3811	0.4108	0.5157	0.5705	0.6254	0.7303	0.7598

on its implementation in political science and international relations. It is designed to highlight the practical history of how results are obtained in Bayesian analysis over the course of time. The hope is that readers will see both the principled theoretical advantages of thinking ‘Bayesianly’ and the practical ease with which results can today be produced and extended.

Bayesian inference is characterized by the explicit use of probability for describing uncertainty, which means probability models (likelihood functions) for data given parameters and probability distributions (PDFs and PMFs) for parameters. From this basis, inference proceeds from inference for unknown values conditioned on observed data with the use of inverse probability with Bayes’ Law to describe the full distribution of the unknown quantities with this update. Probability statements lie at the heart of Bayesian analysis. Everything a Bayesian does not know for a fact is modeled with probability distributions. At the core setup of Bayesian analysis, prior knowledge informs a specified probability model, which is then updated by conditioning on observed data and whose fit to the data is evaluated distributionally. The Bayesian paradigm fits closely with the core tenets of scientific discovery: current theories form the basis of stated prior information and informative evidence from data collection have the ability to update our theories. Contrary to traditional statistical thinking, however, it constitutes a different way of thinking about uncertainty that is strictly probability based, eschewing assumptions like an unending stream of independent and identically distributed data.

In the sections above, we identified three clear historical eras in the development of Bayesian methods: classical, modern, and postmodern. The classical era lasted until 1990 and was characterized by a determination that philosophical correctness should be recognized but was tempered with the challenges of estimating models with realistically

large and complicated social science data. Unfortunately at the time, it was not hard to create logical and mathematical arguments that showed the superiority of Bayesian inference over more traditional methods, but it was very hard, if not impossible, to apply these arguments empirically.

Gelfand and Smith changed this state of the Bayesian world and ushered in the modern era in 1990. They discovered Gibbs sampling, a tool with its roots in engineering and image restoration. Aided by improvements and availability in computing power, Gibbs sampling replaced analytical derivation of marginals from a joint with large numbers of draws sampled by the computer. Together with the Metropolis–Hastings algorithm unearthed from statistical physics, Gibbs sampling solved Bayesians’ problems and became known collectively as Markov chain Monte Carlo. Thanks to MCMC, models were no longer too complex for marginalization of joint posteriors to create regression tables. MCMC revolutionized Bayesian inference, released decades of frustration, and led to countless Bayesian applications and publications.

Finally, the postmodern era began in the early 21st century when researchers realized the full potential of MCMC beyond the estimation of previously inestimable models. It gradually became apparent that Bayesian stochastic simulation could also be exploited for enhanced purposes, such as model checking and model comparison. The consequence of this realization was a number of tools designed to extend the reach of MCMC, such as poster predictive checks, Hamiltonian Monte Carlo, Bayes Factor calculations, and Bayesian nonparametrics. As a result, it is now possible to not only easily produce Bayesian results, but also to extend the Bayesian paradigm and its application far beyond model estimation alone. This makes the Bayesian inferential process extraordinarily useful in political science and international relations.

## Note

- 1 Our thanks to the methodology reading group at American University: Le Bao, Ryan DeTamble, Michael Heseltine, Daisy Muibu, Abhishek Regmi, Samantha Senn, Rui Wang, Kumail Wasif, Morten Wendelbo.

## REFERENCES

- Abramowitz, Milton and Stegun, Irene A. (eds.) (1972) *Handbook of Mathematical Functions: With Formulas, Graphs, and Mathematical Tables*. Mineola: Dover Publications.
- Bayes, Thomas (1763) An Essay Towards Solving a Problem in the Doctrine of Chances, *Philosophical Transactions of the Royal Society of London*, 53: 370–418.
- Bernardo, José M. (1979) Reference Posterior Distributions for Bayesian Inference (with Discussion), *Journal of the Royal Statistical Society, Series B (Methodological)*, 41(2): 113–147.
- Birnbaum, Allan (1962) On the Foundations of Statistical Inference, *Journal of the American Statistical Association*, 57(298): 269–306.
- Blackwell, David and MacQueen, James B. (1973) Ferguson Distributions via Polya Urn Schemes, *The Annals of Statistics*, 1(2): 353–355.
- Box, George E.P. and Tiao, George C. (1973) *Bayesian Inference in Statistical Analysis*. New York: John Wiley & Sons.
- Brooks, Stephen P., Dellaportas, Petros and Roberts, Gareth O. (1997) An Approach to Diagnosing Total Variation Convergence of MCMC Algorithms, *Journal of Computational and Graphical Statistics*, 6(3): 251–265.
- Brooks, Stephen P. and Roberts, Gareth O. (1999) On Quantile Estimation and Markov Chain Monte Carlo Convergence, *Biometrika*, 86(3): 710–717.
- Chen, Ming-Hui and Schmeiser, Bruce (1993) Performance of the Gibbs, Hit-and-Run and Metropolis Samplers, *Journal of Computational and Graphical Statistics*, 2(3): 251–272.
- Chib, Siddhartha (1995) Marginal Likelihood from the Gibbs Output, *Journal of the American Statistical Association*, 90(432): 1313–1321.
- Chib, Siddhartha and Jeliazkov, Ivan (2001) Marginal Likelihood from the Metropolis-Hastings Output, *Journal of the American Statistical Association*, 96(453): 270–281.
- Compas, John (1969) Compound Decisions and Empirical Bayes, *Journal of the Royal Statistical Society, Series B*, 31: 397–423.
- Duane, Simon, Kennedy, Anthony D., Pendleton, Brian J. and Roweth, Duncan (1987) Hybrid Monte Carlo, *Physics Letters B*, 195(2): 216–222.
- Escobar, Michael D. and West, Mike (1995) Bayesian Density Estimation and Inference Using Mixtures, *Journal of the American Statistical Association*, 90(430): 577–588.
- Evans, Stephen (1994) Discussion of the Paper by Spiegelhalter, Freedman, and Parmar, *Journal of the Royal Statistical Society, Series A*, 157: 395.
- Ferguson, Thomas (1973) A Bayesian Analysis of Some Nonparametric Problems, *The Annals of Statistics*, 1(2): 209–230.
- Fisher, Ronald A. (1925a) *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd.
- Fisher, Ronald A. (1925b) Theory of Statistical Estimation, *Proceedings of the Cambridge Philosophical Society*, 22: 700–725.
- Gamerman, Dani and Lopes, Hedibert F. (2006) *Markov Chain Monte Carlo*, 2nd edition. New York: Chapman & Hall.
- Gelfand, Alan E. and Smith, Adrian F.M. (1990) Sampling Based Approaches to Calculating Marginal Densities, *Journal of the American Statistical Association*, 85(410): 398–409.
- Gelman, Andrew and Shalizi, Cosma R. (2013) Philosophy and the Practice of Bayesian Statistics, *British Journal of Mathematical and Statistical Psychology*, 66: 8–38.
- Geman, Stuart and Geman, Donald (1984) Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6): 721–741.
- Gilks, Walter R., Richardson Sylvia and Spiegelhalter, David J. (1996) *Markov Chain Monte Carlo In Practice*. New York: Chapman & Hall/CRC.
- Gill, Jeff (1999) The Insignificance of Null Hypothesis Significance Testing, *Political Research Quarterly*, 52(3): 647–74.

- Gill, Jeff (2014) *Bayesian Methods for the Social and Behavioral Sciences*. New York: Chapman & Hall/CRC.
- Gill, Jeff and Casella, George (2009) Nonparametric Priors for Ordinal Bayesian Social Science Models: Specification and Estimation, *Journal of the American Statistical Association*, 104(486): 453–454.
- Gill, Jeff and Torres, Michelle (2019) *Generalized Linear Models: A Unified Approach*, 2nd edition. Thousand Oaks: Sage.
- Gill, Jeff and Waterman, Richard (2004) Solidary and Functional Costs: Explaining the Presidential Appointment Contradiction, *Journal of Public Administration Research and Theory*, 14(4): 547–569.
- Gill, Jeff and Womack, Andrew (2013) The Multilevel Model Framework. Edited by: Marc A. Scott, Jeffrey S. Simonoff and Brian D. Marx, *The Sage Handbook of Multilevel Modeling*. Thousand Oaks: Sage. pp. 3–20.
- Jain, Sonia and Neal, Radford M. (2004) A Split-Merge Markov Chain Monte Carlo Procedure for the Dirichlet Process Mixture Model, *Journal of Computational and Graphical Statistics*, 13(1): 158–182.
- Jeffreys, Harold (1983) *Theory of Probability*. Oxford: Clarendon Press.
- Korwar, Ramesh M. and Hollander, Myles (1973) Contributions to the Theory of Dirichlet Processes, *The Annals of Statistics*, 1(4): 705–711.
- Kyung, Minjung, Gill, Jeff and Casella, George (2010) Estimation in Dirichlet Random Effects Models, *The Annals of Statistics*, 38(2): 979–1009.
- Kyung, Minjung, Gill, Jeff and Casella, George (2011) New Findings from Terrorism Data: Dirichlet Process Random Effects Models for Latent Groups, *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 60(5): 701–721.
- Kyung, Minjung, Gill, Jeff and Casella, George (2012) Sampling Schemes for Generalized Linear Dirichlet Process Random Effects Models. With Discussion and Rejoinder, *Statistical Methods and Applications*, 20(3): 259–290.
- Laplace, Pierre-Simon (1774) Mémoire sur la Probabilité des Causes par le Évènements, *Mémoires de l'Académie Royale des Sciences Présentés par Divers Savans*, 6: 621–656.
- Laplace, Pierre-Simon (1781) Mémoire sur la Probabilités, *Mémoires de l'Académie Royale des Sciences de Paris*, 1778: 227–332.
- Leimkuhler, Benedict and Reich, Sebastian (2005) *Simulating Hamiltonian Dynamics*. Cambridge: Cambridge University Press.
- Liu, Jun S. (1996) Nonparametric Hierarchical Bayes via Sequential Imputations, *The Annals of Statistics*, 24(3): 911–930.
- Lo, Albert Y. (1984) On a Class of Bayesian Nonparametric Estimates: I. Density Estimates, *The Annals of Statistics*, 12(1): 351–357.
- MacEachern, Steven N. and Müller, Peter (1998) Estimating Mixture of Dirichlet Process Models, *Journal of Computational and Graphical Statistics*, 7(2): 223–238.
- Mackenzie, G. Calvin and Light, Paul (1987) *Presidential Appointees, 1964–1984, ICPSR Study 8458*. Ann Arbor, Michigan: Inter-University Consortium for Political and Social Research.
- McCulloch, Charles E., Searle, Shayle R. and Neuhaus, John M. (2008) *Generalized, Linear, and Mixed Models*, 2nd edition. New York: John Wiley & Sons.
- Metropolis, Nicholas, Rosenbluth, Arianna W., Rosenbluth, Marshall N., Teller, Augusta H. and Teller, Edward (1953) Equation of State Calculations by Fast Computing Machines, *Journal of Chemical Physics*, 21(1087): 1087–1091.
- Meyer, Kenneth, Hall, Glen and Offin, Daniel C. (1992) *Introduction to Hamiltonian Dynamical Systems and the N-Body Problem*, 2nd edition. New York: Springer-Verlag.
- Morris, Carl N. (1983) Parametric Empirical Bayes Inference: Theory and Applications, *Journal of the American Statistical Association*, 78(381): 47–55.
- Neal, Radford M. (1993) Probabilistic Inference Using Markov Chain Monte Carlo Methods, *Technical Report CRG-TR-93-1*, Department of Computer Science, University of Toronto.
- Neal, Radford M. (2011) MCMC Using Hamiltonian Dynamics. Edited by: Steve Brooks, Andrew Gelman, Galin Jones and Xiao-Li Meng, *Handbook of Markov Chain Monte Carlo*. Boca Raton: CRC Press. pp. 113–162.
- Neal, Radford M. M. (2003) Slice Sampling, *Annals of Statistics*, 31(3): 705–767.



- O'Hagan, Anthony (1994) *Bayesian Inference*, Volume 2 (Part 2 of Kendall's Advanced Theory of Statistics). London: Edward Arnold.
- Polson, Nicholas (1996) Convergence of Markov Chain Monte Carlo Algorithm. Edited by: James O. Berger, José M. Bernardo, Alexander P. Dawid, Dennis V. Lindley and Adrian F. M. Smith, *Bayesian Statistics 5*. Oxford: Oxford University Press. pp. 297–321.
- Robert, Christian and Casella, George (2004) *Monte Carlo Statistical Methods*, 2nd edition. New York: Springer-Verlag.
- Robert, Christian and Casella George (2011) A Short History of Markov Chain Monte Carlo: Subjective Recollections from Incomplete Data, *Statistical Science*, 26(1): 102–115.
- Robert, Christian and Mengersen, Kerrie L. (1999) Reparameterization Issues in Mixture Estimation and Their Bearings on the Gibbs Sampler, *Computational Statistics and Data Analysis*, 29(3): 325–343.
- Roberts, Gareth O. and Rosenthal, Jeffrey S. (1998) Two Convergence Properties of Hybrid Samplers, *Annals of Applied Probability*, 8(2): 397–407.
- Roberts, Gareth O. and Tweedie, Richard L. (1996) Geometric Convergence and Central Limit Theorems for Multidimensional Hastings and Metropolis Algorithms, *Biometrika*, 83(1): 95–110.
- Sethuraman, Jayaram (1994) A Constructive Definition of Dirichlet Priors, *Statistica Sinica*, 4(2): 639–650.
- Strevens, Michael (2006) The Bayesian Approach to the Philosophy of Science. Edited by: Donald M. Borchert, *Macmillan Encyclopedia of Philosophy*, 2nd edition. Carmel: Pearson. pp. 495–502.
- Teh, Yee W., Jordan, Michael I., Beal, Matthew J. and Blei, David M. (2006). Hierarchical Dirichlet Process, *Journal of the American Statistical Association*, 101(476): 1566–1581.
- Zellner, Arnold (1971) *Introduction to Bayesian Inference in Econometrics*. New York: John Wiley & Sons.



# Laboratory Experimental Methods

Rebecca Morton and Mateo Vásquez-Cortés

This chapter provides an overview of some leading work using experimental methods in political science. Given that experimental methods have become an increasingly important tool for political scientists – research in all subfields of the discipline can use and have used experimental methods – we will narrow the scope of the chapter to explore the design and implementation of laboratory experiments. Nevertheless, we will explore the use of experimental methods to study a wide range of political science questions.

The chapter is divided into three sections. The first section is dedicated to discussing some of the most important methodological issues regarding effective laboratory experiments. The second section of the chapter explores different types of lab experiments; in this section, we explain the differences between laboratory experiments from a political economics perspective and those from a political-psychology perspective, which are two of the main approaches used in laboratory experiments in political science. The last

section expands on new developments in lab experiments: we consider an expansion of the method to new topics like network analysis and emotions, as well as methodological developments such as lab-in-the-field experiments and virtual labs.

## **EXPERIMENTAL POLITICAL SCIENCE**

Experimental research is at the core of many scientific disciplines. In some cases, the findings and norms are hard to grasp without reference to experimental methods. In a growing number of areas, experiments are now a common and important element of the political scientist's toolkit. Especially in recent years, there has been a large expansion in the number of political scientists who view experiments as useful and informative. Political scientists across subfields now increasingly use one or more of the three major experimental methods: laboratory, survey, and field.<sup>1</sup>

Experiments appeal to political science because of their potential to generate stark and powerful causal claims and their ability to inform theoretically grounded empirical claims (Druckman et al., 2006; Morton and Williams, 2010). Experiments facilitate inference through two mechanisms: control of observable confounding factors and random assignment to treatments. The procedure of random assignment ensures that unobservable factors are equally likely to be present across treatment and baseline groups. Through these two methods, experiments are particularly useful in identifying the causal effects of variables of interest as discussed by Bowers and Leavitt (Chapter 41, this *Handbook*).

Additionally, experiments are an important tool in guiding theoretical advancements. Experiments influence theoretical developments as they provide means for studying preferences over institutions, political behavior, and other fundamental factors that might be hard to assess using non-experimental data. Finally, experiments facilitate causal inference and empirical tests of theories through the transparency and reproducibility of their procedures.

In what follows, we focus on the second aspect: the ability of experimental research to guide theoretical development. Given that this *Handbook* dedicates other chapters to field and survey experiments, we highlight the benefits and challenges of laboratory experiments as a tool for political scientists.

### ***Laboratory Experiments in Political Science***

Laboratory experiments provide many advantages. We want to emphasize four: control conditions, replication, multiple manipulations, and measurement of variables of interest.<sup>2</sup>

First, in the laboratory, it is relatively easy to measure subjects' behavioral differences when manipulations are compared and

subjects are randomly assigned to manipulations. In other experimental settings, like the field and Internet, the researcher often cannot correct for the lack of control of the environment in which manipulations take place and has to rely primarily on random assignment. If the environments in which the different manipulations take place vary systematically and this is not observed by the researcher, then the comparison between manipulations is more difficult to determine.<sup>3</sup>

Second, as research transparency becomes more of a norm within political science, so does the ability to explore previous findings through scientific replication. Because the reproducibility of experiments is an essential part of the scientific method, the inability to replicate the studies of other types of experiments has potentially grave consequences. By contrast, replication of laboratory experiments is often feasible. Since replication has become a major element in scientific advancement in political science, laboratory experiments provide one of the most accessible settings for replication, which allows researchers to evaluate the validity and scope conditions of their proposed causal relationships. Laboratory conditions are similar across different environments, and experimenters can replicate studies keeping most relevant factors equal. No other experimental setting provides such good conditions for replication as laboratory experiments. Most think of the standard student-subject pool used in laboratory experiments as a disadvantage since it is not generally representative of nonstudents in the area, the fact that it is similar and constantly renewing facilitates the scientific replication of previous results in different times and locations.

A third advantage of laboratory experiments is that the researcher can create rich environments that do not often exist in naturally occurring situations. For example, in the laboratory, the researcher can test different voting mechanisms and evaluate the potential benefits of approval voting as opposed to majority rule, holding the choices that are

before voters constant. It would be quite difficult to convince jurisdictions to randomly adopt both of these procedures simultaneously or find jurisdictions facing the same choices but randomly assign their voting rules. Recently, for example, Bassi (2015) and Bouton et al. (2016) explored the benefits of different voting mechanisms, using laboratory experiments.<sup>4</sup>

Fourth, the facilities of the laboratory allow for the measurement of brain activity and emotional responses, which are as important for different research questions as they are hard to measure without adequate equipment. Political scientists use fMRI equipment and skin-conductance measures to gather information that is relevant for political decisions, like partisanship and preferences for social welfare. For instance, Westen et al. (2006) and Amodio et al. (2007) measure neural responses to study partisan political judgments, and Aarøe and Petersen (2013) measure blood glucose levels to study preferences for social-welfare policies. These variables are difficult to obtain outside a laboratory.<sup>5</sup>

Finally, laboratory experiments are a powerful tool for observing behavior in cases where theory makes unclear predictions or does not make predictions at all. For example, laboratory experiments have been used to select over multiplicity of equilibria in theoretical settings. We can experimentally test if players are sufficiently patient for the amount of cooperation that can be sustained in equilibrium.<sup>6</sup> To do so, the researcher can create situations in the laboratory in which players repeatedly interact for an indefinite period of time by using random procedures to determine when the game will end. By varying the probability that the game will end, the researcher can manipulate the benefits of patience in order to determine the extent that subjects demonstrate patience.

Laboratory experiments usually follow a distinct pattern. After formulating a research question, the researcher chooses a design that answers the question based on the treatment variable(s), variation between subjects (when

we examine the differences between individuals) vs variation within subjects (when we examine the variability of a particular value for individuals in a sample), and the required number of independent observations. This design stage of an experiment is probably the most important step in experimental research: the better the design, the cleaner the results from the experiment and the less need for complex post-experiment statistical analysis in order to determine what we have learned from the study. If the study addresses a research question that is based on a formal theoretical model and wishes to evaluate the model, then ideally the experiment is designed to eliminate as many disconnects as possible between the theory's assumptions concerning the environment and the situation facing the subjects, so that the results of the experiment directly speak to the theory.

On the practical side, once the design is formulated, a researcher usually engages in three activities to prepare an experiment: 1) the writing of clear instructions for each of the treatments; 2) the preparation of a 'script' of a session; and 3) the creation of a computer program to measure the variables of interest.

The researcher also has to secure monetary funds to pay subjects when needed. The funds are usually used to cover participation fees and the cost of laboratory time, as well as computer programming in some cases. In the next section, we will explain how monetary incentives affect the experimental design.

Additionally, the researcher has to secure approval from ethical boards. Universities and laboratories have clear protocols that must be satisfied in every study that involves human subjects. The research has to make clear how the proposed study follows the rules of the protocols. Many journals now verify the commitment to ethical considerations before publishing any study.

Finally, when all the previous conditions have been satisfied, the researcher moves on to recruit subjects, pilot the activities, improve the design or instructions when needed, and run the experiment.<sup>7</sup>

Conducting a laboratory experiment from start to finish is a demanding activity prone to many practical restrictions.<sup>8</sup> Additionally, there are a number of criticisms of laboratory experiments. A usual point of concern is the generalization of the sample of experimental participants to a larger population of interest. The validity of laboratory studies has inspired some skepticism.<sup>9</sup> Before discussing these concerns, we highlight some key differences between internal validity and external validity, so we can better comment on the potential limitations of laboratory experiments.<sup>10</sup>

Internal validity is an evaluation about the truth of the inference or knowledge claim within a studied target population. Internal validity has three components: statistical (how valid are the statistical relationships observed and is the sample a random selection from the target population studied); causal (can we identify causal inferences in the sample); and construct (how much does the experiment truly test the theoretical research question addressed). In contrast, external validity evaluates the knowledge claim for observations beyond the target population studied. It has to do with whether we can generalize the results beyond the target population studied in the experiment. A third type of validity that is often confused with external validity is ecological validity, which has to do with the degree to which the environment abstracts from and is at variance with the naturally occurring world.

To understand why ecological validity is different from external validity, consider a simple voting experiment. We can make the voting experiment distinct from reality by having subjects engage in choices where instead of casting votes, they take tokens from different token bins where each subject can only take one token. The bin from which the most tokens are taken is declared the winner. Or, we can make the experiment close to reality by having the choices before the voters be labeled candidates or issues that are real in naturally occurring elections in their jurisdiction and using the actual voting equipment used in their local jurisdiction.

Importantly, the choice of whether to use the less ‘real’ tokens vs the more ‘real’ actual candidates does not affect the external validity of the study, since neither choice affects whether the experimental results generalize to a voting jurisdiction in which there are completely different choices before voting and often distinctively different voting procedures or equipment. Both environments can be equally ‘unreal’ to the external jurisdiction. In both environments, real humans are making real choices that can affect real earnings. Some may even argue that the experiment with tokens may be more ‘generalizable’, since it does not have the baggage of the local jurisdiction and is more neutral to other environments.<sup>11</sup>

Ultimately, the external validity of laboratory experiments is something that needs to be addressed empirically, as does the external validity of any empirical study outside the laboratory. All empirical research can only make inferences for the pool of subjects from which it is drawn. To make inferences beyond that subject pool, more experiments must be conducted. When a study is theoretically founded, the theory can be a guide to these subsequent studies. It is an empirical question that can only be answered empirically; it is not something that can be addressed through argumentation or theory alone or be justified based on the design of a single experiment with one target population.

This does not mean that the question of external validity is a minor one or should be ignored. The results of experiments that rely too much on a single target population (such as college students from westernized, educated, industrialized rich democracies (WEIRD)) may not generalize to other populations of interest, and it is a vital question as to whether the results generalize to nonstudents and subjects drawn from non-WEIRD populations. As the development of laboratories in countries such as Kenya (Nairobi Busara Laboratory), India (Nuffield Oxford Centre of Experimental Social Sciences at FLAME University in Pune), and the United

Arab Emirates (Social Science Experimental Laboratory at New York University Abu Dhabi) and ‘lab-in-the-field’ experiments (discussed below) expand, external-validity issues can be addressed.<sup>12</sup>

## **TYPES OF LABORATORY EXPERIMENTS IN POLITICAL SCIENCE**

Laboratory experiments in political science are based on two different and distinct heritages: psychology and economics.<sup>13</sup> Following other authors (Dickson, 2011; Morton and Williams, 2010), we focus on a number of important dimensions in which these two types of experiments differ: stylization, incentives, level of analysis, the role of theory, the use of repetition of tasks, and deception. These differences are largely due to some of the differences in research focuses and types of questions that experimentalists from the two heritages emphasize, as we will discuss below.

Laboratory research in economics tends to give an abstract description of the situation of interest. The roles and decisions that subjects make in the laboratory are usually labeled in neutral terminology. For example, Fehr and Gächter (2000) use neutral language and do not mention fairness, punishment, or other potentially charged terms when studying punishment in games of public-goods provision. In the same vein, in their study about voter turnout, Levine and Palfrey (2007) use labels *X* and *Y* instead of ‘vote’ or ‘abstain’. The main reason for using neutral language is to keep experimental control and avoid reactions from the subjects that the analysis cannot properly measure. Such control is deemed necessary to increase the internal validity of the experiment as an evaluation of economic theory. According to this argument, context-rich settings tend to have many factors that could affect subjects’ attention, influence their behavior, and potentially affect the outcomes of interest. That said, experimentalists in the economics tradition also often use non-neutral

language as well. More importantly, they sometimes explore the impact of context in experiments. For example, Aragonés and Palfrey (2003) conduct their experiment using both neutral language and language that is non-neutral and refers to the choices in their political terms – they find that the behavior is unaffected by the terminology used. The point is that such context might matter and should be something that the experimentalist investigates and studies; it is in itself a treatment.

By contrast, experiments in the psychology tradition try to evoke more contextually rich environments and put more emphasis on the descriptive realism of laboratory scenarios. For example, Levendusky (2013) uses a series of experiments to study the effect of the media on polarization. The study provides subjects with recent news capsules from political TV shows and realistic newspapers editorials. In another study, Brader (2005) recreates political advertisements using genuine ads from a naturally occurring, ongoing campaign. Making the ads as credible as possible was crucial for the study as the psychological mechanisms that the author explores are rooted in the experimental protocol.

Abstract experimental designs have advantages and disadvantages. Stylization can provide a higher degree of control and allow for a clearer definition of theoretical benchmarks that has to pass a hard test imposed by the neutral design. Experiments in this tradition, however, often face skepticism about the ecological validity of their inferences from political scientists who come from the psychology tradition.

Experimental economics usually offers subjects monetary incentives that depend on subjects’ decisions in the laboratory. Monetary incentives can be of much use when researchers want to reward accuracy and induce preferences over the choices that are before the subjects, as assumed in a theoretical model. In contrast, psychological experiments tend not to offer inducements that are conditional to subjects’ actions and give fixed amounts of cash payments or course credit.

In many cases, the research question of interest to political psychologists does not seem to straightforwardly yield itself to using monetary incentives. For example, at first blush it would seem that in the studies of Levendusky (2013) and Brader (2005), political attitudes are the relevant dependent variable, and offering monetary incentives to report one opinion over the other would appear not to be useful. That said, some studies are beginning to incorporate incentives in similar experiments by having subjects express their political attitudes though making monetary choices over contributions to interest groups, as in Haas and Morton (2018).

Monetary incentives can lead to better task performance, as shown by Camerer and Hogarth (1999), and more accurate responses, as demonstrated by Prior and Lupia (2008) in a survey experiment in which monetary rewards incentivized more precise answers to political knowledge questions. However, many psychological studies, such as Ryan and Deci (2000), have found that financial incentives lower task performance by crowding out intrinsic motivations.<sup>14</sup> More nuanced studies of the effects of incentives on task performance demonstrate that the effect appears to be heavily dependent on the size of the incentives and the task involved. Paying subjects small amounts of money, as in Ryan and Deci (2000), can have a negative effect. But as the size of the payment increases, the effect becomes positive. If the experimental game is complex, further increases beyond even standard payments can also have a positive effect.<sup>15</sup>

In game theory, the preferences of an individual are part of the game primitives. Therefore, in a study that is rooted in a game-theoretical model, the researcher wishes to induce in participants the same preferences that are assumed in the theory. In laboratory experiments, this is usually done through the use of monetary incentives. By controlling for the preferences using monetary incentives, the researcher can focus on other aspects of the study, such as consistency with theoretical equilibrium.<sup>16</sup>

Laboratory experiments in economics and psychology also differ in the level of analysis and the role of formal theory. In a typical economic experiment, subjects are participating in a game of interaction with each other, and the emphasis is typically in comparing subject behavior in such interactive situations to formally derived game-theoretic equilibrium predictions. Since in most economics style experiments, the researcher is interested in measuring behavior in a steady state, after learning about the environment and the game has subsided, subjects usually play the game repeatedly and the researchers often focus on analyzing behavior in the later periods. In psychological experiments, subjects are usually engaging in an individualized decision-theoretic task that they complete only once without interaction with others, and the goal is to compare their behavior with the predictions drawn from a non-formal psychological theory of behavior.

A final point in which economics-style and psychology-style laboratory incentives differ is the acceptance and use of deception. The consensus among experimental economists is to avoid most deception, while in the political-psychology research tradition, deception is seen as a useful tool in experimentation. There are potential advantages and disadvantages of using deception, which are not discussed here due to space limitations.<sup>17</sup>

The multifaceted nature of our subject of study, along with the varying strengths and weaknesses of different research methods, highlights the advantages of experimental political science. In what follows, we discuss two of the most promising directions for experiments in political science using laboratory settings: emotions and network analysis.

## NEW DEVELOPMENTS: TOPICS

As experimentation has increased in political science, many new avenues have been taken.

In this section, we focus on just two of these: analyses of social networks and emotions.

### **Network Analysis**

The study of social networks in political science is a fast-growing topic that has drawn a lot of interest in the discipline. Empirical research on social networks has lagged behind theory, however, and we still lack empirical validation for the large number of theoretical findings accumulated in the last decade. The main reason for this lag is related to the challenges involved with causal identification of the impact of network structure on behavior.

First of all, observational data on social networks is usually unavailable and incomplete, and it is often hard to make inferences based on incomplete data. This *Handbook* contains current methods to address data collection of relation data (Calvo et al., Chapter 30, this *Handbook*), network theory (Victor and Khwaja, Chapter 45, this *Handbook*), and estimation methods in networks (Schoeneman and Desmarais, Chapter 46, this *Handbook*). Second, even if we have complete data, it is hard to impossible to find exogenous factors able to affect network structure that are independent from outcomes of interest, generating severe endogeneity issues: when links in a network are created based on characteristics that are also correlated with what is being measured, it is hard to know whether a correlation in behavior across subjects connected in the network is the result of a social interaction effect or simply from the selection into the network.

Experimental research of social networks is growing in social science because experiments allow the causal identification of network-structure effects.<sup>18</sup> Moreover, experiments provide an invaluable tool to validate existing theoretical findings in general and how individuals use network information in particular. In what follows, we highlight the uses of laboratory experiments on

social networks that are particularly relevant in political science: network formation and information transmission.

As mentioned before, there are several challenges to identification in social networks with observational studies. This is particularly true of studies that look at network formation: since social network data encodes interactions in real life, it is often extremely hard – if not impossible – to find sources of exogenous variation in network structure. Laboratory experimentation on social networks is a particularly important tool as it can easily manipulate not only network structure (which can be chosen by the researcher) but also the formation of networks (manipulate incentives to create links with other participants). Given that observational studies have to rely on the empirically difficult problem of selection into the network, laboratory experiments provide a tool to study network formation that has several advantages over other methods. A host of experiments have successfully been used to study network formation (Caldara et al., 2014; Callander and Plott, 2005; Carrillo and Gaduh, 2012; Goeree et al. 2009).

Most of these papers study the costly decision to form links with other subjects in a context in which being connected is valuable. Studies vary the costs of forming links across agents and the extent to which links can be created unilaterally. Some of the results of these experiments can be exported outside the lab. For example, Goeree et al. (2009) argue that agents' heterogeneity in terms of linking costs is a major determinant of the predominance of star-like structures in real-life social networks, that is, structures in which a single agent is directly connected to all other peripheral agents.

Another avenue of research in laboratory experiments on social networks is on the topic of social learning. Opinions about new technologies, job opportunities, and political candidates are largely transmitted through social networks. Several studies have experimentally studied the transmission of information



in networks (see for example Çelen et al., 2010; Chandrasekhar et al., 2015; Choi et al., 2012) to understand how agents filter and use information they receive from neighbors.

All the studies mentioned above follow the economic tradition in laboratory experiments. For example, Chandrasekhar et al. (2015) study different models of social learning in networks where agents can update their guess about the state of the world according to Bayesian learning (sophisticated learners) or simply following the guesses of others (naïve learners).

Chandrasekhar et al. (2015) study different models of social learning in networks. After observing a signal, subjects guess the binary state of the world. In subsequent rounds, agents observe their network neighbors' previous guesses before guessing again. The models of learning that they study separates sophisticated from naïve learners based on the use of the information agents receive from their network. The lab sessions were conducted in Indian villages and in a high-ranked Mexican university. The laboratory experimental analysis abstracts from other important unobserved determinants of learning behavior. As the researchers explain, the laboratory experiment allows the researchers to prioritize studying fundamentals of learning behavior over analyzing the effect of confounders of such fundamentals.

The possibility of using exogenous network change to study causal links between network structures and other outcomes is exciting. Such an inquiry requires that the underlying change to the network is not directly correlated with the outcome of interest. Laboratory experiments facilitate the manipulation of network formation and structure to better understand causal relations. In other words, the network structure can be manipulated by the researcher while controlling other variables. For example, when studying coordination, Cassar (2007) keeps constant elements like number of players and number of connections and manipulates the structure of the network. Similarly,

Choi et al. (2011) vary network structure in dynamic public-goods game. In any other application, the network structure can be exogenously imposed by the experimenter.

The number of possible implementations of network structure and learning in networks is also a growing and exciting area of research. The described studies provide the fundamental tools that can be used to apply the laboratory methodology to politically relevant questions.

### **Emotions**

Increasingly, social scientists are using experiments to study the role of emotions in a wide variety of political phenomena. Studies have looked at the impact of emotions such as fear, anxiety, anger, and enthusiasm (for a review, see Valentino et al., 2011). For obvious reasons, most of these studies depart from psychology-style experiments and study questions related to voting, framing effects, and public opinion. However, there is a growing trend in the literature that applies known psychological findings into models of political economy. This research agenda departs from the rational-choice framework and includes psychological insights such as the effect of emotions.

Recently, Acharya et al. (2018), Lupia and Menning (2009), Minozzi (2013), Aldama et al. (2018), and Little and Zeitzoff (2017) incorporated ideas from psychology into political-economy models. Accordingly, recent research tries to incorporate psychological elements into what traditionally was economics-style laboratory experimental political science, which has been the testing ground for political-economy theories.

Myers and Tingley (2016) study the effect of a set of emotions on trust. The authors manipulate emotions using a common psychological method: the Autobiographical Emotional Memory Task (AEMT).<sup>19</sup> The method asks subjects to recall a wide variety of emotions and focuses them on the emotion

that the researcher wants them to experience. Then, they measure trust using the common economic-based trust game. Similarly, Hopfensitz and Reuben (2009) focus on the effect of emotions for social punishment. The authors highlight the role of emotions for cooperative behavior and propose a game-theoretical framework that captures how anger may motivate subjects to punish free riders. The measure of emotions comes from self-reported data.

Studying the effect of emotions on political outcomes is not new.<sup>20</sup> What is novel in these approaches is the combination of methods from economic-style and psychology-style laboratory experiments. These studies rely on laboratory applications of formal theory and psychology that are common in political science. As noted by Albertson and Gadarian (2016), there are two main challenges: manipulating emotions and isolating the impact of the emotions of interests (say anxiety), from other emotions that might be affected by the same manipulation (like fear).

We expect that the intersection of these two types of experiments, along with the adoption of technologies to measure emotions and neural responses in the lab, will be one of the avenues for future research in laboratory experiments in political science. The advancement of experiments that take elements from both traditions is also likely to change methodological debates presented in the previous section.

## NEW DEVELOPMENTS

In 2012, Druckman identified the three most prominent experimental methods in political science: laboratory, surveys, and field. Since then, the reduced cost and increased feasibility of digital field experiments has led to an increase in experimentation using different platforms. We focus here on two recent advancements that relate to laboratory experiments: lab-in-the-field and virtual labs.

A lab-in-the-field experiment takes place when subjects participate in a common physical location but the experimenter, to some degree, brings the laboratory to the subjects' natural environment (Morton and Williams, 2010). Lab-in-the-field experiments are now often used in political science (for recent studies, see Viceisza, 2016, and Gneezy and Imas, 2017). Taking the lab experiment from the university into different habitats of social interaction allows for the comparison of the lab results to those with different populations, while keeping constant other important aspects of the experiment. There are two important insights derived from this point: first, lab experiments in the field can be used as robustness checks of the results obtained in the lab; second, they can provide information about behavior across cultures and contexts, which overcomes common criticisms about the validity of laboratory experiments, as discussed earlier.

In this section, we are going to highlight how lab-in-the-field experiments can be used to complement traditional field experiments by providing reliable measures of variables of interests and to test theoretical models with new populations. These reasons contrast with the contextual advantages in which lab-in-the-field developed: usually, a lab-in-the-field experiment was seen as an appropriate methodological tool when researchers wish to conduct a traditional laboratory experiment and take advantage of a particular naturally occurring situation in the field.<sup>21</sup>

Lab-in-the-field experiments can be used for all of the same purposes as standard laboratory experiments. Moreover, lab-in-the-field experiments can be combined with standard field experiments to unpack mechanisms behind causal relations identified by field experiments (Gneezy and Imas, 2017). As such, they have been used to elicit, measure, and identify parameters associated with characteristics such as preferences, beliefs, and social norms. In particular, when there are reasons to expect that risk, time, and ambiguity preferences play an

important role, scholars rely on laboratory procedures to elicit and estimate those factors. Additionally, lab-in-the-field techniques provide incentive-compatible mechanisms through which subjects credibly rely on their social norms (preferences, trust, and reciprocity). This helps overcome difficulties in answering abstract questions and the social desirability biases, which make self-reports problematic.<sup>22</sup>

There are a number of noteworthy recent lab-in-the-field experiments on pro-social behavior and cooperation outside of the traditional laboratory. For example, Baldassarri and Grossman (2013) study whether group attachment and position in social networks affects pro-social behavior within in-groups. The study took place in rural communities in Uganda and used different variants of the dictator game to show that group attachment positively affects pro-social behavior. Also interested in pro-social norms, Gilligan et al. (2014) unpack the effects of wartime violence on social cohesion. Using public-good contributions, trust games, and altruistic giving, the authors measure pro-social motivation among communities affected by Nepal's ten-year civil war. Finally, Alexander and Christia (2011) look at the effect of ethnic diversity on cooperation with a sample of students from groups affected by conflict (Croats and Bosnians).<sup>23</sup> All three studies are examples of a lab-in-the-field approach that take full advantage of the methodology in that they examine an understudied population while using known laboratory methods to measure behavior.

An important implication of validating theories across populations and using laboratory methodologies to measure otherwise difficult to elicit variables is that researchers rely on standardized validated methods. This point is important because it ensures that results are comparable and replicable with known laboratory studies and also across other contexts. However, there are costs to using methods that may not be best simply because they were used previously. Developing new

strategies with students can be validated through several experiments, whereas in the field, new techniques may be constructed of other factors. It is important to be aware of cultural differences when developing the design of a lab-in-the-field study. For example, the population of interest in some of the studies mentioned above can exhibit large variation in levels of literacy and mathematical ability, which will limit the use of more complex elicitation methods.

Another advantage of using lab-in-the-field techniques is to test theories with a relevant population. Experiments that test theories in the lab typically use a convenience sample of undergraduate students, and sometimes it is assumed that behavior in the lab can be generalized to the relevant population. While most theories of human behavior are not specific to particular populations (for example, issues of pro-social behavior are important for all humans), some theories are particularly relevant to a specific subset of humans. That is, suppose that the theory concerns the behavior of lobbyists; a lab-in-the-field experiment conducted with actual lobbyists may seem a better test of the theory. Indeed, Potters and Van Winden (1996, 2000) conducted such an experiment with both lobbyists and students.<sup>24</sup> When using lab-in-the-field techniques, the researcher can test a theory with the relevant population, while having the advantage of studying behavior in a controlled setting. The lab-in-the-field evidence allows the researcher to isolate variations in individuals' motivations vs other types of effects, given that it is often difficult for some populations to participate in research activities in standard laboratory environments.

Virtual-lab experiments are conducted similarly to laboratory experiments, but the interactions between the experimentalists and the subjects and with each other takes place online or virtually. Most online survey experiments are essentially psychological virtual-lab experiments. In contrast, online economics experiments in which subjects

make interactive choices in a game situation are more distinctive and more complex to design, since it is often difficult to have subjects online at the same time as one another for a sufficient time period in order to participate in online simultaneous games. Hence, there can be less opportunity for subjects to engage in learning and repetition here than in the laboratory. Moreover, subjects may be less likely to believe that they are participating in a game with others when they are doing so virtually and cannot observe that there are others in the game, as they would be able to in a standard laboratory experiment. For a discussion of some of the issues involved in designing and implementing virtual-laboratory experiments, see Arechar et al. (2018).

## CONCLUSION

In this chapter, we have considered some elements of laboratory experiments in political science. After a description of the basic elements of this important research toolkit, we discussed the two heritages from which laboratory experiments in political science arise and how they differ on stylization and use of incentives. We then introduced two new developments that follow from these traditions: laboratory experiments with social networks and laboratory experiments with emotions. When relevant, we have shown how these new developments take elements from the traditional laboratory experiments. Finally, we have provided a brief introduction to lab-in-the-field experiments and also stressed their ability to complement the findings from traditional laboratory experiments.

Good experiments should be analyzed in light of theoretical motivations and be complemented with adequate econometric method. We believe that while the adoption of new technologies allows researchers to use different platforms to conduct randomized experiments, the laboratory methodology still has several advantages that make it the

most appropriate method to study several topics. In particular, using carefully designed laboratory methodology can increase our understanding of theory-oriented inquiries, provide information about the mechanisms behind behavior, and the effectiveness of potentially larger interventions.

As technology advances provide useful platforms to develop experiments, the laboratory setting not only remains the most advantageous method to study several topics but also provides the most effective tools to address several academic purposes.

While we recognize and celebrate the use of different platforms and methods to develop experimental research in political science, we should not overestimate their advantages. The laboratory enables experimentalists to have a degree of control over the experiment that is hardly achievable in any other setting. Similarly, other factors such as compliance and bundled treatments, which can be problematic in other types of experiments, are less of a problem in laboratory experiments. Finally, given the possibility to control the experimental conditions, factors such as the context, the location, and the time can be manipulated in the laboratory in a way that cannot be controlled in the field, for example.

## Notes

- 1 A proof of the growing expansion of experimental methods in the discipline is the new *Journal of Experimental Political Science*, JEPS, which exclusively publishes experimental research.
- 2 Some of these reasons have been explained in Morton and Williams (2010).
- 3 The level of control applies to individual and group experiments as well.
- 4 Additionally, researchers can use experiments as a complement to theory in order to determine the effect of different forces on complex situations where theory is impractical. For example, in economics, laboratory experiments have been used to test the performance of different auctions in order to sell spectrum rights. Although these experiments were grounded in theory, the results provide new insights for the development of new theoretical studies.

- 5 Some topics that are particularly important for political science, like deliberation, require subjects to be physically in the same space. See Karpowitz et al. (2012) for an example in which subjects debate under different group compositions.
- 6 Dal Bó and Fréchet (2011) study the conditions under which cooperation can be sustained in repeated games.
- 7 The analysis of results from laboratory experiments is beyond the scope of this chapter, but we encourage the reader to look at 'The Experimentalist's To Do List' section in Morton and Williams (2010).
- 8 This is not to say that other empirical methods are less demanding in terms of time and money.
- 9 See, for instance, the widely cited Sears (1986).
- 10 A full discussion can be found in Druckman et al. (2006) and Morton and Williams (2010).
- 11 Ecological validity is more likely to relate to internal validity, since high ecological validity arguably may increase the salience of the experiment to subjects and thus to results that are more statistically valid by reducing the variance in subjects' choices.
- 12 For a full discussion of the participation of students in laboratory experiments in politics see Druckman and Kam (2011).
- 13 There is a third heritage in experimental political science, statistical methods, but experimentalists who come from this tradition tend to be field experimentalists, not laboratory experimentalists. See Morton and Williams (2010).
- 14 Intrinsic motivation involves engaging in a behavior because it is inherently enjoyable and interesting, while extrinsic motivation refers to doing something because it led to separable outcomes.
- 15 Economists worry that subjects will not believe an experimenter if they expect that deception might be used based on past experience in the laboratory or the laboratory's reputation among other subjects. Such disbelief can lead to greater variance in behavior. That said, economists do not typically reveal to subjects the exact purpose of an experiment but usually use vague descriptions such as describing an experiment as one on decision-making. See the review in Morton and Williams (2010) and also Bassi et al. (2011).
- 16 The importance of inducing preferences was posited by Vernon Smith (1976) in what is known as induced value theory.
- 17 Chapter 13 in Morton and Williams (2010) has a comprehensive review.
- 18 For a recent review, see Choi et al. (2016).
- 19 The AEMT is not the only method to induce emotions in the lab. There are other methods such as having the subject watch movie clips and pictures and interact with confederates.
- 20 See Albertson and Gadarian (2016) for a recent review.
- 21 That is, for example, the motivation in Whitt and Wilson's (2007) study of Katrina victims.
- 22 Additionally, Gneezy and Imas (2017) note that the data generated using laboratory techniques in the field can help identify participants that will benefit from future potential intervention. Obviously, these variables may be relevant for different reasons depending on the research questions.
- 23 Lab-in-the-field experiments are also an important tool if researchers are interested in comparing results between different contexts. If there are theoretical reasons to expect that cultural differences will be important for an outcome of interests, running the same experiments in different contexts can potentially provide information of the relevance of the contexts. Research on the difference of the willingness to cooperate (Gächter et al., 2010) and engage in anti-social behavior (Herrmann et al., 2008) uses similar implementations in different populations.
- 24 See Potters and Van Winden (2000) for an example of an experiment conducted with both lobbyists and undergraduate students.

## REFERENCES

- Aarøe, L. and M. B. Petersen (2013). Hunger games: Fluctuations in blood glucose levels influence support for social welfare. *Psychological Science* 24 (12), 2550–2556.
- Acharya, A., M. Blackwell and M. Sen (2018). Explaining preferences from behavior: A cognitive dissonance approach. *The Journal of Politics* 80 (2), 400–411.
- Albertson, B. and S. K. Gadarian (2016). Did that scare you? Tips on creating emotion in experimental subjects. *Political Analysis* 24 (4), 485–491.
- Aldama, A., M. Vásquez-Cortés and L. Young (2018). Fear and citizen coordination against dictatorship. *Journal of Theoretical Politics* 31 (1), 103–125.
- Alexander, M. and F. Christia (2011). Context modularity of human altruism. *Science* 334 (6061), 1392–1394.
- Amodio, D. M., J. T. Jost, S. L. Master and C. M. Yee (2007). Neurocognitive correlates of liberalism and conservatism. *Nature Neuroscience* 10 (10), 1246–1247.

- Aragones, E. and T. Palfrey (2004). The effect of candidate quality on electoral equilibrium: An Experimental study. *American Political Science Review* 89 (1), 77–90.
- Arechar, A. A., S. Gächter and L. Molleman (2018). Conducting interactive experiments online. *Experimental Economics* 21 (1), 99–131.
- Baldassarri, D. and G. Grossman (2013). The effect of group attachment and social position on prosocial behavior: Evidence from lab-in-the-field experiments. *PLoS One* 8 (3), e58750.
- Bassi, A. (2015). Voting systems and strategic manipulation: An experimental study. *Journal of Theoretical Politics* 27 (1), 58–85.
- Bouton, L., M. Castanheira and A. Llorente-Saguer (2016). Divided majority and information aggregation: Theory and experiment. *Journal of Public Economics* 134, 114–128.
- Brader, T. (2005). Striking a responsive chord: How political ads motivate and persuade voters by appealing to emotions. *American Journal of Political Science* 49 (2), 388–405.
- Caldara, M., and M. McBride, (2014). *An experimental study of network formation with limited observation*. Unpublished Manuscript.
- Callander, S. and C. R. Plott (2005). Principles of network development and evolution: An experimental study. *Journal of Public Economics* 89 (8), 1469–1495.
- Camerer, C. F. and R. M. Hogarth (1999). The effects of financial incentives in experiments: A review and capital-labor-production framework. *Journal of Risk and Uncertainty* 19 (1–3), 7–42.
- Carrillo, J. and A. Gaduh (2012). *The strategic formation of networks: Experimental evidence*. Unpublished Manuscript.
- Cassar, A. (2007). Coordination and cooperation in local, random and small world networks: Experimental evidence. *Games and Economic Behavior* 58 (2), 209–230.
- Çelen, B., S. Kariv and A. Schotter (2010). An experimental test of advice and social learning. *Management Science* 56 (10), 1687–1701.
- Chandrasekhar, A. G., H. Larreguy and J. P. Xandri (2015). Testing models of social learning on networks: Evidence from a lab experiment in the field. *National Bureau of Economic Research, NBER Working Paper* No. 21468.
- Choi, S., D. Gale and S. Kariv (2012). Social learning in networks: A quantal response equilibrium analysis of experimental data. *Review of Economic Design* 16 (2–3), 135–157.
- Choi, S., D. Gale, S. Kariv and T. Palfrey (2011). Network architecture, salience and coordination. *Games and Economic Behavior* 73(1), 76–90.
- Choi, S., S. Kariv and E. Gallo (2016). Networks in the laboratory. *The Oxford Handbook of the Economics of Networks*. Oxford University Press, New York, 440–478
- Dal Bó, P. and G. R. Fréchet (2011). The evolution of cooperation in infinitely repeated games: Experimental evidence. *American Economic Review* 101 (1), 411–429.
- Dickson, E. (2011). Economics vs. psychology experiments. *Cambridge Handbook of Experimental Political Science*. Cambridge University Press, New York, 58–69
- Druckman, J. N. (2012). Experimenting with politics. *Science* 1207808 (335). 1177–1179.
- Druckman, J. N., D. P. Green, J. H. Kuklinski and A. Lupia (2006). The growth and development of experimental research in political science. *American Political Science Review* 100 (4), 627–635.
- Druckman, J. N. and C. D. Kam (2011). Students as experimental participants: A Defense of the 'Narrow Data Base'. *Cambridge Handbook of Experimental Political Science*, vol. 1. Cambridge University Press, New York, 41–57.
- Fehr, E. and S. Gächter (2000). Cooperation and punishment in public goods experiments. *American Economic Review* 90 (4), 980–994.
- Gächter, S., B. Herrmann and C. Thöni (2010). Culture and cooperation. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 365 (1553), 2651–2661.
- Gilligan, M. J., B. J. Pasquale and C. Samii (2014). Civil war and social cohesion: Lab-in-the-field evidence from Nepal. *American Journal of Political Science* 58 (3), 604–619.
- Gneezy, U. and A. Imas (2017). Lab-in-the-field: Measuring preferences in the wild. *Handbook of Economic Field Experiments*, vol. 1. Elsevier, Amsterdam, 439–464.
- Goeree, J. K., A. Riedl and A. Ule (2009). In search of stars: Network formation among

- heterogeneous agents. *Games and Economic Behavior* 67 (2), 445–466.
- Haas, N. and R. B. Morton (2018). Saying versus doing: A new donation method for measuring ideal points. *Public Choice* 176 (1), 79–106.
- Herrmann, B., C. Thöni and S. Gächter (2008). Antisocial punishment across societies. *Science* 319 (5868), 1362–1367.
- Hopfensitz, A. and E. Reuben (2009). The importance of emotions for the effectiveness of social punishment. *The Economic Journal* 119 (540), 1534–1559.
- Karpowitz, C. F., T. Mendelberg and L. Shaker (2012). Gender inequality in deliberative participation. *American Political Science Review* 106 (3), 533–547.
- Levendusky, M. S. (2013). Why do partisan media polarize viewers? *American Journal of Political Science* 57 (3), 611–623.
- Levine, D. K. and T. R. Palfrey (2007). The paradox of voter participation? A laboratory study. *American Political Science Review* 101 (1), 143–158.
- Little, A. T. and T. Zeitzoff (2017). A bargaining theory of conflict with evolutionary preferences. *International Organization* 71 (3), 523–557.
- Lupia, A. and J. O. Menning (2009). When can politicians scare citizens into supporting bad policies? *American Journal of Political Science* 53 (1), 90–106.
- Minozzi, W. (2013). Endogenous beliefs in models of politics. *American Journal of Political Science* 57 (3), 566–581.
- Morton, R. B. and K. C. Williams (2010). *Experimental Political Science and the Study of Causality: From Nature to the Lab*. New York. Cambridge University Press.
- Myers, C. D. and D. Tingley (2016). The influence of emotion on trust. *Political Analysis* 24 (4), 492–500.
- Potters, J. and F. Van Winden (1996). Comparative statics of a signaling game: An experimental study. *International Journal of Game Theory* 25(3), 329–353.
- Potters, J. and F. Van Winden (2000). Professionals and students in a lobbying experiment: Professional rules of conduct and subject surrogacy. *Journal of Economic Behavior & Organization* 43(4), 499–522.
- Prior, M. and A. Lupia (2008). Money, time, and political knowledge: Distinguishing quick recall and political learning skills. *American Journal of Political Science* 52 (1), 169–183.
- Ryan, R. M. and E. L. Deci (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology* 25 (1), 54–67.
- Sears, D. O. (1986). College sophomores in the laboratory: Influences of a narrow data base on social psychology's view of human nature. *Journal of personality and social psychology* 51 (3), 515–530.
- Smith, V. L. (1976). Experimental economics: Induced value theory. *The American Economic Review* 66 (2), 274–279.
- Valentino, N. A., T. Brader, E. W. Groenendyk, K. Gregorowicz and V. L. Hutchings (2011). Elections Night's alright for fighting: the role of emotions in political science. *The Journal of Politics*, 73 (1), 156–170.
- Viceisza, A. C. (2016). Creating a lab in the field: economics experiments for policymaking. *Journal of Economic Surveys* 30 (5), 835–854.
- Westen, D., P. S. Blagov, K. Harenski, C. Kilts and S. Hamann (2006). Neural bases of motivated reasoning: An fmri study of emotional constraints on partisan political judgment in the 2004 us presidential election. *Journal of Cognitive Neuroscience* 18 (11), 1947–1958.
- Whitt, S. and R. Wilson (2007). The dictator game, fairness and ethnicity in postwar Bosnia. *The American Journal of Political Science* 51 (3), 655–668.

# Field Experiments on the Frontier: Designing Better

Betsy Sinclair

## INTRODUCTION

‘Leave no trace’ is a motto used to characterize outdoor ethics – the notion that when humans interact with the outdoors, they should minimize their impact on the natural landscape. The phrase is also appropriate for field experimenters, whose goals should jointly be to extract as much scholarly insight as possible from each experiment but also to minimize the disruption of their experiment on the political landscape. As field experiments have grown in popularity – Desposato’s (2016) latest estimate is that 25% of all articles in the *American Journal of Political Science* contain an experiment – experimenters are under increasing pressure to ensure their experiments do not disrupt politics and maintain high ethical standards.

There are rising concerns about field experimental ethics. Some have to do with sample size – Bertrand and Mulainathan (2004) sent 5,000 fictitious résumés to potential

employers. Was the sample size worth the lost productivity? Others deal with questions of consent. What about experiments that contact public officials or elected representatives, individuals for whom the Institutional Review Board guidelines provide a near-blanket exemption for research? Does the scholarly value of the experiment offset the lost time and public funds spent replying to audit studies? What if an experiment affected an election outcome?

These are questions that continue to challenge the research community. Most of the responses from scholars have been to think broadly about individuals who might be impacted or about the fairness of who is likely to benefit from experimentation. This chapter aims to turn the conversation to one place where the tools of social science can help increase the quality of our experiments: experimental design.

Experimental ethics and the ‘leave no trace’ principle suggest that through experimental design, it is possible to design better



experiments. Our aims should be to maximize the scholarly value from each endeavor while also minimizing disruption, which may imply minimizing sample size. There are risks in engaging in experimental research in terms of disrupting the political landscape, and there are also costs that experimental researchers impose on subjects, whether deliberately or accidentally. Yet, there are also very real risks of failing to learn about our political world and carefully understand cause and effect. Some experimental designs leave less of a trace – and gain more knowledge – than others. This chapter provides a brief overview of five experimental designs that are on the frontier of the field experiment community's use. While infrequently used by political scientists, they have broad support from biomedical researchers, Google Analytics, and psychologists as strategies that can be employed to strengthen research impact. In the following section, this chapter will introduce each of these strategies and discuss the benefits of each. The chapter concludes with a call for expanding the EGAP conference model to researchers presenting research in the design phase, as a way for our field experiment community to continue to design better.

## EXPERIMENTAL DESIGNS

Many field experiments merit a very simple design – the randomized-control trial – in which the units of observation (e.g. individuals, groups, institutions, states) are randomly assigned to treatment and control groups. The randomization ensures that every observation has the same probability of being assigned to the treatment group. This implies that the outcomes are identical in expectation, so we are then able to draw causal inferences based upon the Rubin Causal Model (Rubin, 1978). Let  $i$  be an observation,  $Y_0$  be the outcome if  $i$  is exposed to control, and  $Y_1$  be the outcome if  $i$  is exposed to treatment.

The treatment effect is the difference between the two potential states of the world:

$$\tau_i = Y_{i1} - Y_{i0}.$$

Averaging across our units of observation, we define the average treatment effect (ATE) as

$$ATE = E(\tau_i) = E(Y_{i1}) - E(Y_{i0}).$$

With a handful of assumptions (typically made in this context: exclusion restriction, monotonicity, and SUTVA), we can then estimate this quantity using our randomized data as

$$E(Y_{i1} | T_i = 1) - E(Y_{i0} | T_i = 0).$$

This design is tremendously attractive in that it is highly intuitive: by simply subtracting the average outcome in the control group from the average outcome in the treatment group, it is possible to estimate the effect of the treatment. The comparison is made between the treatment and control groups. Yet, this is not the only possible design for field experiments. This chapter explores several alternative designs that researchers may want to consider.<sup>1</sup> In particular, many of these alternative designs have clear advantages in terms of the intensity and exposure for the intervention for the target population – some designs may be able to produce equally valid inferences, for example, with smaller sample sizes, and others may be able to allow a more dynamic adjustment of the treatment protocol. What is important to note about the conventional randomized-control trial is that the relevant comparison to draw causal inferences is made between observations. This chapter will briefly present several of these alternative designs that are not frequently employed in political science research and encourage researchers to think broadly about the optimal research design for their experimental framework, particularly in cases where their designs may permit variability in measuring outcomes over time.

## ***Crossover and Stepped-Wedge Designs***

Many experiments consider the unit of analysis to be a cluster (e.g. a school, a family, a village, a precinct), and experiments that employ this feature are called cluster-randomized trials. In this context, the clusters – and not the individuals – are randomized to treatment and control. They are frequently employed when individual randomization is not possible because of concerns regarding spillover or logistical hurdles (for an example, see Paluck and Green, 2009). These experiments are typically costly in terms of sample size because analysis is conducted at the cluster level. That is, inferences are conducted by looking between clusters.

In a crossover design, every cluster will instead receive both treatment and control. Randomization occurs with respect to timing – the order of the interventions is randomized for each cluster. This has a clear advantage in that it improves the efficiency of the cluster-randomized trial, as it permits the researcher to leverage variation both between and within clusters. Yet, for perhaps obvious reasons, it is not always prudent to conduct an experiment based on a crossover design, as this experimental setup assumes the absence of carryover effects. That is, the estimated treatment effects are independent of the order in which the treatment was assigned.

In many cases, it is completely unrealistic and unreasonable to assume that the effects of the treatment will have fully disappeared by the time the control wave has started. Thus, an extension of the crossover design is a stepped-wedge design. In this framework, clusters switch from control to treatment in only one direction and at different points in time. These distinct time points are the ‘steps’ of the experiment. Each cluster is randomly assigned a point in time to switch from control to treatment, and by the close of the experiment, all clusters will have switched from control to treatment.

There are clear ethical benefits to this design, first and foremost being that all

clusters in the experiment will eventually receive treatment – assuming the treatment is a normatively positive good – then this design ensures that all communities in the experiment will eventually benefit from having participated in the trial. It may also expand the number of institutional collaborators for researchers. For example, as reported in Gerber et al. (2011), the Perry campaign was willing to participate in the randomization of advertisements into media markets, conditional upon all markets receiving their advertising message. Another feature of this design is that clusters act as their own controls, as they receive both the control and treatment conditions. Thus, a second ethical benefit of a stepped-wedge trial is directly attributable to its statistical power – because this experimental design leverages both between and within cluster comparisons, there is more statistical power and thus the experiment requires a smaller sample size than a cluster-randomized experiment (Woertman et al., 2013; Baio, 2015).<sup>2</sup> It may be that the researcher would like to minimize their impact in a community, and thus minimizing the experimental size poses a clear advantage. There are practical reasons to favor a stepped-wedge experiment as well – in terms not of minimizing but maximizing the capacity to deliver treatment. The size of the experimental population may be limited by a budget constraint due to the capacity of the research time. A stepped-wedge design maximizes the delivery of treatment by the research staff. A final note is that this design ensures that even if the experiment were to stop during the experiment (occasionally unexpected things happen in field experiments), this design would permit inferences, because of the random assignment to clusters over steps.

A stepped-wedge design also contains some risk – it requires that the treatment be delivered without variability across waves. If, for example, the experimental treatment involved auditing, then it is possible that the auditors would become more experienced by the final wave and thus would not deliver a

constant treatment. Additionally, there can be no shifts in the underlying population during the experimental window.

### ***Multi-armed Bandit and Adaptive Design***

In the context where an experiment could potentially compare – or want to compare – an arsenal of competing treatments, again the randomized-control trial may not be the optimal choice. A multi-armed bandit design may help the researcher identify the most promising treatment. Like the stepped-wedge design, this type of experiment requires the possibility that it be conducted across multiple waves. The principle argument for a multi-armed bandit experiment is that of efficiency: this design will help the researcher ascertain the most effective treatment by adapting the proportion of subjects over multiple waves of the experiment.

The intuition for the multi-armed bandit comes from slot-machine gambling – the intuition is that there is an arsenal of competing slot machines (the bandits), where the researcher is tasked with simultaneously estimating each slot machine payout (explore) and using the best slot machine to maximize winnings (exploit).<sup>3</sup> Each exploration costs money that could otherwise have helped generate winnings, so there is a tension between exploration and exploitation. In political science language, a campaign might be interested in sending mailers, canvassers, robo-calls, or social media advertisements across a variety of channels to potential voters in order to moderate support. How should the campaign best optimize the use of its resources? Similarly, it is under time pressure to find the right answer *quickly* and then allocate as many of its resources to that strategy as possible. In this case, a multi-armed bandit experiment may be appropriate and, in some contexts, arguably more ethical, particularly when the design needs to produce results with speed.

How does this work? In brief, the researcher allocates some portion of experimental subjects to the explore phase, and in the first wave, subjects are allocated equally across all waves. After the first wave, the outcomes are evaluated and then, based upon the rates of success across treatments, new subjects are re-allocated in the second wave. So, for example, if one treatment arm was three times more efficacious than all others in the first wave, it would receive three times more subjects in the second wave. This process repeats in the third wave, with the proportion of subjects for each treatment adjusted by considering the outcomes in all previous waves.<sup>4</sup> Eventually, not only is the optimal treatment arm discovered but the majority of subjects are assigned to this arm.

Like with the stepped-wedge design, there can be no shifts in the underlying population during the period of experimentation. For a long multi-armed bandit process conducted over the course of an election campaign, during which new information may emerge about candidates, this could be problematic. Finally, multi-armed bandit processes may struggle if no one treatment is superior to the others (Offer-Westort et al., 2019).

### ***Spillover***

Frequently, field experiments administer treatment that consists of information. Most field experiments make the assumption that there are no spillovers – known as the non-interference assumption (part of SUTVA). Yet, in instances where researchers have designed their experiment to account for the possibility of information spillover, they have found at least suggestive (if not compelling) evidence that spillover does take place, most likely between close social ties (Nickerson, 2008; Bond et al., 2012; Sinclair et al., 2012).

Hierarchical design, where the social structures are explicitly theorized and randomization is conducted based upon those

structures, allows not only estimation of treatment effects but also spillovers (Hudgens and Halloran, 2008; Sinclair et al., 2012). Essentially, hierarchical design requires a weaker assumption than non-interference. Specifically, the assumption is based upon the particular social model used but generally permits subjects to spill over their treatment to subjects in nearby geographic spaces. That is, the assumption is that there will be only the modeled spillovers.

There are other strategies to explicitly acknowledge the possibility of spillover, too. These include Aronow and Samii's (2015) strategy to explicitly assign treatment based upon network structure and Bowers et al.'s (2013) strategy to consider modeling the explicit dosage an individual would receive based upon the possible spillover. What all these strategies have in common, however, is that they require the researcher to articulate a theory about where spillovers are likely to occur.

There is tremendous value in designing an experiment to account for spillovers. Beyond the social science knowledge we could glean from a better understanding of the social transmission of political information, designing an experiment that accounts for the potential of spillovers also gives us a way to understand the broad impacts of our research. While institutional review boards consider the impact of research on those we contact, they do not consider the impact on those who are socially connected to those we contact. If our treatments are contagious – if they spill over into adjacent neighborhoods, precincts, or ZIP codes – then not only does it provide critical social science knowledge, it also helps us better understand our research impact.

The range of experiments for which we do not have estimates of spillover – which instead fall into the family of experiments where information is delivered to a randomly assigned group of individuals – is vast. For example, they include experiments where the information focuses on corruption (Ferraz and Finan, 2008; Chong et al., 2011), violence (Collier and Vicente, 2014), and voting

(Gerber et al., 2008; Sinclair et al., 2013). In each of these instances, these experiments may have impacted the family and friends of those contacted. Imagine instead if we knew what the spillover estimates looked like in these contexts! Whether hierarchical or otherwise, incorporating estimates of spillover into a research-design protocol allows not only the estimation of the social effect but also the capacity to understand who may inadvertently be affected by the experiment.

### ***Selective Trials***

A lively and ongoing debate regarding the external validity of field experiments (Deaton, 2010; Imbens, 2010; Deaton and Cartwright, 2018) generated a broad set of interest in the extent to which the experimental subjects themselves could be modeled in order to increase external validity. Selective trials offer one such strategy (Chassang et al., 2012): extending the randomized-control trial to estimate the degree of effort needed to be deployed by subjects if the results are to be generalized.

In short, the outcome observed by each subject is a consequence of her treatment assignment and effort. Yet, her effort is private information. Insight into that information can be gleaned via mechanism design, by charging a price for treatment: subjects first send a message to the researcher, revealing the price they would pay for treatment, and then once price is drawn from a distribution of possible prices, subjects are allowed to purchase the treatment if the price is below their message. The willingness to pay is a good instrument for estimating marginal causal effects (Heckman and Vytlacil, 2005), as willingness to pay is a good signal of effort. The aim of this additional feature is to disentangle the treatment effect from the effort effect (associated with the belief of the subject as to the efficacy of the treatment). This has the clear benefit of increasing the external validity of the experiment, given that it demonstrates the

value the subjects of the experiment place on the treatment.

Apart from the logistical hurdles of implementing a selective trial design in the field, it may also be the case that the subjects may not play dominant strategies (in which case their messages would not be good signals of their actual effort).

### ***Causal Mechanisms***

Randomized-control trials allow us to make inferences about whether the treatment causes the outcome. However, they frequently do not tell us how and why the treatment causes the outcome. The how and why components are the pieces we refer to as causal mechanisms – the causal pathway from treatment to outcome. Many times, the variable that was manipulated – or can be manipulated – is not the mediator but rather the treatment variable.<sup>5</sup> Imai and Yamamoto (2013) suggest a strategy to elicit causal-mediation effects via research design. For example, they propose a parallel design where in one half of the sample, treatment is randomized, the mediator is measured, and the outcome is measured. In the second half of the sample, treatment is randomized, the mediator is randomized, and the outcome is measured. In this framework (with a consistency assumption), the ACME is identified. While many mediators are simply not easily manipulable, there are other strategies that can nudge variation, such as encouragement designs (Imai and Yamamoto, 2013; Imai et al., 2013).

Not every experiment can manipulate a mediator. But, for those where manipulation is possible, it transforms the randomized trial into one that reveals not only whether the treatment causes the outcome but also how and why. This chapter argues that when experiments take place, we should squeeze all possible knowledge out of that endeavor. Learning everything we can about a treatment ensures that the experiment produces the most knowledge possible for future scholars.

### **CONCLUSION**

As the community of scholars begins to consider the impact of field experiments on the world at large – and the ethics of field experiments – it is important to recognize that ethics and experimental design are inherently and deeply intertwined. The scholarly value of a study is deeply associated with its design – and this chapter suggests five strategies that are at the frontier of field experimental research in political science, but which can augment research designs. By minimizing our impact on the world while maximizing our scholarly insights, we can design better and more ethical field experiments.

This chapter has explored five potential strategies for extending field experiments to the frontier of our discipline. Each have clear ethical arguments as well as practical ones. The stepped-wedge design permits greater flexibility with institutional partners. Instead of asking organizations to provide a control group – which can be challenging unless there is a binding budget constraint that prevents them from contacting all experimental subjects – a stepped-wedge design ensures that all subjects will receive treatment, just at randomly varying times. This research design has the added benefit of improving statistical power and, in some cases, researcher efficiency.

Adaptive designs – typically framed as multi-armed bandit designs – can quickly ascertain the most productive treatment arm. Much of the experimental work that happens in the field experiment community occurs in the developing world, where resources are scarce and quick improvements in policy can truly have life-changing outcomes for experimental subjects. So, whether the experimental treatment options involve de-worming drugs vs school uniforms vs teacher photographs to study school attendance or a myriad of media channels to contact potential donors, a multi-armed bandit design can expedite the learning process.

Many field experiments surely have treatments that spill over from one individual or community to another – particularly when

that treatment is information. As an academic community, we have few estimates of these effects. Designing experiments to measure spillover not only allows the research community to accumulate knowledge about contagion but also ensures that researchers are aware of the impact their research has on those who are not explicitly part of the study.

Selective trials offer one strategy to increase the generalizability of our experiments. There will surely continue to be ongoing debate on the merits of estimating an average treatment effect for a specific population, and by pushing our research to think carefully about the ways in which the underlying population would otherwise seek out the treatments – and at what cost – is one way to consider the potential impacts these experiments could have more broadly.

Finally, embedding causal-mediation analysis into experimental work, when possible, allows us to better understand the causal process. If there are ways to do additional experimentation, such as the parallel design advanced by Imai and Yamamoto (2013), it may be possible to embed causal-mediation analysis into a conventional randomized-control trial. Each study that makes this choice increases our knowledge of the world around us.

Research-design choices are challenging choices. Recently, an experimental community – EGAP – has grown, which challenges researchers to design better by encouraging them to present research designs in advance of starting an experiment. These conversations push scholars to think more broadly – and arguably more ethically – about the experiments they are likely to conduct. Software, such as DeclareDesign,<sup>6</sup> makes this easier for researchers to accomplish. This chapter has suggested only five possible designs that are used infrequently – although they are generally applicable to both survey and field experimentation. Finding the right design requires a research community. The EGAP model should be extended to all major political science conferences, so that researchers are strongly encouraged to present research at

the design stage. By working collaboratively, we can continue to design better, stronger, richer, and more ethical experiments.

## Notes

- 1 We can also define the average treatment effect among the treated (ATT) as

$$ATT = E(\tau_i | T_i = 1) = E(Y_{i1} | T_i = 1) - E(Y_{i0} | T_i = 1)$$

- 2 A stepped-wedge design could also be considered at the individual level, allowing the researcher to draw from both within and between variation. Indeed, repeat measures on individuals produces a powerful experimental design, although it requires the researchers to be able to measure the subject's outcomes at each wave in the experimental design.
- 3 Multi-armed bandit experiments are used by Google Analytics to place advertisements because of their efficiency relative to randomized-control trials (Whittle, 1980) and are increasingly employed in medical trials (Chin, 2016).
- 4 For more detail and a technical explanation, please see Offer-Westort et al., 2019.
- 5 For example, Brader et al. (2008) vary ethnic cues to evaluate their effect on immigration attitudes. They expect that this effect is manipulated by anxiety. Based upon the methodological innovation of Imai et al. (2012), it is possible to calculate the average causal-mediation effect (ACME), assuming sequential ignorability.
- 6 <https://declaredesign.org/about/>

## REFERENCES

- Aronow, P. M. and Samii, C. 2015. *Estimating Average Causal Effects Under Interference Between Units*. Working paper.
- Baio, G., Copas, A., Ambler, G., Hargreaves, J., Beard, E. and Omar, R. Z. 2015. 'Sample Size Calculation for a Stepped Wedge Trial'. *Trials*, 16: 354.
- Bertrand, M. and Mullainathan, S. 2004. 'Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination'. *American Economic Review*, 94(4): 991–1013.
- Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D. I., Marlow, C., Settle, J. E. and Fowler, J. H. 2012. 'A 61-Million-Person Experiment in

- Social Influence and Political Mobilization'. *Nature*, 489: 295–298.
- Bowers, J., Fredrickson, M. M. and Panagopoulos, C. 2013. 'Reasoning about Interference Between Units: A General Framework'. *Political Analysis*, 21: 97–124.
- Brader, T., Valentino, N. and Suhay, E. 2008. 'What Triggers Public Opposition to Immigration? Anxiety, Group Cues, and Immigration Threat'. *American Journal of Political Science*, 52(4): 959–978.
- Chassang, S. Padro, G. and Snowberg E. 2012. 'Selective Trials: A Principal-Agent Approach to Randomized Controlled Experiments'. *The American Economic Review*, 102(4): 1279–1309.
- Chin, R. 2016. *Adaptive and Flexible Clinical Trials*. New York: CRC Press.
- Chong, A., De La O, A., Karlan, D. and Wantchekon, L. 2011. *Looking beyond the incumbent: The effects of exposing corruption on electoral outcomes*. NBER Working Paper No. w17679. Available at [www.nber.org/papers/w17679.pdf](http://www.nber.org/papers/w17679.pdf) (Accessed on 17 January 2020).
- Collier, P. and Vicente, P. 2014. 'Votes and Violence: Evidence from a Field Experiment in Nigeria'. *The Economics Journal*, 124(574): 327–355.
- Deaton, A., 2010. 'Instruments, Randomization, and Learning about Development'. *Journal of Economic Literature*, 48(2): 424–455.
- Deaton, A. and Cartwright, N. 2018. 'Understanding and Misunderstanding Randomized Controlled Trials'. *Social Science and Medicine*, 210: 2–21.
- Desposato, S. (ed.) 2016. *Ethics and experiments: Problems and Solutions for Social Scientists and Policy Professionals*. New York: Routledge.
- Ferraz, C. and Finan, F. 2008. 'Exposing Corrupt Politicians: The Effects of Brazil's Publicity Released Audits on Electoral Outcomes'. *The Quarterly Journal of Economics*, 123(2): 703–745.
- Gerber, A. S., Gimpel, J. G., Green, D. P. and Shaw, D. R. 2011. 'How Large and Long-Lasting are the Persuasive Effects of Televised Campaign Ads? Results from a Randomized Field Experiment'. *The American Political Science Review*, 105(1): 135–150.
- Gerber, A., Green, D. and Larimer, C. 2008. 'Social Pressure and Vote Turnout: Evidence from a Large-Scale Field Experiment'. *The American Political Science Review*, 102(1): 33–48.
- Heckman, J. and Vytlacil, E. 2005. 'Structural Equations, Treatment Effects, and Econometric Policy Evaluation'. *Econometrica*, 73(3): 669–738.
- Hudgens, M. G. and Halloran, M. E. 2008. 'Toward Causal Inference with Interference'. *Journal of the American Statistical Association*, 103(482): 832–843.
- Imai, K. and Yamamoto, T. 2013. 'Identification and Sensitivity Analysis for Multiple Causal Mechanisms: Revisiting Evidence from Framing Experiments'. *Political Analysis*, 21(2): 141–171.
- Imai, K., Tingley, D. and Yamamoto, T. 2013. 'Experimental Designs for Identifying Causal Mechanisms'. *Journal of the Royal Statistical Society Series A*, 176(1): 5–51.
- Imai, K., Tingley, D. and Yamamoto, T. 2012. 'Unpacking the Black Box of Causality: Learning about Causal Mechanisms from Experimental and Observational Studies'. *American Political Science Review*, 105(4): 765–789.
- Imbens, G. W., 2010. 'Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009)'. *Journal of Economic Literature*, 48(2): 399–423.
- Nickerson, D. 2008. 'Is Voting Contagious? Evidence from Two Field Experiments'. *The American Political Science Review*, 201(1): 49–57.
- Offer-Westort, M., Coppock, A. and Green, D., 2019. *Adaptive Experimental Design: Prospects and Applications in Political Science*. Working paper.
- Paluck, E. L. and Green, D. P. 2009. 'Deference, dissent, and dispute resolution: An Experimental Intervention Using Mass Media to Change Norms and Behavior in Rwanda'. *American Political Science Review*, 103(4), 622–644.
- Rubin, D. B. 1978. 'Bayesian Inference for Causal Effects: The Role of Randomization'. *Annals of Statistics*, 6(1): 34–58.
- Sinclair, B., McConnell, M. and Green, D. P. 2012. 'Detecting Spillover Effects: Design and Analysis of Multilevel Experiments'. *American Journal of Political Science*, 56(4): 1055–1069.
- Sinclair, B., McConnell, M. and Michelson, M. 2013. 'Local Canvassing: The Efficacy of Grass-roots Voter Mobilization'. *Political Communication*, 30(1): 42–57.
- Whittle, P. 1980. 'Multi-Armed Bandits and the Gittins Index'. *Journal of the Royal Statistical Society, Series B*, 42(2): 143–149.
- Woertman, W., de Hoop, E., Moerbeek, M., Zuidema, S., Gerritsen, D. and Terrenstra, S. 2013. 'Stepped Wedge Designs could Reduce the Required Sample Size in Cluster Randomized Trials'. *Journal of Clinical Epidemiology*, 66(7): 752–758.

# Field Experiments, Theory, and External Validity

Anna M. Wilke and Macartan Humphreys

## INTRODUCTION

In a recent study, Yeh et al. (2018) report on the first randomized control trial that tests whether there are health benefits from wearing parachutes when jumping out of airplanes and helicopters. The authors find no evidence of benefits from wearing parachutes. A potential weakness of the study is that, in order to protect human subjects, the airplane and helicopter used for the trial were small, stationary, and grounded. For critical readers this detail might raise flags about whether we can generalize from this study to other applications of interest.

Of course, Yeh and colleagues meant their study as a joke. It is *obvious* that you cannot learn anything about effects in realistic situations from this experiment. We think, however, that it is not entirely obvious why it is obvious. The experiment lacked external validity, perhaps, because the ‘target’ applications of interest involve planes that are in flight and far from the ground. The sample of

jumps is not *like* the population of jumps out of an aircraft that we are interested in. Yet we might not have that concern were the plane in the air but not in flight – for example, if it were attached to a large crane, even though in this case, too, the study would not be like the target applications of interest. Why would we be less concerned? The reason is that our determination of ‘likeness’ depends on a prior model of how things work – in this case that falling at speed is a key mechanism that is interrupted by parachutes and that speed depends on the height from which you jump. So, it seems, a theory is used to assess external validity.

This chapter is about such theories and the role they play in assessing external validity. We are motivated by persistent concerns with field experimental approaches, which have seen an explosion in political science in the last 15 years (Druckman, et al., 2011<sup>1</sup>; Grose, 2014). Core concerns are that experimental results lack external validity and are disconnected from theory. As in the parachute



example, these two concerns are deeply interrelated. Put differently, the charge is that political scientists are doing parachute experiments from stationary planes without knowing it.

In the following sections, we describe these concerns and review strategies to address them. We begin by introducing a stylized example that we draw on throughout the chapter. We then review challenges that arise from a weak connection to theory and limited external validity and describe approaches to address these. Finally, we extend our example to illustrate opportunities and risks of parametric structural estimation with experimental data, an approach that links estimation closely to theoretical models and that, in principle, allows for inferences that go beyond those available with design-based approaches alone.

## PRELIMINARIES

### *A Running Example*

Throughout this chapter, we will make use of a stylized example of a fictitious field experimental study that seeks to explain bargaining outcomes. We imagine a study of the bargaining process between taxi drivers and customers. See Michelitch (2015) and Castillo et al. (2013) for existing studies of this kind.

We assume we want to explain why some taxi customers have to pay higher taxi fares than others. We start with a simple theory that says that taxi prices are determined by three variables: whether the customer or the driver makes the first offer, how many offers and counteroffers can be made before bargaining breaks down (the number of bargaining rounds), and the behavioral ‘type’ of the customer.

Whether the customer makes the first offer might be thought of as linked to the customer’s identity, assertiveness, or skills. Skilled negotiators, for example, may be more likely

to approach the driver with an offer. How long bargaining can continue may depend on contextual factors such as the competitiveness of the taxi market. Finally, customers may behave in different ways. Some customers may be ‘rational’ decision makers in the narrow sense that they seek to maximize their material gains from the bargaining process. Others may seek to follow established fairness norms. For simplicity, we presume that taxi drivers are always of the first type.

With this theory in mind, we imagine running a field experiment that focuses, first, on the effect of making the first offer. Suppose we recruit a sample of taxi customers, provide each of them with the same sum of money and ask them to negotiate a taxi fare for a certain distance. Taxi customers can keep whatever they do not spend on the ride. We randomly assign half of the taxi customers to approach the driver with an offer (‘move first’) while the other half is instructed to ask the driver for a price (‘move second’). Additionally, we will imagine being able to control for how long bargaining can continue. We will assume players have common knowledge over endowments and types.

This example is certainly contrived. But it has the advantage that it draws on simple and well understood theoretical models, which makes the example useful for demonstrating core ideas and for walking through the development and use of structural models. One should not infer, however, that models of this form are always so simple to set up and estimate.

### *What We Mean by Theory*

Critics worry that a disconnect from theory limits the types of inferences that can be drawn from experimental research (Card et al., 2011; Deaton and Cartwright, 2018; Harrison, 2014; Huber, 2017).

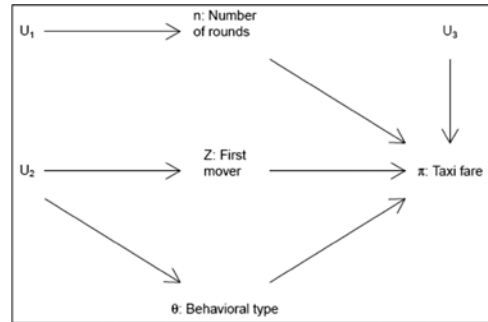
What is theory? Conscious that the term is used to mean very different things in different research communities,<sup>2</sup> we will make use

of a quite simple notion of theory: by theory we will mean a set of general claims about the causal relations among a set of variables, from which more specific claim (for instance, about a case) can be deduced under the supposition that the theory is true.<sup>3</sup> For instance, if theory *T* says that all taxi customers who make the first offer end up paying a lower fare than those who let the driver make the first offer, then the specific claim ‘this customer will pay less if she makes the first offer’ follows from the theory. Theory *T* is not very deep; it is little more than a proposition. Yet, theory *T* itself could be implied by a deeper theory, for example a claim about equilibrium offers in a class of bargaining games (of which we take the taxi game to be an instance). We thus treat depth as a property of a theory and not as part of the definition of theory.

Drawing on work by Pearl (2009) and the treatment in Humphreys and Jacobs (2017), we think it useful to distinguish between three ‘levels’ of theory that differ in the specificity with which they describe the relations between variables. Even though, in practice, not all theories will neatly fall under one of these levels, this classification helps organize our thinking about theories and theory development.

**Level 1:** Level 1 theories identify variables of interest and specify the structure of causal relationships between them. Examples are non-parametric structural equation models or causal DAGs (Directed Acyclic Graphs) (Dawid 2010).<sup>4</sup>

Figure 53.1 shows an example of a causal DAG for the Level 1 version of our simple bargaining model. The graph lists the variables that the theory considers and the structure of relations between these variables. An arrow from *A* to *B* means that, in some situations, a change in *A* induces a change in *B*. There are unobserved factors  $U_3$  that affect the taxi fare and unobserved factors  $U_2$  that affect both, whether the customer makes the first offer and the customer’s behavioral type. The Level 1 version of our theory also contains information about which variables do not affect each other. For example, the



**Figure 53.1** A directed acyclic graph (DAG) for a non-parametric bargaining model

number of rounds for which bargaining may continue is determined by unobserved factors  $U_1$ , but these unobserved factors do not have a direct impact on any of the other variables in the model.

Note that although the theory involves many substantive claims, it has no implications for effect sizes or even whether or not variables interact with each other to produce outcomes.

**Level 2:** rather than simply specifying the qualitative structure of the causal relationships between variables, a theory may contain statements about the functional form of these relationships. These might be qualitative statements about marginal effects, for example: ‘making the first offer in taxi bargaining reduces the price that the customer has to pay’. Alternatively, a Level 2 model may contain fully specified functional relations,  $f = f_1, f_2, \dots, f_n$ , one for each endogenous variable.

For instance, a parametric bargaining model might add to the Level 1 model the following functional relationship between the number of possible bargaining rounds  $n$ , the customer’s first-mover status  $z_i$ , the customer’s behavioral type  $\theta_i$  and  $\pi_i$ , the taxi fare paid by customer  $i$ :

$$\pi_i = \theta_i(z_i\omega + (1 - z_i)(1 - \omega)) + (1 - \theta_i)\phi + u_3$$

where  $\omega = \sum_{t=2}^n (-1)^t \delta^{t-1}$  is the taxi price predicted by the Rubinstein (1982) bargaining

solution under the assumption that taxi bargaining follows an alternating offers protocol with  $n$  possible bargaining rounds, that customers and drivers act rationally, and that the customer gets to make the first offer. This representation normalizes the ‘pie’ that the customer and driver are bargaining over to 1. In the context of our running example, we can treat  $\omega$  as the share of her endowment that a rational customer ends up paying to the driver if she moves first.  $\delta$  stands for a discount factor (see the section ‘Illustration of a Parametric Structural Model Connecting Theory and Experimentation’ for more details). This Level 2 version of our theory implies that rational customers ( $\theta_i = 1$ ) pay the price implied by the Rubinstein bargaining solution. This price depends on whether customers go first ( $z_i = 1$ ) or second ( $z_i = 0$ ). Non-rational customers ( $\theta_i = 0$ ) insist on giving the driver some share  $\phi$  of their endowment, irrespective of whether they go first or second. The last term,  $u_3$ , is a random disturbance.

This theory is not well motivated, but, as is well known, the behavior of rational types can be derived from a more complex Level 2 model that fully specifies the extensive form of the bargaining game and allows for a richer characterization of optimal behavior (e.g. offers and responses in every period). We provide this motivation below.

Unlike a Level 1 theory, the Level 2 theory *can* make claims about the values of endogenous variables given the values of exogenous variables. However, even a Level 2 theory might not be enough to make claims about the average effects of treatments.<sup>5</sup>

**Level 3:** a still more fully specified model might add assumptions about how the set of unobserved exogenous variables  $U$  is distributed. A theory that specifies all of these elements is called a probabilistic causal model by Pearl (2009).

Note that a Level 3 model implies a Level 2 model and a Level 2 model implies a Level 1 model, but the converse is not true. Note

also that starting from a Level 1 model, one could use data to develop a Level 2 or a Level 3 model.

### ***What We Mean by External Validity***

The most common external validity critique holds that findings from field experiments may not tell us much about processes in contexts other than the one where the findings were generated (Lucas, 2003). Many discussions of threats to external validity focus on whether the study population is sufficiently similar to a *target* population of interest. This framing gives rise to what might be the more useful term, ‘target validity’. Can we make claims to a particular target of inference (Westreich et al., 2018)?

More broadly, Cronbach and Shapiro (1982) describe four dimensions along which studies may differ from their inferential targets, in terms of units, treatments, outcomes, and settings (collectively, ‘UTOS’). Using the language of causal models, we can distinguish three of these four dimensions. Units and contexts are equivalents in the language of causal models.

Let  $W$  denote an observed or unobserved collection of background variables;  $Y$  and  $V$  denote outcomes, and  $Z$  and  $X$  denote treatments. The question about units and contexts is:

- What can we infer about the distribution  $\Pr(Y = y|Z = z, W = w)$  from our knowledge of  $\Pr(Y = y|Z = z, W = w)$ ? For instance, what does an experiment in a place with competitive taxi markets tell us about the effect of moving first in taxi bargaining when taxi competition is low?

The treatments question is:

- What can we infer about the distribution  $\Pr(Y = y|Z = z)$  from our knowledge of  $\Pr(Y = y|X = x)$ ? For instance, what does a study

of the effect of moving first in taxi bargaining on prices tell us about the effect of an increase in taxi competition on prices?

The outcomes question is:

- What can we infer about the distribution  $\Pr(V = v | X = x)$  from our knowledge of  $\Pr(Y = y | X = x)$ ? For instance, what do we learn from the impact of moving first in taxi bargaining on taxi fares about the effect of moving first on the duration of taxi negotiations?

## CONCERNS AROUND THEORY AND EXTERNAL VALIDITY IN FIELD EXPERIMENTS

### *Why Worry About Theory?*

We can distinguish at least four worries that result from weak connections between theory and experimental research. Two relate to the applicability of findings and two, more fundamentally, to the orientation of research.

#### *Undefined scope: units and settings*

As in the parachute example, without theory, researchers will not have the information needed to think through which units and settings are ‘similar enough’ to a given study environment in order to license the transportation of results.<sup>6</sup> The concern might be less severe if experiments use random samples from the target population of interest and produce homogeneous results within strata. However, if such conditions do not hold and researchers do not have access to theory, they may be at a loss about what broader inferences they are licensed to draw.

#### *Undefined scope: treatments and outcomes*

It is often not feasible to experimentally evaluate all relevant versions of a particular

intervention or to examine all outcomes of interest. Target inferences may then require extrapolation to effects of treatments, or effects on outcomes that were not studied by an experiment. For example, field experiments tend to involve versions of interventions that have a lower dosage or shorter time horizons than target applications. Similarly, field experiments sometimes examine effects on proxies of ultimate outcomes of interest. Consider, for instance, a study that shows that information changes attitudes of voters but does not feature data on actual voting behavior. Making inferences to voting behavior then requires understanding how attitudes affect voting (among those for whom information affects attitudes).

These two worries illustrate the connection between theory and external validity. Unaccompanied by theory, experimental results do not give a handle on how to extrapolate out of sample or make inferences to other relations of interest. Weak connection to theory leaves open the question of whether and how results can be generalized.

The next two concerns relate to the questions being asked.

#### *Restrictive estimands*

Experimental research largely limits itself to estimating variants of the average treatment effect. By the magic of the linearity of expectations, random assignment lets one estimate average differences between outcomes in treatment and control conditions by looking at the differences in average outcomes in treatment and control groups. However, many questions of interest are not summaries of average treatment effects. Consider, for example, questions about so-called ‘causes of effects’ as opposed to ‘effects of causes’. Imagine a field experiment that randomly assigns individuals to go first in bargaining. Randomization does not, in general, let researchers answer the question: ‘Knowing that individual *A* made the first offer, what is the probability that she would have paid a

higher price had she not made the first offer?' Murtas et al. (2017) show that, although this quantity is not identified, bounds around this probability can be tightened by drawing on theoretical knowledge about the variables that mediate the relationship between the treatment and the outcome of interest (even if such mediators are not observed).

### *The point of empirical work is to learn about theories*

The point of research, critics argue, is to understand *how* things work. In other words, the *goal* is to develop theory. Engaging in research while ignoring theory is thus missing the point. Experimentalists, in this view, are too often satisfied with 'black-box' accounts that do not go beyond causal relations between variables. Deaton (2010), for instance, advocates for a more explicit focus on the usage of causally identified empirical work to test hypotheses that are derived from a lower-level theory in order to learn about the validity of the theory itself, and not just to generate a case-specific estimate.

Subsequently, we will explore in more detail some of the ways in which researchers have responded to these worries.

### *Why Worry about External Validity?*

Some of the concerns around external validity apply to all research. In general, empirical work does not ever draw on a random sample of the units to which inferences might be made – if only because we want to make inferences to future events from past events. In a sense, experiments can be thought of as case studies and, hence, as facing challenges that affect all case study research.

Beyond that, however, there are ways in which experiments are especially vulnerable to external validity concerns. Through experimentation, researchers alter the world to make it amenable to study. A core worry is

that, in doing so, researchers actively create distance to target environments.

### *Threats arising from control*

Insofar as experimenters control the conditions of an experimental study, they risk orchestrating environments that differ from the target environment in ways that researchers might not be aware of.

Experiments may, by design, hold relevant variables constant at typical levels. For example, suppose we are interested in both the effect of going first in taxi bargaining and the effect of increasing the number of possible bargaining rounds. Say we randomly assign taxi customers to go first and also vary for how long bargaining can continue. An experiment of this kind allows us to estimate the effect of additional bargaining rounds among customers who have been assigned to go first or second. Yet, this effect may differ from, say, the effect among those who actually choose to go first outside our experiment. Voluntary first movers may be systematically different from other customers. A design that randomizes who goes first provides no information about who would have gone first outside the experiment and, as a consequence, makes it impossible to obtain effect estimates among this group without further assumptions.

In addition, *how* variables are controlled can matter. The process of random treatment assignment in itself may have consequences. For instance, voters might not reward a politician for patronage if they know that patronage was distributed at random, precisely because rewards normally result from the information communicated by a transfer rather than by the transfer itself (see also the treatment in Mesquita and Tyson, 2019). The problem here is an exclusion restriction violation.

### *Threats arising from selection*

Very often, neither the sites nor the subjects of experimental studies are randomly selected. Especially for experiments that depend on

partnerships with government or other implementing partners, site selection tends to be contingent on the willingness of partners to participate. Yet, governments that are willing to assess anti-corruption interventions, for example, may be fundamentally different from those that are not. The problem is akin to that in medical studies where subjects who are willing to participate tend to be those most likely to stick to regimens. See Allcott (2015) for an example of how bias can arise from non-random site selection.

Problems can also arise if experimentalists study treatments, or variants of treatments, that are amenable to manipulation but wish to make inferences to treatments that are not (Thelen and Mahoney 2015). A common example is a situation in which a small scale intervention is studied – because it is feasible – though the target of inference is the effect of a large scale intervention (Bold et al. 2013).

### *Threats arising from inferential strategies*

External claims from experiments can sometimes be rendered difficult by the fact that common approaches to analyzing experimental data implicitly assume that the target of inference is the sample estimand, even if this is not always explicitly stated.

Randomization inference, for instance, can be used to calculate exact  $p$ -values – but only under the assumption that all variability comes from assignment processes within the sample and not from the selection of the sample itself. Similarly, at least under the assumption of constant effects, clustering standard errors at the level of treatment assignment can produce confidence intervals with the correct coverage. Yet, this approach implicitly assumes that study units have not been randomly sampled in a clustered way. Suppose, for example, that a study randomly samples a set of schools and assigns an intervention on the classroom level. In this case, clustering might have to be performed at levels above the level of assignment – though

this analysis strategy is not common practice (Abadie et al., 2017).

To be clear, these are ways in which the tools of experimental analysis tend to orient researchers towards sample inference, even though, in principle, approaches that focus on population inference can certainly be employed with experimental data.

## **THE PLACE OF THEORY IN EXPERIMENTAL RESEARCH**

In practice, experimentalists can employ a range of strategies to combine theory and experimental work (or not). We discuss six of these.

### ***Strategy 1: Push Back – Experimentation Obviates the Need for Theory***

A first response to the critique that experimental research tends to be atheoretical is unapologetic. The absence of theory is a strength. As characterized by Heckman (1991), the ability to dispense with theory was, if anything, a motivation to engage in experimentation. That you can find out whether democratic institutions cause growth without having to assume a model of human behavior is remarkable, and to be celebrated. The identification of average causal effects is not possible from observational data without a model that tells you which variables you should or should not condition on (Pearl, 2009). An experiment removes much of this model dependence.<sup>7</sup>

This response leaves open the question of how experiments can be used to learn about theories. Moreover, despite the remarkable ability of experiments to identify average treatment effects under minimal assumptions, there are quantities that cannot be learned from an experiment. Both topics are discussed in more detail below.

### **Strategy 2: Use Theory as a Helper for Inference**

One strategy for bringing experiments together with theory is to use theories to draw inferences that could not be drawn based on experimental data alone. Consider the following examples:

- 1 *Transportation of results to other settings.* Suppose we conduct an experiment at a site *A* and we would like to use our treatment effect estimates to learn about the treatment effect at site *B*. Unless site *A* was randomly chosen from some population of sites that also contains site *B*, our experimental data alone do not speak to whether and how transportation from *A* to *B* is possible, unless additional assumptions are made. We discuss in the section ‘Strategy 6: Formal Transportation’ how a causal model can help answer these questions and provide an example in the section ‘Illustration of a Parametric Structural Model Connecting Theory and Experimentation’.
- 2 *Predicting the effects of other treatments.* Many causes of theoretical interest cannot be experimentally manipulated. Similarly, it is often prohibitively expensive to evaluate all policy-relevant variations of an intervention with an experiment. One solution is to make use of a structural causal model that serves as the basis for the extrapolation of estimates from a single, possibly small-scale experiment to the effect of a different intervention or of the same intervention at a larger scale. Todd and Wolpin (2006), for example, use a structural causal model to extrapolate estimates of the effect of a randomized school subsidy program in Mexico to the effects of similar programs with different subsidy schedules. See ‘Illustration of a Parametric Structural Model Connecting Theory and Experimentation’ for an example.
- 3 *Inferences to unidentified causal quantities.* There are causal quantities that cannot be estimated without additional assumptions, even when the treatment of interest has been randomized. For instance, as noted above, while the ‘causes of effects’ estimand is not identified by randomization, randomization can be used to generate upper and lower bounds. Dawid et al. (2019) show how these bounds can sometimes be tightened if we measure the values of mediators

through which the treatment affects outcomes. In short, knowing that *X* causes *Y* through *M* may improve our inferences about whether *X* caused *Y* for a particular unit.

### **Strategy 3: Use Theory as a Helper for Design**

Apart from helping with inference, theories can provide guidance for various aspects of experimental design. For example:

*Site selection:* causal models may help us decide where to run an experiment. We may want to choose, for instance, a site that allows results to be transported to as many other settings as possible. As we discuss in the section ‘Strategy 6: Formal Transportation’, causal models provide guidance on the extent to which an experiment conducted in one setting will be informative for treatment effects in other settings. Alternatively, if our aim is to test a causal model (see the section titled ‘Strategy 5: Use Experiments to Put Theories to the Test’), we may consider contexts for which the model’s predictions differ from the predictions of alternative models.

*Treatments:* causal models may help researchers decide which treatments to implement. If the aim is to test a causal model, we would obviously like to randomize causes relevant to the model. If the aim is to estimate the parameters of a structural model (see ‘Strategy 6: Use Experiments to Estimate Structural Models’), additional treatments can sometimes help with the identification of model parameters other than the effect of the treatment itself (DellaVigna, 2018).

*Sampling:* in the section ‘Strategy 3: Exploit Variation Within Studies’, we describe a way of transporting treatment effect estimates from a setting *A* to a setting *B* by calculating weighted averages of estimates within subgroups. Crucial for our ability to use this strategy is that our subject pool in setting *A* contains enough subjects in each subgroup to estimate treatment effects within these groups. Prior knowledge of

a causal model that tells us the dimensions along which treatment effects vary can help us design a sampling strategy that achieves this goal.

*Random assignment:* the same considerations affect the way in which we assign units to treatment conditions. Being able to estimate effects in a subgroup requires a sufficient number of treated and untreated subjects in this subgroup. We can fix the number of treated subjects in each subgroup by assigning treatment within blocks. Causal models thus help with the design of blocking schemes. In the presence of spillovers, causal models can also help decide how to allocate units across experimental conditions in order to maximize statistical power (Bowers et al., 2018).

*Measurement:* a causal model can help assess which covariates need to be measured in setting  $A$  and  $B$  in order to be able to transport treatment effect estimates from  $A$  to  $B$ . A causal model can also give us guidance on which variables we need to measure if we would like to bound our estimates of the probability of causation (see point 3 in the previous section).

#### **Strategy 4: Use Experiments as Building Blocks of Theories**

A fourth approach is to think of the ability of experiments to identify average treatment effects as an opportunity for inductive learning about theories. Being able to claim that  $X$  causes  $Y$  already establishes a theory of sorts. Beyond that, researchers sometimes stitch experimental results together to form more elaborate theories.

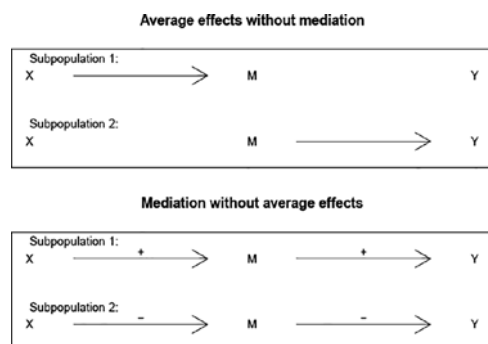
Imagine that we run our experiment that randomizes whether customers make the first offer in taxi bargaining. Moreover, suppose that we also find a way to randomly vary how many offers and counter-offers can be made before bargaining breaks down. For the sake of the argument, assume that the customers in our experiment are a random sample

from the population of interest. From this experiment, we could obtain estimates of the average treatment effect of going first for different numbers of bargaining rounds and of the average treatment effect of additional bargaining rounds conditional on whether the customer goes first. Even without prior knowledge of the causal model presented in the section ‘What we Mean by Theory’, we could use these estimates to ‘piece together’ the functional relationship between moving first, the number of bargaining rounds and the expected taxi fare in the population of interest.

Unfortunately, things become more complicated when causal models are slightly more complex. Consider an effort to establish that  $X$  causes  $Y$  through  $M$  by stitching together estimates of the effect of  $X$  on  $M$  and the effect of  $M$  on  $Y$  (Imai et al., 2011; Green et al., 2010).

Figure 53.2 shows two problems one might run into.

- 1 Even if a treatment  $X$  has an average treatment effect on a variable  $M$  and  $M$  has an average treatment effect on  $Y$ ,  $M$  might not be a mediator of the relationship between  $X$  and  $Y$  for any unit. For example, as in the upper panel of Figure 53.2,  $X$  may affect  $M$  among some set of units and  $M$  may affect  $Y$  among another set of units, but these sets may not overlap.



**Figure 53.2** The hazards of trying to stitch experimental results together to form a theory



- 2 Even if a treatment  $X$  has no average treatment effect on a variable  $M$ , and  $M$  has no average effect on  $Y$ ,  $M$  may mediate the relationship between  $X$  and  $Y$  for every unit. Such a situation could arise, as in the lower panel of Figure 53.2, if the effects of  $X$  on  $M$  and of  $M$  on  $Y$  have opposite signs in two subpopulations and therefore offset each other.

The broader take away is that there are limits to the extent to which experiments enable us to learn about causal models from empirical observation alone. Even if we were able to randomize all variables in a model, we would not be able to recover many causal quantities of theoretical relevance.

### **Strategy 5: Use Experiments to Put Theories to the Test**

Rather than seeking to *generate* theories, experimentation could focus on ‘selecting’ theories. Indeed, for many, the falsification of theories is what science is all about – at least in principle (Lakatos, 1970).

Experiments are useful for theory testing because they allow for valid statistical tests of hypotheses about causal effects. Of course, testable implications of theories can take many forms. Our theory about bargaining, for example, implies a correlation between the number of bargaining rounds and whether a customer moves first once we condition on the fare that the customer has to pay.<sup>8</sup> One advantage of implied causal relationships, however, is that they are often consistent with a smaller set of alternative theories. For example, imagine an alternative model according to which the number of bargaining rounds has a causal effect on whether the customer moves first. Both our model and the alternative model are consistent with the finding that the number of bargaining rounds and the customer’s first-mover status are correlated conditional on the taxi fare, but only one of them is consistent with the finding that the number of bargaining rounds causes the customer to move first. In short, being able to

test hypotheses about causal effects is essential for our ability to empirically distinguish causal models.

Perhaps the most common way of using field experiments for theory testing proceeds as follows:

- 1 Derive claims about marginal effects from a causal model, for example: ‘Making the first offer has a non-zero (positive or negative) effect on the price that the customer has to pay’.<sup>9</sup>
- 2 Design and implement an experiment to test a specified null hypothesis that is inconsistent with the claim (such as: making the first offer has a zero effect on prices).<sup>10</sup>

Recent work has pushed further on what can be done with testing. In particular, work in the tradition of Fisher (1935) and Rosenbaum (2002; 2010) has explored the potential of using experiments to test a set of more elaborate causal models against each other (see Bowers et al., 2013). More elaborate here means that these models specify a functional relationship for treatment and outcome for every unit conditional on one or more model parameters. For example, one of the simplest such models entails that going first adds the same constant  $\tau$  to the taxi price for all customers. A more complex model may specify that the taxi price is given by  $\pi_i = \alpha_i + \tau'z_i + \beta s_i$ , where  $\tau'$  is a constant effect of going first,  $s_i$  is the number of other customers in the same shared taxi who go first,  $\beta$  is a constant marginal effect of each additional first mover in a taxi and  $\alpha_i$  is the price that the customer pays when she and all other taxi passengers go second. Fisherian inference proceeds by hypothesizing specific values for the parameters in the model (e.g.  $\tau' = -0.2$  and  $\beta = -0.1$ ) and testing the null hypothesis that the assumed model and parameter values are correct. The same test is performed for a grid of parameter vectors for any given model and, also, for different models. For example, one could compare the  $p$ -value associated with a constant additive treatment effect model with  $\tau = -0.5$  to the

$p$ -value associated with the more complex model with  $\tau' = -0.2$  and  $\beta = -0.1$ .

In a recent application, Ichino et al. (2013) use this approach to test two different agent-based models of how party agents who aim to rig the voter registration process react to the placement of observers at randomly selected voter registration centers. They demonstrate how the close connection between the causal model of interest and the null hypotheses that are being tested facilitates learning about the causal model. They also show, however, that the wholesale rejection of one model for another is often impossible, since different models can imply the same distribution of outcomes for some values of their respective parameters.

### **Strategy 6: Use Experiments to Estimate Structural Models**

An approach that has been the norm in economics in the past (Heckman, 1991) and seems in the process of making a comeback (DellaVigna, 2018) is to use field experiments for the *estimation* of structural models. In the section ‘Illustration of a Parametric Structural Model Connecting Theory and Experimentation’, we demonstrate some of the advantages and pitfalls of this approach using our stylized example theory of bargaining. Here, we highlight the main points.

1 *Building a structural model*: the basis for structural estimation is typically a decision-theoretic or game-theoretic model from which the equations that link exogenous and endogenous variables in the model can be derived. A key step towards the specification of a model that can be estimated is the modeling of heterogeneity. For example, below we derive the taxi fare that a customer pays as a deterministic function of three variables (the customer’s behavioral type, the number of possible bargaining rounds and whether the customer gets to make the first offer) and of the customer’s discount factor, a parameter that we seek to estimate. Naturally, it seems unrealistic

that every subject in a real-world experiment would behave exactly in accordance with these functions. The DAG in Figure 53.1 represents this idea through the arrow that points from the unobserved variable  $U_3$  into  $\pi$ . Structural models explicitly incorporate such heterogeneity, typically by allowing for random shocks to the utility of players, for random implementation errors or for heterogeneity in model parameters (DellaVigna, 2018). Usually, these models assume that this heterogeneity follows a particular distribution.

- 2 *Identification*: in order to estimate them unambiguously, the parameters of a structural model need to be identified, i.e. the distribution of data should only be implied by one set of parameter values. If more than one set of parameter values can generate the same distribution of data, then we cannot distinguish true and false parameters from each other even with infinite data and even if the true data generation process is well captured by the assumed model. Sometimes, additional assumptions about functional forms or distributions are required for the purpose of identification.
- 3 *Estimation and extrapolation*: the estimation of the model parameters can be performed using one of various methods including Maximum Likelihood Estimation (MLE), Generalized Method of Moments (GMM), or Bayesian approaches. A key advantage of structural estimation is that the resulting estimates, together with the model, can subsequently be used for various extrapolations (e.g. towards the effects of the same treatment in other settings or the effects of different treatments) that go beyond what could be learned from experimental data alone.
- 4 *Theory dependence and cross-validation*: whether such inferences will be misleading depends, of course, on the extent to which the model itself is a good approximation to reality. Crucial to the extrapolation of results to other settings is often the assumption that the parameters of a model are ‘structural’ in the sense that they do not vary across settings (Acemoglu, 2010). For example, we estimate a discount factor which captures the extent to which individuals value the future relative to the present. In order to predict treatment effects for other settings, we need to assume that the players in these other settings have the same discount factor. Yet, individuals may not necessarily place the same value on the future in all contexts.

In response to such concerns, researchers can assess the sensitivity of their estimates to alternative assumptions and cross-validate their models in various ways. If we believe that certain parameters are indeed structural, one way to validate a model is to compare the resulting estimates of, say, the discount factor to estimates of the same parameters from other studies. Another possibility is to make predictions for other settings and compare those to actual data from these settings. See Martinez et al. (2017) and DellaVigna et al. (2017) for examples. In the section 'Illustration of a Parametric Structural Model Connecting Theory and Experimentation', we illustrate how the ability of experiments to obtain unbiased estimates of average treatment effects can sometimes be helpful for cross-validating a structural model.

## SIX STRATEGIES TO ADDRESS EXTERNAL VALIDITY CONCERNS

Below, we review six strategies that researchers have used to address the issue of external validity. While some of these approaches treat external validity as an empirical question, most of them draw, in some way or the other, on prior theoretical knowledge to determine whether and how the results from an experiment can be generalized.

### ***Strategy 1: Push Back – Researchers Should Focus on Sample Effects***

The first response is to ignore concerns over external validity and limit claims to sample effects. One justification for this, touched on already, is that empirical research in social science should be about testing general theoretical propositions and not about estimating effects. As long as a proposition applies to a sample, results from the sample can be declared consistent or inconsistent with the proposition even if they are not in other ways representative of a population.

### ***Strategy 2: Claim a Bellwether***

A second response is to assert that the case studied is in some way *especially* informative for other cases. Researchers often claim that their experiment is 'ideal' in some sense. This might mean that it is typical in some way, or that it is atypical in an informative way.

Say that we choose to run an experiment at a site that has some value on background variable  $X$ , and that we have good reason to believe that the causal effect of interest is decreasing in  $X$ . Then:

- 1 finding a positive effect in a location with a high value of  $X$  is informative for the claim that effects are in general positive in the target population;
- 2 finding a negative effect in a location with a high value of  $X$  is not very informative for the claim that effects are in general negative in the target population;
- 3 finding a positive effect in a location with a low value of  $X$  is not very informative for the claim that effects are in general positive in the target population;
- 4 finding a negative effect in a location with a low value of  $X$  is informative for the claim that effects are in general negative in the target population;
- 5 finding a positive (negative) effect in a location with a typical (e.g. modal) value of  $X$  is informative for the claim that effects are in general positive (negative).

Claims 1 and 4 are strong because the case is atypical. Claims 2 and 3 are weak because the case is atypical. Claim 5 is strong because the case is typical.

$X$  here could be any characteristic of the experiment including properties of the treatment. For example, we may expect treatment effects to increase in the intensity of the intervention or decrease in its scale.

Critically, claiming inference from unusualness (or typicality) depends on prior beliefs about the distribution of effects as a function of selection criteria, here  $X$ . Justification, or at least articulation, of these beliefs is needed to assess these claims. Pre-existing theoretical knowledge can help in this regard. A more

empirical approach might be to elicit beliefs (e.g. from prospective research consumers) about the *ranking* of effects across a set of cases, including the case at hand.

Note that, given these considerations, the *ex post* informativeness of an experiment is not the same as its *ex ante* informativeness. You may not be wise spending resources to search for a positive treatment effect in a setting with a high value of  $X$ , for example, since the chances of finding such an effect are small. But if you find it, such a result is highly informative.<sup>11</sup>

### Strategy 3: Exploit Variation within Studies

A third approach is to use variation in effects *within* a study site to justify claims that go beyond the study site.

The simplest approach is to identify relevant dimensions along which the study site differs from the target site and to demonstrate that there is no heterogeneity in effects along these dimensions within the study at hand.

If there is heterogeneity along such dimensions, one possibility is to use weighting or propensity score subclassification estimators to estimate effects for the target population using variation in the study population. See Kern et al. (2016) for an assessment of several such approaches.

The general idea is to identify a set of strata across which effects vary, estimate effects for each of these strata in the study population, and take a weighted average of the resulting estimates where the weights correspond to the share of subjects in each stratum in the target population.

For intuition, say you undertake a study in location  $A$ . You want to make a claim for target site  $B$ . In site  $A$ , one third of subjects are young and two thirds are old. In site  $B$ , it is the reverse. The strategy is to estimate the effect separately for young and old subjects in  $A$  and then calculate a weighted average of these group-specific estimates using the proportion

of young and old subjects in site  $B$ . Let  $\hat{\tau}_O^A$  and  $\hat{\tau}_Y^A$  be the respective treatment effect estimates among old and young people in setting  $A$ . Suppose we learn from study  $A$  that  $\hat{\tau}_Y^A = 1$  and  $\hat{\tau}_O^A = \frac{1}{4}$ . The overall treatment effect estimate in  $A$  is given by  $\hat{\tau}^A = \frac{1}{3} \times 1 + \frac{2}{3} \times \frac{1}{4} = \frac{1}{2}$ . For site  $B$ , we estimate the treatment effect to be  $\hat{\tau}_B = \frac{2}{3} \times 1 + \frac{1}{3} \times \frac{1}{4} = \frac{3}{4}$ .

This approach depends on many assumptions, formalized in Pearl and Bareinboim (2014) and Tipton (2013). Most obviously the support of  $B$  must be a subset of the support of  $A$  – i.e. for any stratum in  $B$  for which one wants to generate estimates there should be a corresponding stratum in  $A$ . You cannot estimate effects for, say, a co-ed school based on estimates obtained from an experiment in a boys-only school.

More substantively, you need to be willing to assume that the distribution of effects within each stratum is the same in the two populations. This claim is most easily made when the study sample is itself a random draw from the study population. Yet, even in such an ideal case, one has to worry about how spillovers work not just in the study population but also in the target population (Tipton, 2013).

Finally, this approach has implications for how one can design the study sample in  $A$  to facilitate inference to target  $B$  (see Tipton and Peck, 2017).

### Strategy 4: Exploit Variation across Studies

A fourth strategy is to design studies so that they can feed into meta-analyses that seek to make broader claims. A common approach is to think of a superpopulation of cases with effects drawn from a common distribution:

$$\tau_j \sim f(\mu, \sigma)$$

The interest is in learning about the expectation of the superpopulation effects,  $\mu$ , but also

the variation in effects  $\sigma$ . A single study drawn randomly from a population can give an unbiased, but noisy, estimate for  $\mu$  and says nothing about  $\sigma$ . Multiple comparable studies can provide tighter estimates of both  $\mu$  and  $\sigma$ .

Here, the external gains from a study operate through complementarities with other studies. Yet, up until recently, only few topics in political science had generated a large enough set of comparable experiments to allow for a meta-analysis. Initiatives that encourage the co-ordinated implementation of experiments on the same topic in various contexts – such as the ‘Metaketa’ projects – seek to address this problem (Dunning et al., 2018).

Meta-analyses can also include a more systematic analysis of treatment effect heterogeneity across studies. Vivaldi (2019), for example, finds that experiments implemented by governments tend to have smaller effects than those implemented by non-governmental actors. Ultimately, such results can become useful for theory development.

### **Strategy 5: Cross Validation**

A fifth approach used in Dunning et al. (2018) and Coppock et al., (2018) treats external validity as an empirical question.<sup>12</sup> Dunning et al. (2018) ask: do research consumers *in fact* update inferences for out-of-sample estimands? They gather results from multiple related experiments and assess whether exposing research consumers to the findings makes them update their beliefs about effects in studies they have not been informed about, and whether updating goes in the right direction. Coppock et al. (2018) assess empirically whether the results from online samples are *in fact* consistent with what we know from representative samples (and vice versa). They find that they are and attribute this to low effect heterogeneity. One could engage in a similar kind of exercise using a single study by analyzing whether results estimated on a non-random subsample correspond to those in the rest of the

study. Results will depend on the degree of effect heterogeneity.

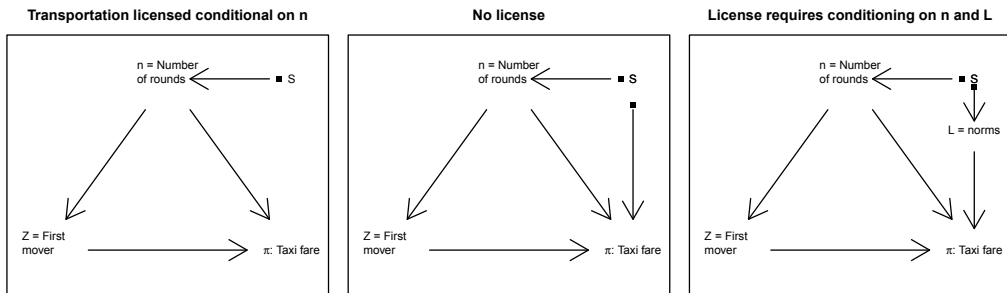
This empirical approach gives, in some sense, an iron clad answer to the question of external validity. Yet, it has an obvious shortcoming: *checking* external validity requires already knowing the target estimand. One can make external claims, and check whether the claims are correct, but this establishes the claim to external validity only in cases in which it is not needed. Put differently, you do not know whether the claim for external validity itself is externally valid outside of the test set.

### **Strategy 6: Formal Transportation**

The last approach we consider uses causal models to formally justify – or ‘license’ – claims to external validity. This is an example of the use of theory discussed in the section ‘Strategy 2: Use Theory as a Helper for Inference’. Pearl and Bareinboim (2014) develop a framework in which researchers provide a causal model and the associated DAG for their study population and then represent the ways in which the target population differs from the study population as a set of ‘selection’ nodes that are the origins of arrows that point into nodes in the original DAG. This representation of differences makes it possible to assess whether there is a weighting strategy that allows inference to the target population. The assessment is, of course, conditional on the model.

To illustrate, imagine first that two sites share a common causal model relating the taxi fare, the number of bargaining rounds, and whether the customer makes the first offer. Say, for instance, data on taxi bargaining is generated in Kenya and one wants to make inferences to Somalia. Figure 53.3 displays the corresponding DAG.

We are interested in the effects of the customer being the first mover on the taxi price. We know that the average effect of moving first depends on how long bargaining can



**Figure 53.3** Three selection graphs

continue before negotiations break down. In the first graph in Figure 53.3, the selection variable ‘ $S$ ’ characterizes the differences between Kenya and Somalia suggesting that there is a difference in bargaining rounds between these sites. Perhaps the taxi market is more competitive in Kenya and, as a consequence, drivers and customers are more likely to separate if they cannot reach an agreement after a small number of offers and counter-offers. If this is an accurate characterization of the differences between these sites, then results can be transported using a strategy much like that described in ‘Strategy 3: Exploit Variation Within Studies’: re-weight the estimates from Kenya using information on propensities for bargaining to break down after a specific number of rounds for a given customer-driver interaction. The ability to use this strategy requires – beyond the model being right – (a) that the range of bargaining rounds in Somalia is also present in Kenya, and (b) that we have data about the distribution of this variable in both contexts. There are thus gains from implementing experiments in places with wide variation and in gathering data about distributions of variables out of sample.

The implications of this framework extend further, however. Say that the differences between Kenya and Somalia are those depicted in panel 2 (Figure 53.3). In this case, these differences extend to how order of play and bargaining protocol affect outcomes.

Average effects could be different in these sites which would prevent extrapolation even with complete data. In graph theoretic terms, the reason is that there is no set of variables that you can condition on that ‘ $d$ -separates’ the selection variable from the outcome.

Say, finally, that we have a model that tries to explain the differences in effects via some specified intermediate variable, as in the third panel. The selection graph in the third panel can be interpreted as saying that the (conditional) effect of moving first on the taxi price is different in Somalia because there are different norms in Somalia that moderate this effect. If so, we can now get good estimates of effects in Somalia by conditioning not just on the likelihood of bargaining breaking down after a specific number of rounds but also on a measure of norms. We re-weight by finer strata. Formally, conditioning on norms now separates the selection node  $S$  from the outcome.

This logic captures the core elements of Theorem 2 in Pearl and Bareinboim (2014): Let  $Z$  denote a stratum. The strata-specific causal effect of  $X$  on  $Y$  is transportable from one graph to another if conditioning on  $Z$  produces independence between the outcome  $Y$  and the set of selection nodes  $S$ .

The gains from this framework are two-fold. First, it becomes possible to state justifications for transportation in terms of independence relations between variables, which are statements that can then be

assessed empirically. Second, given a justification, there is clarity on the set of variables for which data should be gathered to calculate stratum level causal effects.

That said, to our knowledge, the framework cannot be used to extrapolate from findings on the effects of one treatment to the effects of another, which, in principle, is possible with parametric models (see sections ‘Strategy 6: Use Experiments to Estimate Structural Models’ and ‘Illustration of a Parametric Structural Model Connecting Theory and Experimentation’).

As far as we know, the Pearl and Bareinboim (2014) framework has not yet been used within political science.

### **ILLUSTRATION OF A PARAMETRIC STRUCTURAL MODEL CONNECTING THEORY AND EXPERIMENTATION**

We now return to our example experiment on taxi bargaining and use it to walk through the logic of estimating a simple parametric structural model. Recall that, in this example, we are interested in why some taxi customers end up paying more than others. According to our Level 1 theory illustrated in Figure 53.1, there are three variables that directly affect our outcome of interest: the number of rounds for which bargaining can continue, whether the customer gets to make the first offer, and the customer’s type. We begin by further developing this theory into a Level 2 theory in the form of a parametric structural model. We then illustrate how we would turn our Level 2 theory into a Level 3 theory and estimate it using simulated data. Finally, we highlight four possible benefits of the approach.

#### ***Setting Up and Estimating a Structural Causal Model***

We will assume that the taxi bargaining process can be captured by a standard alternating

offers bargaining game with complete information (Rubinstein, 1982). A customer and a driver take turns in making offers of how to divide a pie. We think of the pie as the unit endowment that we provided to the customer. For simplicity, we assume that the size of the pie is known to both players. Whoever moves first makes a suggestion on how to divide the pie. For example, if the customer moves first, she may say ‘I will pay you 0.2 and keep 0.8 for myself’. The second player decides whether to accept or reject this offer. If the second player accepts, each player receives the share of the pie that the offer allocated to her (e.g. the customer pays 0.2 to the driver and keeps 0.8 for herself). If the second player rejects, the game moves to the next round and the second player gets to make an offer which the first player can accept or reject. If no agreement is reached, bargaining ends after  $n$  rounds and both players receive a payoff of 0. We might imagine, for example, that customers have to pay the endowment back to us if they do not secure a ride.

To capture preferences about time, we assume that players discount their payoffs at a rate of  $0 \leq \delta \leq 1$ . If they reach an agreement in round 1 and a player receives an amount  $x$ , the player will value this amount at  $x$ . If the same agreement is reached in period 2, the player will value the amount at  $\delta x$ ; in period 3 she will value the amount at  $\delta^2 x$  etc.

The standard solution for rational players is found via backwards induction for a finite number of bargaining rounds  $n$ . Suppose that  $n = 1$ , i.e. the player who goes first gets to make a take-it-or-leave-it offer. Since both players know that there will be no second round, the first mover will take the entire pie for herself. The second player will accept, since she will receive a payoff of 0 anyway, irrespective of what she does. Let’s consider  $n = 2$ , i.e. there can be one offer and a counter-offer. We already know that, in the second round, the second-mover will be able to take the whole pie of 1 for herself (again, the other player will receive 0 irrespective of whether she accepts or rejects the offer).

Seen from the first period, the second mover knows that she can achieve  $\delta \times 1$  by rejecting the first mover's offer. In the first period, the first mover will therefore offer exactly  $\delta$  to the second mover and the second mover will accept. Letting  $\pi_n^j$  denote the equilibrium share paid by player  $j$  in an  $n$ -round game, the first mover in a two-round game gives  $\pi_2^1 = \delta$  to the second mover and keeps  $1 - \pi_2^1 = 1 - \delta$  for herself. The same logic can be applied to  $n = 3, n = 4$  etc.<sup>13</sup>

In the infinite version of the game, the optimal solution involves offering an amount

so that the receiver is indifferent between accepting and moving on to the next round of the infinite game in which she would offer the same amount; i.e. we seek an offer  $\pi_\infty^1$  such that  $\pi_\infty^1 = \delta(1 - \pi_\infty^1)$  which implies  $\pi_\infty^1 = \delta / (1 + \delta)$ .

Note that, in this model, agreement is always reached in the first period irrespective of the number of possible bargaining rounds  $n$ .

The top panel of Table 53.1 summarizes the price that a rational customer should pay according to this model depending on the

**Table 53.1 Equilibrium prices for first ( $\pi_n^1$ ) and second ( $\pi_n^2$ ) moving customers and average treatment effect ( $\tau_n$ ) of the customer moving first for games with  $n$  possible rounds**

Rational customers			
	$\pi_n^1$	$\pi_n^2$	$\tau_n = \pi_n^1 - \pi_n^2$
$n = 1$	0	1	-1
$n = 2$	$\delta$	$1 - \delta$	$2\delta - 1$
$n = 3$	$\delta(1 - \delta)$	$1 - \delta(1 - \delta)$	$2\delta(1 - \delta) - 1$
$n = \infty$	$\frac{\delta}{1 + \delta}$	$1 - \frac{\delta}{1 + \delta}$	$2 \frac{\delta}{1 + \delta} - 1$
Behavioral customers			
	$\pi_n^1$	$\pi_n^2$	$\tau_n = \pi_n^1 - \pi_n^2$
$n = 1$	$\frac{3}{4}$	$\frac{3}{4}$	0
$n = 2$	$\frac{3}{4}$	$\frac{3}{4}$	0
$n = 3$	$\frac{3}{4}$	$\frac{3}{4}$	0
$n = \infty$	$\frac{3}{4}$	$\frac{3}{4}$	0
Population with share $q$ of behavioral customers			
	$E(\pi_n^1)$	$E(\pi_n^2)$	$\tau_n = E(\pi_n^1 - \pi_n^2)$
$n = 1$	$q \frac{3}{4}$	$q \frac{3}{4} + (1 - q)$	$-(1 - q)$
$n = 2$	$q \frac{3}{4} + (1 - q)\delta$	$q \frac{3}{4} + (1 - q)(1 - \delta)$	$(1 - q)(2\delta - 1)$
$n = 3$	$q \frac{3}{4} + (1 - q)\delta(1 - \delta)$	$q \frac{3}{4} + (1 - q)(1 - \delta(1 - \delta))$	$(1 - q)(2\delta(1 - \delta) - 1)$
$n = \infty$	$q \frac{3}{4} + (1 - q) \frac{\delta}{1 + \delta}$	$q \frac{3}{4} + (1 - q) \left(1 - \frac{\delta}{1 + \delta}\right)$	$(1 - q) \left(2 \frac{\delta}{1 + \delta} - 1\right)$



number of rounds  $n$  and on whether the customer moves first ( $\pi_n^1$ ) or second ( $\pi_n^2$ ).

We are interested in the effect,  $\tau_n$ , of being able to make the first offer on the price paid in an  $n$  round game. This effect simply equals the difference between  $\pi_n^1$  and  $\pi_n^2$  (see the last column of Table 53.1).

So far, we have only considered the behavior of rational players. Yet, according to our causal model, customers – though not drivers – can be of different types. In this stylized example, we assume that there are some norm following customers who always pay  $\frac{3}{4}$  of the pie to the driver and keep  $\frac{1}{4}$  for themselves, irrespective of the number of bargaining rounds or whether they go first or second. As can be seen in Table 53.1, the treatment effect of going first on the price paid is always 0 for norm following customers. We assume that a customer's behavioral type is not observed by the researchers and that the share of norm followers in the population of customers is  $q$ .

Ultimately, we are interested in the average treatment effect of the customer going first on the taxi fare. In order to make predictions about this effect, we need to add estimates of  $q$ , the distribution of rational and norm following types to our Level 2 model. As displayed at the bottom of Table 53.1, the population-level average of the price paid by customers who move first or second is just a weighted average of the price paid by rational customers and the price paid by norm following customers. Accordingly, since the average treatment effect among norm followers is zero, the average treatment effect in the population is just the proportion of rational customers  $(1 - q)$  times the predicted treatment effect among rational customers.

Subsequently, we focus on  $n = 2$  and  $n = \infty$ . Note that the model produces interesting heterogeneity: from the table we see that  $\tau_2 > \tau_\infty$  for  $\delta > 0$ ,<sup>14</sup> though  $\tau_2$  and  $\tau_\infty$  may differ in sign and  $\tau_\infty$  may be larger in absolute value. In the extreme case when  $\delta$  is close to 1,  $\tau_2$  is close to  $(1 - q)$  and  $\tau_\infty$  is close to 0.

### *Taking the model to data*

Our Level 2 model makes predictions about average taxi fares as a function of whether the customer moves first, the number of bargaining rounds and two model parameters, the discount factor  $\delta$ , and the share  $q$  of norm followers in the population. Suppose we have run our taxi bargaining experiment that randomly assigns customers to make the first offer in a place where  $n = 2$  or  $n = \infty$ , how can we use these data to get estimates of the model parameters,  $\delta$  and  $q$  and estimates of treatment effects  $\tau_n$ ?

We do so by using the model's equilibrium predictions to motivate a data generating process that describes the probability of observing any data given a set of parameter values. We then rely on MLE to find estimates of  $\delta$  and  $q$ . The same could be achieved through various other estimation methods, including GMM or Bayesian approaches.

An obvious challenge in using real world data to estimate the parameters of a model is that the real world might produce data that are inconsistent with an overly simple model. For this reason, it is generally necessary to allow for a stochastic component that can render all data *possible*, even if improbable. In the DAG in Figure 53.1, the idea that observed taxi fares may deviate from the predictions in Table 53.1 is captured by the node  $U_3$ . We can think of  $U_3$  as representing factors other than the number of bargaining rounds, whether the customer moves first and the customer's behavioral type that also affect the price a customer pays.

What could be sources of  $U_3$  in the context of our model? There are multiple possibilities here. In his review of structural estimation in behavioral economics, DellaVigna (2018) discusses the three most common ways in which researchers incorporate heterogeneity in their models. First, it is sometimes assumed that the actors in the model receive an unobserved utility shock which generates heterogeneity in how they behave. Second, we can imagine that there is heterogeneity

in some of the parameters of the model. Imagine, for example, that instead of a fixed number,  $\delta$  is a random variable that follows, say, a beta distribution in the population of interest. Individuals would then be heterogeneous in their  $\delta$  which, in turn, would result in heterogeneity in the way in which the customer's endowment is divided. A final possibility is to assume that individuals make implementation errors.

The last interpretation fits well with our example. Imagine that players attempt to act in accordance with the model but make random mistakes when dividing the endowment. As a consequence, the prices that get paid are not always the ones predicted by the model.

To estimate our model via MLE, we need to be specific about the distribution of these implementation errors. Here, we assume that the price paid by a given customer is a draw from a beta distribution that is centered on whatever price the model predicts for this customer. The beta distribution is a natural choice in this case, since it is defined on the interval  $[0,1]$ , just like the shares of the endowment that customers pay as a result of the taxi bargaining process. The distribution has two parameters,  $\alpha$  and  $\beta$ , that control its

shape. The mean of the beta distribution can be expressed as a function of these parameters,  $\mu = \frac{\alpha}{\alpha + \beta}$ . In turn, we can write the two parameters as  $\alpha = \kappa\mu$  and  $\beta = \kappa(1 - \mu)$ . In this way, we can model  $\mu$ , the mean of the distribution, as a function of the variables in our model. Specifically,  $\mu$  will depend on:

- customer  $i$ 's treatment status  $z_i$ , where  $z_i = 1$  if the customer has been randomly assigned to go first and  $z_i = 0$  if the customer has been randomly assigned to go second;
- customer  $i$ 's (unobserved) behavioral type  $\theta_i$ , where  $\theta_i = 1$  if the customer is rational and  $\theta_i = 0$  if the customer is a norm follower; and
- the number of rounds  $n \in \{2, \infty\}$  for which bargaining can continue.

Thus,  $\kappa$  enters the model as a new parameter that describes the variance of the distribution of prices but not its mean. It is of some substantive interest in that it captures how close behavior is to the Level 2 model predictions.

Using this parameterization and the predictions of the model, we can write down the likelihood given the observed prices paid by  $N$  customers in our experimental subject pool (where we use subscript  $i$  to denote individuals):

$$\Lambda(q, \delta, \kappa | \pi) = \prod_{i=1}^N \left( q \text{Beta} \left( \pi_i \mid \frac{3}{4}\kappa, \frac{1}{4}\kappa \right) + (1-q) \text{Beta}(\pi_i \mid \mu_i\kappa, (1-\mu_i)\kappa) \right)$$

$$\mu_i = \begin{cases} z_i\delta + (1-z_i)(1-\delta) & \text{for } n=2 \\ z_i \frac{\delta}{1+\delta} + (1-z_i) \left( 1 - \frac{\delta}{1+\delta} \right) & \text{for } n=\infty \end{cases} \tag{1}$$

$z_i = 1$  with probability .5 for all  $i$

A couple of features are noteworthy: first, the assumed data generating process that underlies the likelihood function reflects that customers are randomly assigned to go first; hence  $z_i = 1$  with probability  $\frac{1}{2}$  for all units. Second, we do not need to know a customer's

behavioral type  $\theta_i$  in order to compute the likelihood, which is fortunate as we do not observe this characteristic. Instead, the joint probability distribution is simply a mixture between a beta distribution with mean  $\frac{3}{4}$  and a beta distribution with mean equal to

the predicted price paid by rational customers, conditional on whether they go first or second. The weight on each of these distributions is given by the share of norm followers,  $q$ .

**Estimation and inferences**

Now that we have equation (1) in hand, we use standard MLE approaches to find the values for  $q$ ,  $\delta$  and  $\kappa$  that maximize the likelihood of the data we observe in our experiment.

Once our analysis is set up, it is easy to inquire into how well this procedure works. Suppose our model is correct. We can find out how well we recover the model parameters using the following steps: (a) simulate data using draws from the data generating process described above; (b) perform the estimation; and (c) compare the estimates to the parameter values assumed during the simulation. By repeating these steps many times, we can figure out whether our estimates are, on average, correct and how variable they are.

In practice, we use the R package `DeclareDesign` for this exercise (Blair et al., 2018). We provide all code for the declaration and diagnosis of this model in supplementary material.

The simulations involve (a) imagining that we have run an experiment that randomly assigns taxi customers to make the first offer, either in a setting where bargaining can continue for two rounds ( $n = 2$ ) or in a setting where bargaining can continue indefinitely ( $n = \infty$ ); (b) using a maximum likelihood estimator to estimate  $\delta$ ,  $\kappa$  and  $q$  in either setting; and (c) drawing inferences regarding treatment effects using the parameter estimates. This last step is done by referring to the last column of Table 53.1. For example, with estimates of  $\hat{q} = 1/2$  and  $\hat{\delta} = 4/5$ , we would predict a treatment

$$\text{effect of } \hat{\tau}_\infty = \frac{1}{2} \left( 2 \frac{\frac{4}{5}}{1 + \frac{4}{5}} - 1 \right) = -\frac{1}{18} \approx -0.06$$

for a setting where bargaining can continue indefinitely.

For each estimand-estimator pair, we then report the expected bias (which simply equals the difference between the estimand and the mean estimate across all simulations).

In Table 53.2, we report the estimates of model parameters and in Table 53.3, we report estimates of average treatment effects, comparing those generated from parameter estimation to those generated using a simple difference-in-means estimator. In each case  $MLE_2$  indicates estimates obtained from the maximum likelihood estimators that take data from a 2 period game and  $MLE_\infty$  indicates estimates obtained from the maximum likelihood estimators that take data from an  $n = \infty$  period game. Correspondingly,  $DIM_n$  indicates estimates obtained using the difference-in-means estimator and data generated in a setting with  $n$  possible bargaining rounds.

We recover unbiased estimates of  $q$  and  $\delta$ , irrespective of whether we use data from a setting with  $n = 2$  or  $n = \infty$ . The estimates of  $\kappa$  are slightly biased. Moreover, our estimates of model parameters allow us to

**Table 53.2 Estimation of model parameters using the correct model**

Estimand	Estimator	Estimand value	Bias
$\delta$	$MLE_2$	0.80	-0.00
$\delta$	$MLE_\infty$	0.80	0.00
$\kappa$	$MLE_2$	6.00	0.04
$\kappa$	$MLE_\infty$	6.00	0.03
$q$	$MLE_2$	0.50	-0.00
$q$	$MLE_\infty$	0.50	-0.00

**Table 53.3 Estimation of average treatment effects using difference-in-means (DIM) and parameter estimation (MLE)**

Estimand	Estimator	Estimand value	Bias
$\tau_2$	$MLE_2$	0.30	-0.00
$\tau_2$	$DIM_2$	0.30	0.00
$\tau_\infty$	$MLE_\infty$	-0.06	-0.00
$\tau_\infty$	$DIM_\infty$	-0.06	-0.00

recover unbiased estimates of average treatment effects.

Why, in this case, are we able to obtain unbiased estimates of the model parameters of interest? To see the intuition, recall that, according to our model (see Table 53.1), the average prices paid by first and second moving customers in a setting with  $n = 2$  are given by the following functions:  $E(\pi_2^1) = q\frac{3}{4} + (1-q)\delta$  and  $E(\pi_2^2) = q\frac{3}{4} + (1-q)(1-\delta)$ . It turns out that, if we knew the values of  $E(\pi_2^1)$  and  $E(\pi_2^2)$ , these two equations could, in most cases, be solved for the unique values of  $q$  and  $\delta$ . For example, if  $E(\pi_2^1) = \frac{1}{2}$  and  $E(\pi_2^2) = \frac{3}{4}$ , then we know that  $q$  must equal  $\frac{1}{2}$  and  $\delta$  must equal  $\frac{1}{4}$ .<sup>15</sup> An experiment that randomly assigns customers to go first or second provides us with unbiased estimates of  $E(\pi_2^1)$  and  $E(\pi_2^2)$ . As a consequence, we are able to recover estimates of  $q$  and  $\delta$ .

Even though this result seems encouraging, one may wonder why we would want to estimate the parameters of our model in the first place. After all, as can be seen in Table 53.3, using a simple difference-in-means estimator yields unbiased estimates of the average treatment effect of moving first irrespective of whether we perform our experiment in a context with  $n = 2$  or  $n = \infty$ . What more can we learn from this exercise about why some taxi customers pay higher prices than others?

In the next four sections, we use this example to demonstrate what we see as some of the potential benefits of combining experimental data and structural estimation. The first section relates to the core argument of this chapter – that stronger theoretical connections can provide (or at a minimum, clarify) the bases for stronger external claims. The other three sections point to ways in which experiments can be strengthened through structural modeling and vice versa.

### **Benefit 1: Theory Allows for Answers to a Wider Set of Questions**

A key benefit of the structural model is that it allows us to answer a more varied array of questions than can typically be addressed by design-based inference, in particular questions regarding different settings, different treatments, and different outcomes.

To illustrate, our causal model suggests that the average treatment effect of moving first varies with the number of rounds for which bargaining can continue. This is confirmed in Table 53.3, where the effect estimate that we obtain in a setting with  $n = 2$  potential bargaining rounds differs from the estimate obtained when bargaining can continue indefinitely. In other words, the effect estimates do not travel across settings.

This is where the structural model may help. Imagine that we conduct our experiment in a setting where bargaining continues for  $n = 2$  rounds and we obtain estimates of  $q$  and  $\delta$ . We can now estimate the treatment effects for the  $n = \infty$  case by consulting the equations in Table 53.1. Thus, provided that our model is correct, we can use an experiment conducted in a single setting to predict effects for other settings where effects may be different.

Table 53.4 shows estimates based on this approach for our model.

Beyond generalization across settings, the model also helps us predict the effects of alternative treatments. The equations in Table 53.1 imply, for example, that the average treatment effect of a change that allows bargaining to continue for  $n = 3$  rounds instead of  $n = 2$  rounds on the average price

**Table 53.4 Extrapolation**

<i>Estimand</i>	<i>Estimator</i>	<i>Estimand value</i>	<i>Bias</i>
$\tau_2$	$MLE_2$	0.30	-0.00
$\tau_2$	$MLE_\infty$	0.30	0.00
$\tau_\infty$	$MLE_2$	-0.06	-0.00
$\tau_\infty$	$MLE_\infty$	-0.06	-0.00

paid by a customer who moves first is  $-(1 - q)\delta^2$ . Based on our estimates of  $\hat{q} = 1/2$  and  $\hat{\delta} = 4/5$ , we would thus predict that such a change reduces the price paid by a customer who makes the first offer by  $\frac{8}{25}$ .

Giving more structure to our model has considerably expanded the set of inferences that we would be able to draw from a single experiment. Of course, these inferences are only warranted if our structural causal model is the correct one. This may well be an unrealistic assumption. Among other things, it entails that the structural parameters  $\delta$  and  $q$  are indeed structural in the sense that they do not vary across settings. If the distribution of behavioral types varies across settings, for example, our predictions will not be correct. We will show below how reliance on the wrong model can lead to biased inferences and, more positively, how randomized experiments can sometimes draw attention to flaws in a model.

### ***Benefit 2: Theory Provides Pointers to Better Design***

Sticking for the moment with the assumption that our stylized model is correct, we can use the model to draw lessons for how to design our hypothetical experiment. This is an example of the use of theory described in the section ‘Strategy 3: Use Theory as a Helper for Design’.

Consider, for example, settings in which only take-it-or-leave-it offers are possible, i.e.  $n = 1$ . We can see from Table 53.1 that neither  $E(\pi_1^1)$  nor  $E(\pi_1^2)$  depends on  $\delta$ .

An experiment in a setting where  $n = 1$  would thus allow us to estimate  $q$  but not  $\delta$ . Without an estimate of  $\delta$ , we will not be able to generalize to other settings or treatments in the manner described above.

Similarly, not all equations for  $E(\pi_n^1)$  and  $E(\pi_n^2)$  can be easily solved for  $q$  and  $\delta$ . In the case of  $n = 3$ , for example, one set of estimates of  $E(\pi_3^1)$  and  $E(\pi_3^2)$  can be consistent

with two different solutions for  $q$  and  $\delta$ . In order to maximize the inferences that can be drawn, the experiment should be conducted in a setting that enables us to uniquely identify both  $q$  and  $\delta$ .

The model also has implications for case selection if we are interested in maximizing power. Suppose, for example, that we believe it likely that  $\delta$  is close to 1 and that the share of norm followers  $q$  is not very large prior to running our experiment. We would then expect  $|\tau_2| > |\tau_\infty|$ , which suggests that we will be better powered to detect a treatment effect if we run our experiment in a setting where bargaining must end after two rounds than in a setting where bargaining can continue indefinitely. While it may generally seem undesirable to let design considerations determine estimands rather than the other way around, greater power here may also imply the ability to obtain more precise estimates of the model’s structural parameters that may be used to generalize towards other estimands of interest.

### ***Benefit 3: Experimental Data Can Make It Possible to Improve Inferences from Theory***

We are used to the idea that randomization helps identify treatment effects (Gerber et al. 2014), but it can also play a key role in parameter identification. Instead of substituting for randomization, model-based inference, here, exploits randomization.

In our example, the assumption of randomization is built into the likelihood function and employing this model in a context without randomization could lead to biased parameter estimates. To illustrate, imagine that the probability of going first is related to a player’s behavioral type. Specifically, we imagine a data generating process according to which norm followers go first with probability  $p_n$  whereas rational players go first with probability  $p_r$ . We thus have a joint probability distribution over  $z$  and  $\pi$  – where  $z_i$

depends on the unknown type and  $\pi_i$  depends on both the unknown type and  $z_i$ .

We now re-do our simulations, changing the data generating process in this way (setting  $p_n = 0.8$  and  $p_r = 0.5$ ), but maintaining the estimation strategy, i.e. we still choose  $q$ ,  $\delta$ , and  $\kappa$  to maximize the likelihood of our observed data according to equation (1). Thus, we (incorrectly) analyze the data *as if* customers were randomly assigned to go first or second.

Table 53.5 contains the resulting estimates of the model parameters and average treatment effects. We also include treatment effect estimates obtained from a difference-in-means estimator. In the absence of random assignment, all our estimates are biased.

Intuitively, without random assignment, we do not achieve unbiased estimates of  $\pi_n^1$  and  $\pi_n^2$ , since the relationship between the treatment (going first) and the outcome (the price that a customer pays) is confounded by an unobserved variable, the customer's type. Without being able to estimate  $\pi_n^1$  and  $\pi_n^2$ , however, we cannot uncover  $\tau_n$  and neither  $q$  nor  $\delta$ . Of course, this problem could be solved by conditioning on the customer's type if we could observe it. Alternatively, one might extend the model in an attempt to incorporate confounding.

**Table 53.5 Diagnosis of design without randomization**

<i>Estimand</i>	<i>Estimator</i>	<i>Estimand value</i>	<i>Bias</i>
$\tau_2$	$DIM_2$	0.30	0.11
$\tau_2$	$MLE_2$	0.30	0.11
$\tau_2$	$MLE_\infty$	0.30	0.12
$\tau_\infty$	$DIM_\infty$	-0.06	0.08
$\tau_\infty$	$MLE_2$	-0.06	-0.03
$\tau_\infty$	$MLE_\infty$	-0.06	0.04
$\delta$	$MLE_2$	0.80	-0.01
$\delta$	$MLE_\infty$	0.80	0.13
$\kappa$	$MLE_2$	6.00	-0.09
$\kappa$	$MLE_\infty$	6.00	-0.04
$q$	$MLE_2$	0.50	-0.22
$q$	$MLE_\infty$	0.50	0.01

In short, randomization can help identify parameters of interest, which can subsequently be used for extrapolation. This point gains importance in more complex models where identification of the parameters of interest becomes more difficult. See DellaVigna (2018) and Card et al. (2011) for more on how additional treatments can help identify parameters of a structural model.

#### ***Benefit 4: Experimental Data Can Help Improve Theory***

So far, we have assumed the model is right. But of course, we know the model is wrong (Box, 1976). The question is how to assess the consequences of relying on a model that is incorrect and how to react if the model is misleading. One way to do so is to take a model seriously, confront it with data, and then step back to see whether the theory did violence to the data. Poor fit can be suggestive of the need to improve a model (Browne and Cudeck, 1993; Gelman et al., 1996).

We illustrate by imagining that we had mistakenly not considered the possibility that there are norm following types and instead assumed that all customers acted rationally, i.e.  $q = 0$ . Essentially, this means ignoring the bottom two panels of Table 53.1 and considering the top panel only. Table 53.6 displays the results of changing our estimation strategy accordingly (we use  $MLE'$  rather than  $MLE$  to denote the new estimators).

The first thing to note is that, because we have returned to the scenario where we randomly assign a customer to make the first offer, the difference-in-means estimator recovers unbiased estimates of the average treatment effects of going first for both cases,  $n = 2$  and  $n = \infty$ . Our estimates of  $\delta$ , however, are biased irrespective of whether we rely on data from an experiment in the  $n = 2$  or  $n = \infty$  context. Moreover, we naturally do not obtain any estimates of  $q$ , since we

**Table 53.6 Diagnosis of design with incorrect model (assume  $q = 0$ )**

<i>Estimand</i>	<i>Estimator</i>	<i>Estimand value</i>	<i>Bias</i>
$\tau_2$	$DIM_2$	0.30	0.00
$\tau_2$	$MLE'_2$	0.30	-0.04
$\tau_2$	$MLE'_\infty$	0.30	0.52
$\tau_\infty$	$DIM_\infty$	-0.06	-0.00
$\tau_\infty$	$MLE'_2$	-0.06	-0.17
$\tau_\infty$	$MLE'_\infty$	-0.06	0.01
$\delta$	$MLE'_2$	0.80	-0.17
$\delta$	$MLE'_\infty$	0.80	0.11
$\kappa$	$MLE'_2$	6.00	-3.83
$\kappa$	$MLE'_\infty$	6.00	-3.05
$q$	$MLE'_2$	0.50	

assume  $q = 0$ . It is thus not surprising that our predictions for treatment effects in other settings based on our estimates of  $\delta$  are severely biased. Yet, we do not only obtain biased estimates of the treatment effect in a *different* context. Our prediction of the treatment effect for the *same* context is also biased. Based on our estimate of  $\delta$  obtained from an experiment run in a place where  $n = 2$ , for example, we would predict a treatment effect of 0.26 for this same place. Yet, the actual treatment effect, which we recover using the difference-in-means estimator, is 0.3.

This pattern shows how random assignment can help us detect problems with our theoretical model. Knowing that we have an unbiased estimate of the average treatment effect allows us to use this estimate to validate the prediction of our theoretical model. The larger the gap between our experimental estimates and our predictions, the weaker our confidence that our theoretical model is correct.

A similar strategy is used by Todd and Wolpin (2006) who evaluate the effects of a randomized school subsidy program in Mexico. They fit a structural model using

households who did not receive the subsidy and validate the model by comparing the predicted effect of the subsidy program to the experimental estimates. Similarly, DellaVigna et al. (2016) use not only new, but also existing, experimental results to validate their model. Specifically, the authors develop a structural model of voting based on the idea that individuals derive pride from telling others that they voted or face costs when lying about whether they voted or not. The authors rely on an experiment with several randomized treatments to estimate the parameters of the model. Subsequently, they compare the effects predicted by their model to the results of one new and various existing get-out-the-vote campaigns.

As encouraging as this strategy is, we note that, in general, there is no guarantee that a model that is wrong will yield observably wrong predictions (though it may still misguide). For example, imagine that the true model is one where a share  $q > 0$  of customers behaves non-rationally and that norm

**Table 53.7 Diagnosis of design with incorrect yet observationally equivalent model**

<i>Estimand</i>	<i>Estimator</i>	<i>Estimand</i>	<i>Bias</i>
$\tau_2$	$DIM_2$	0.30	0.00
$\tau_2$	$MLE'_2$	0.30	0.01
$\tau_2$	$MLE'_\infty$	0.30	0.49
$\tau_\infty$	$DIM_2$	-0.06	0.00
$\tau_\infty$	$MLE'_2$	-0.06	-0.15
$\tau_\infty$	$MLE'_\infty$	-0.06	0.00
$\delta$	$MLE'_2$	0.80	-0.15
$\delta$	$MLE'_\infty$	0.80	0.10
$\kappa$	$MLE'_2$	6.00	-2.62
$\kappa$	$MLE'_\infty$	6.00	-0.09
$q$	$MLE'_2$	0.50	

followers always pay a share of 1/2 of their endowment (instead of 3/4). Further suppose again that we as researchers have in mind the wrong model where  $q = 0$ , i.e. everyone behaves rationally. Table 53.7 contains the results from a simulation based on these assumptions. Note first that our estimates of  $\delta$  are, as one would expect, severely biased. Nonetheless, our prediction of the treatment effect in a context with  $n = 2$  based on experimental data from the same context is only very slightly biased and the prediction of the treatment effect for  $n = \infty$  based on data from this context is not biased at all. Predictions *across* contexts are severely biased, which we would, however, not discover if we conducted only a single experiment in one context.

This example highlights that the ability to use unbiased experimental estimates to validate our model does not guarantee that we will find out if our model is wrong. Our model might yield the correct treatment effect estimates for the wrong reason; it may behave well in sample, but poorly out of sample.

## CONCLUSION

Even though field experiments have become immensely popular in political science, there are ongoing worries about how much we can learn from them. Two interrelated concerns are that experiments enjoy limited external validity and that they are disconnected from theory.

Throughout this chapter, we have highlighted the connection between these critiques and discussed various ways in which theories can help researchers learn more from their experiments. Examples range from selection diagrams that help assess how results can be transported from one setting to another to parametric structural models that allow for the extrapolation of

treatment effects to other settings or even other treatments.

Fundamentally, we think there is scope for researchers to do a lot better on these fronts. Tools already exist and we have illustrated how they can be put to use, albeit for a very simple problem.

We close, however, with a worry. There is nothing new to the idea that theory can be a powerful aid to inference. In fact, in some accounts, social experimentation first came up as a method in economics at a time when the dominant mode of inference was structural estimation. Heckman (1991) describes the vision of early experimentalists who were brought up in the structural tradition as one in which the primary goal of an experiment was not the non-parametric identification of an average treatment effect but the estimation of a structural model that could subsequently be used to assess the welfare consequence of numerous other experiments that had not actually been implemented. The turn towards ‘atheoretical’ experimentation was in many ways motivated by concerns about the over-dependence of results on highly parametric and often unrealistic structural models.

Asking experimentalists to accept more theoretical assumptions in exchange for the ability to make broader claims may thus seem like going in circles. The right response, we think, is use with caution. Like any powerful tool, causal models are easy to misuse and should be handled with care. Such care includes being transparent about robustness of conclusions to alternative assumptions, engaging in routine model validation, and validating extrapolations whenever possible.

## APPENDIX

Code for simulations is provided at: <https://gist.github.com/macartan/1ccf6ff3f72042a73701332333a8f27e>



## ACKNOWLEDGEMENTS

We would like to thank Donald Green, James Fearon, Alan Jacobs, and Richard Nielsen for their generous comments on earlier versions of this chapter.

## Notes

- 1 We recognize that the same issues can arise for other experimental and also for non-experimental approaches and believe many of the ideas described in this chapter apply more broadly.
- 2 In their review of the role that theory plays in experimental studies published in top economics journals, for example, Card et al. (2011) classify an experimental study as one that draws on theory only if the study explicitly includes mathematical expressions. Other accounts appear to have a less stringent view that encompasses informal statements about causal relationships derived from prior knowledge (Huber, 2017).
- 3 See Pearl and Mackenzie (2018) for other notions of theory that do not involve claims about causality.
- 4 For a discussion of related alternatives, see, for example, Robins (2003).
- 5 The reason is that average effects of one variable can depend upon levels of another variable which depends on the distribution of exogenous variables. Consider the effect of 'going first'. This effect depends on the customer's behavioral type. Since the theory does not specify the distribution of these types in the population, we would not be able to make statements about the average effect of going first based on this model alone.
- 6 See Mesquita and Tyson (2019) for a treatment of 'commensurability' which can be used to assess whether research designs capture quantities that are of theoretical interest.
- 7 Even with an experiment, however, the identification of average causal effects cannot be achieved completely without assumptions. Experiments must invoke some form of the Stable Unit Treatment Value Assumption (SUTVA) (Gerber and Green, 2012).
- 8 The reason is that the taxi fare is a 'collider' on the path from the number of rounds to whether the customer moves first (Pearl, 2009)
- 9 In practice, the models that are invoked vary in their complexity from several informal statements about hypothesized causal relationships (see e.g. Chong et al., 2014; Olken, 2010) to fully

specified decision-theoretic or game-theoretic models from which empirical predictions are derived in the form of comparative statics (see e.g. Avdeenko and Gilligan, 2015; Blattman and Annan, 2016).

- 10 In practice, researchers tend to either test a null hypothesis of no average treatment effect if they rely on the Neyman (1933) tradition of hypothesis testing, or the sharp null hypothesis of no treatment effect for any unit if they follow the Fisherian (Fisher, 1935) approach. See Bowers and Leavitt, Chapter 41, this *Handbook*, for more on the differences between these approaches.
- 11 This logic links to case selection criteria that are used in qualitative research (Van Evera, 1997).
- 12 See also Pritchett and Sandefur (2015), Vivaldi (2019), Dehejia et al. (2015), and Bisbee et al. (2017).
- 13 In the general solution for  $n > 1$  possible rounds, a customer who moves first pays  $\pi_n^1 = \sum_{t=2}^n (-1)^t \delta^{t-1}$ , where  $t$  indexes the bargaining round. The price paid by a customer who moves second  $\pi_n^2$  is just  $1 - \pi_n^1$ .
- 14  $\tau_2 > \tau_\infty \leftrightarrow 2\delta - 1 > 2\delta(1 + \delta) - 1 \leftrightarrow \delta > 0$ .
- 15 In general,  $q = 2(E(\pi_2^1) + E(\pi_2^2) - 1)$  and  $\delta = \frac{E(\pi_2^1) + 3E(\pi_2^2) - 3}{4(E(\pi_2^1) + E(\pi_2^2)) - 6}$ . Note that for the special case of  $E(\pi_2^1) = \frac{3}{4}$  and  $E(\pi_2^2) = \frac{3}{4}$ , the system implies  $q = 1$  and is consistent with any value of  $\delta$ . Moreover, there are values of  $E(\pi_2^1)$  and  $E(\pi_2^2)$  for which this system of equations has no solution. For example, there is no combination of  $0 \leq q \leq 1$  and  $0 \leq \delta \leq 1$  that can produce  $E(\pi_2^1) = 1/2$  and  $E(\pi_2^2) = 1/4$ .

## REFERENCES

- Abadie, Alberto, Susan Athey, Guido W Imbens, and Jeffrey Wooldridge. 2017. When Should You Adjust Standard Errors for Clustering? NBER Working Paper Series. Working Paper 24003. Cambridge, MA: National Bureau of Economic Research. <https://www.nber.org/papers/w24003>
- Acemoglu, Daron. 2010. Theory, General Equilibrium, and Political Economy in

- Development Economics. *Journal of Economic Perspectives* 24(3): 17–32.
- Allcott, Hunt. 2015. Site Selection Bias in Program Evaluation. *The Quarterly Journal of Economics* 130(3): 1117–65.
- Avdeenko, Alexandra, and Michael J Gilligan. 2015. International Interventions to Build Social Capital: Evidence from a Field Experiment in Sudan. *American Political Science Review* 109(3): 427–49.
- Bisbee, James, Rajeev Dehejia, Cristian Pop-Eleches, and Cyrus Samii. 2017. Local Instruments, Global Extrapolation: External Validity of the Labor Supply–fertility Local Average Treatment Effect. *Journal of Labor Economics* 35(S1): S99–S147.
- Blair, Graeme, Jasper Cooper, Alexander Coppock, and Macartan Humphreys. 2018. Declaring and Diagnosing Research Designs. *American Political Science Review* 113(3): 838–59.
- Blattman, Christopher, and Jeannie Annan. 2016. Can Employment Reduce Lawlessness and Rebellion? A Field Experiment with High-Risk Men in a Fragile State. *American Political Science Review* 110(1): 1–17.
- Bold, Tessa, Mwangi Kimenyi, Germano Mwabu, Alice Ng'ang'a, and Justin Sandefur. 2013. Scaling up What Works: Experimental Evidence on External Validity in Kenyan Education. Center for Global Development Working Paper No. 321. Washington, DC: Center for Global Development. [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=2241240](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2241240)
- Bowers, Jake, Bruce A Desmarais, Mark Fredrickson, Nahomi Ichino, Hsuan-Wei Lee, and Simi Wang. 2018. Models, Methods and Network Topology: Experimental Design for the Study of Interference. *Social Networks* 54: 196–208.
- Bowers, Jake, Mark M Fredrickson, and Costas Panagopoulos. 2013. Reasoning About Interference Between Units: A General Framework. *Political Analysis* 21(1): 97–124.
- Box, George EP. 1976. Science and Statistics. *Journal of the American Statistical Association* 71(356): 791–9.
- Browne, Michael W, and Robert Cudeck. 1993. Alternative Ways of Assessing Model Fit. In: *Testing Structural Equation Models*, edited by Kenneth A. Bollen and J. Scott Long, 136–62. Newbury Park: Sage.
- Card, David, Stefano DellaVigna, and Ulrike Malmendier. 2011. The Role of Theory in Field Experiments. *Journal of Economic Perspectives* 25(3): 39–62.
- Castillo, Marco, Ragan Petrie, Maximo Torero, and Lise Vesterlund. 2013. Gender Differences in Bargaining Outcomes: A Field Experiment on Discrimination. *Journal of Public Economics* 99: 35–48.
- Chong, Alberto, Ana L De La O, Dean Karlan, and Leonard Wantchekon. 2014. Does Corruption Information Inspire the Fight or Quash the Hope? A Field Experiment in Mexico on Voter Turnout, Choice, and Party Identification. *The Journal of Politics* 77(1): 55–71.
- Coppock, Alexander, Thomas J Leeper, and Kevin J Mullinix. 2018. Generalizability of Heterogeneous Treatment Effect Estimates Across Samples. *Proceedings of the National Academy of Sciences* 115(49): 12441–6.
- Dawid, A. Philip. 2010. Beware of the DAG!. *JMLR: Workshop and Conference Proceedings* 6 (*Causality: objectives and assessment.*): 59–86.
- Cronbach, Lee J, and Karen Shapiro. 1982. *Designing Evaluations of Educational and Social Programs*. San Francisco: Jossey-Bass.
- Dawid, Philip, Macartan Humphreys, and Monica Musio. 2019. Bounding Causes of Effects with Mediators. *arXiv Preprint arXiv:1907.00399*. <https://arxiv.org/abs/1907.00399>
- Deaton, Angus. 2010. Understanding the Mechanisms of Economic Development. *Journal of Economic Perspectives* 24(3): 3–16.
- Deaton, Angus, and Nancy Cartwright. 2018. Understanding and Misunderstanding Randomized Controlled Trials. *Social Science & Medicine* 210: 2–21.
- Dehejia, Rajeev, Cristian Pop-Eleches, and Cyrus Samii. 2019. From Local to Global: External Validity in a Fertility Natural Experiment. *Journal of Business & Economic Statistics*. *Forthcoming*: 1–48.
- DellaVigna, Stefano. 2018. Structural Behavioral Economics. In: *Handbook of Behavioral Economics - Foundations and Applications* 1,

- Volume 1, edited by B. Douglas Bernheim, Stefano DellaVigna, and David Laibson, 614–717. Amsterdam: Elsevier.
- DellaVigna, Stefano, Attila Lindner, Balázs Reizer, and Johannes F Schmieder. 2017. Reference-Dependent Job Search: Evidence from Hungary. *The Quarterly Journal of Economics* 132(4): 1969–2018.
- DellaVigna, Stefano, John A List, Ulrike Malmendier, and Gautam Rao. 2016. Voting to Tell Others. *The Review of Economic Studies* 84(1): 143–81.
- Druckman, James N, Donald P Green, James H Kuklinski, and Arthur Lupia. 2011. *Cambridge Handbook of Experimental Political Science*. New York: Cambridge University Press.
- Dunning, Thad, Guy Grossman, Macartan Humphreys, Susan D Hyde, Craig McIntosh, and Gareth Nellis. 2018. *Metaketa I: Information, Accountability, and Cumulative Learning*. Cambridge: Cambridge University Press.
- Fisher, Ronald Aylmer. 1935. *The Design of Experiments*. Edinburgh; London: Oliver & Boyd.
- Gelman, Andrew, Xiao-Li Meng, and Hal Stern. 1996. Posterior Predictive Assessment of Model Fitness via Realized Discrepancies. *Statistica Sinica* 6(4): 733–60.
- Gerber, Alan S, and Donald P Green. 2012. *Field Experiments: Design, Analysis, and Interpretation*. New York; London: WW Norton.
- Gerber, Alan S, Donald P Green, Edward H Kaplan, Ian Shapiro, Rogers M Smith, and Tarek Massoud. 2014. The Illusion of Learning from Observational Research. In *Field Experiments and Their Critics: Essays on the Uses and Abuses of Experimentation in the Social Sciences*, edited by Dawn Langan Teele, 9–32. New Haven, CT: Yale University Press.
- Green, Donald P, Shang E Ha, and John G Bullock. 2010. Enough Already About ‘Black Box’ Experiments: Studying Mediation Is More Difficult Than Most Scholars Suppose. *The Annals of the American Academy of Political and Social Science* 628(1): 200–8.
- Grose, Christian R. 2014. Field Experimental Work on Political Institutions. *Annual Review of Political Science* 17: 355–70.
- Harrison, Glenn W. 2014. Cautionary Notes on the Use of Field Experiments to Address Policy Issues. *Oxford Review of Economic Policy* 30(4): 753–63.
- Heckman, James J. 1991. Randomization and Social Policy Evaluation. NBER Technical Working Paper Series. Technical Working Paper No. 107. Cambridge, MA: National Bureau of Economic Research. <https://www.nber.org/papers/t0107>
- Huber, John D. 2017. *Exclusion by Elections: Inequality, Ethnic Identity, and Democracy*. Cambridge; New York: Cambridge University Press.
- Humphreys, Macartan, and Alan Jacobs. 2017. Qualitative Inference from Causal Models. *Draft Manuscript (Version 0.2)*. Retrieved November 27, 2017.
- Ichino, Nahomi, Jake Bowers, and Mark M Fredrickson. 2013. Ethnicity and Electoral Fraud in New Democracies: Modelling Political Party Agents in Ghana. *Unpublished Manuscript*. <http://www.jakebowers.org/PAPERS/Ichino-Bowers-Fredrickson-Ghana.pdf>
- Imai, Kosuke, Luke Keele, Dustin Tingley, and Teppei Yamamoto. 2011. Unpacking the Black Box of Causality: Learning About Causal Mechanisms from Experimental and Observational Studies. *American Political Science Review* 105(4): 765–89.
- Kern, Holger L, Elizabeth A Stuart, Jennifer Hill, and Donald P Green. 2016. Assessing Methods for Generalizing Experimental Impact Estimates to Target Populations. *Journal of Research on Educational Effectiveness* 9(1): 103–27.
- Lakatos, Imre. 1970. Falsification and the Methodology of Scientific Research Programmes. *Criticism and the Growth of Knowledge* 4: 91–196.
- Lucas, Jeffrey W. 2003. Theory-Testing, Generalization, and the Problem of External Validity. *Sociological Theory* 21(3): 236–53.
- Martinez, Seung-Keun, Stephan Meier, and Charles Sprenger. 2017. Procrastination in the Field: Evidence from Tax Filing. *Unpublished Manuscript*. [https://csap.yale.edu/sites/default/files/files/bsw\\_9-5-17\\_cs.pdf](https://csap.yale.edu/sites/default/files/files/bsw_9-5-17_cs.pdf)
- Mesquita, Ethan Bueno de, and Scott A. Tyson. 2019. The Commensurability Problem: Conceptual Difficulties in Estimating the Effect of Behavior on Behavior. *Working Paper*. <https://>

- [www.dropbox.com/s/p9grrztu6rdj5s3/shocks\\_violence.pdf?dl=0](http://www.dropbox.com/s/p9grrztu6rdj5s3/shocks_violence.pdf?dl=0)
- Michelitch, Kristin. 2015. Does Electoral Competition Exacerbate Interethnic or Interpartisan Economic Discrimination? Evidence from a Field Experiment in Market Price Bargaining. *American Political Science Review* 109(1): 43–61.
- Murtas, Rossella, Alexander Philip Dawid, and Monica Musio. 2017. New Bounds for the Probability of Causation in Mediation Analysis. *arXiv Preprint arXiv:1706.04857*. <https://arxiv.org/abs/1706.04857>
- Neyman, Jerzy, and Egon S. Pearson. 1933. On the Problem of the Most Efficient Tests of Statistical Hypotheses. *Philosophical Transactions of the Royal Society of London Series A, Containing Papers of a Mathematical or Physical Character* 231: 289–337.
- Olken, Benjamin A. 2010. Direct Democracy and Local Public Goods: Evidence from a Field Experiment in Indonesia. *American Political Science Review* 104(2): 243–67.
- Pearl, Judea. 2009. *Causality: Models, Reasoning and Inference*. 2<sup>nd</sup> edition. Cambridge, New York: Cambridge University Press.
- Pearl, Judea, and Elias Bareinboim. 2014. External Validity: From Do-Calculus to Transportability Across Populations. *Statistical Science* 29(4): 579–95.
- Pearl, Judea, and Dana Mackenzie. 2018. *The Book of Why: The New Science of Cause and Effect*. New York: Basic Books.
- Pritchett, Lant, and Justin Sandefur. 2015. Learning from Experiments When Context Matters. *American Economic Review* 105(5): 471–75.
- Robins, James M. 2003. Semantics of Causal Dag Models and the Identification of Direct and Indirect Effects. In: *Highly Structured Stochastic Systems: 70 – 83*, edited by Peter J. Green, Nils Lid Hjort, and Sylvia Richardson. New York: Oxford University Press.
- Rosenbaum, Paul R. 2002. *Observational Studies*. New York: Springer.
- Rosenbaum, Paul R. 2010. *Design of Observational Studies*. New York: Springer.
- Rubinstein, Ariel. 1982. Perfect Equilibrium in a Bargaining Model. *Econometrica: Journal of the Econometric Society* 50(1): 97–109.
- Thelen, Kathleen, and James Mahoney. 2015. Comparative-Historical Analysis in Contemporary Political Science. In *Advances in Comparative-Historical Analysis*, edited by James Mahoney and Kathleen Thelen, 3–36. Cambridge: Cambridge University Press Cambridge.
- Tipton, Elizabeth. 2013. Improving Generalizations from Experiments Using Propensity Score Subclassification: Assumptions, Properties, and Contexts. *Journal of Educational and Behavioral Statistics* 38(3): 239–66.
- Tipton, Elizabeth, and Laura R Peck. 2017. A Design-Based Approach to Improve External Validity in Welfare Policy Evaluations. *Evaluation Review* 41(4): 326–56.
- Todd, Petra E, and Kenneth I Wolpin. 2006. Assessing the Impact of a School Subsidy Program in Mexico: Using a Social Experiment to Validate a Dynamic Behavioral Model of Child Schooling and Fertility. *American Economic Review* 96(5): 1384–1417.
- Van Evera, Stephen. 1997. *Guide to Methods for Students of Political Science*. Ithaca, NY: Cornell University Press.
- Vivaldi, Eva. 2019. How Much Can We Generalize from Impact Evaluations? Unpublished Manuscript. <http://evavivaldi.com/wp-content/uploads/How-Much-Can-We-Generalize.pdf>
- Westreich, Daniel, Jessie K Edwards, Catherine R Lesko, Stephen R Cole, and Elizabeth A Stuart. 2018. Target Validity and the Hierarchy of Study Designs. *American Journal of Epidemiology* 188(2): 438–443.
- Yeh, Robert W, Linda R Valsdottir, Michael W Yeh, Changyu Shen, Daniel B Kramer, Jordan B Strom, Eric A Secemsky, et al. 2018. Parachute Use to Prevent Death and Major Trauma When Jumping from Aircraft: Randomized Controlled Trial. *BMJ* 363: k5094.



# Survey Experiments and the Quest for Valid Interpretation

Gustavo Diaz, Christopher Grady,  
and James H. Kuklinski

When Diana Mutz wrote *Population-Based Survey Experiments* in 2011, she stressed one theme throughout the book. That theme: the use of large random samples with experiments embedded in them is an ideal means by which to generate causal generalizations. The embedded experiment provides the needed leverage to identify true cause and effect, and the random sample of a national population ensures that the results can be generalized to the population from which the sample was drawn.

Mutz's logic remains as compelling today as it was when she wrote the book. However, two significant changes have occurred. First, the increasing influence of the causal inference movement has changed political scientists' priorities with respect to data collection and analysis. Because causal inference emphasizes making the right comparisons, not generalization, researchers increasingly search for unique, often local, research opportunities, thus avoiding the costs and delays associated with the

collection of random samples of national populations. Their practice resonates with Campbell's long-ago assertion about social scientific practices: 'There was gross overvaluing of, and financial investment in, external validity, in the sense of representative samples at the nationwide level. In contrast, the physical sciences are so provincial that they have established major discoveries like the hydrolysis of water... by a single water sample' (Campbell, 1988, cited in Rosenbaum, 1999).<sup>1</sup> As a result, individual scholars' own research programs have progressed quickly, and, more significantly, these same scholars have been able to respond to and build on others' work in rapid-fire fashion.<sup>2</sup>

Second, measurement has emerged as a distinct and very active area of experimental survey research, with some of the discipline's best methodologists working in it.<sup>3</sup> Much of the effort has focused on the measurement of sensitive attitudes. The ingenuity of the designs that scholars have used to identify 'true' attitudes has been nothing short of

remarkable. We put quotes around true attitudes, since one of the key developments in measurement has been a continuing change in conceptions of what true attitudes are.

In both areas of experimental survey research, inference and measurement, scholars seek to interpret their results correctly (i.e. validly). In the case of causal inference, the goal is to reach proper conclusions about relationships between treatments and outcomes. In the case of measurement, the goal is to construct methodological approaches that increase confidence in the measurement of concepts such as prejudice, given that respondents often seek to hide their true views or do not consciously understand what they truly feel and think.

In both types of survey experimental study, scholars might misinterpret the empirical results. But why expect scholars to misinterpret their results? After all, scholars have been describing well-designed experiments as the gold standard by which to estimate true causal relationships for centuries. And no one would doubt that most political scientists can capably design strong survey experiments these days.

Considering causal inference experiments first, we identify and discuss three potential sources of misinterpretation of results: factors not included in the experiment moderate the basic treatment-outcome relationship; some people enter the experiment already having been treated in the very external world the researcher seeks to understand; and, respondents enter the experiment with different experiences, which are typically unknown to the experimenter and which shape the way respondents interpret the treatments. As part of our discussion, we evaluate some increasingly complex methods that scholars have proposed to overcome the sources of misinterpretation.

With respect to survey experiments designed to uncover true attitudes when social desirability might be coloring the respondent's true beliefs and feelings, we identify and discuss several problems that might undermine the increasingly complex designs that scholars

have brought to bear, thus leading them to misinterpret the results. The biggest obstacle to proper interpretation of results is the lack of full respondent anonymity, as viewed by the respondents themselves. Other problems, such as contamination from earlier questions in the survey, stem from the survey context, not necessarily from features of the experiment.<sup>4</sup> With respect to implicit attitudes, the fundamental problem that can impair interpretations of the results, in addition to some of those discussed in the preceding paragraph, arises when the treatments do not prime the concepts researchers think they have primed.

In both areas of survey experimental research, the designs of the experiments have increased in complexity over time. Two obvious questions, which we keep in mind throughout our discussion: has the increasing complexity of experiments increased scholars' capacities to make right inferences about the outside world? And has this increasing complexity wrought its own set of problems, or at least does the potential exist?

To be clear, we define survey experiments as experiments in which the treatments are delivered through a survey instrument. This excludes field experiments that use surveys to measure outcomes (Broockman et al., 2017). Conversely, if respondents enter a laboratory and are assigned to different treatments via different of a survey item, the study meets the definition of a survey experiment.

We have divided our discussion into three sections. First, we discuss reasons why political scientists can inadvertently misinterpret their results when conducting causal inference studies. Second, we undertake the same task with respect to measurement studies. In both instances, we also discuss designs that scholars have begun to propose to avoid misinterpretation. Finally, we pursue some implications of our earlier discussions. The overall purpose of our discussion is to highlight key areas that require the attention of both experts in the field and junior scholars interested in incorporating survey experiments into their research toolkits.

One of the most difficult aspects of writing this chapter is the knowledge that we will not be able to cite the many meritorious studies that warrant citation. While they originated in American politics, survey experiments are now common across all subfields of political science and international relations, exploring topics as different as immigration (Hainmueller and Hopkins, 2015), vote buying (Gonzalez-Ocantos et al., 2012), corruption (Winters and Weitz-Shapiro, 2013), the democratic peace (Tomz and Weeks, 2013), compliance with international treaties (Findley et al., 2017), and support for extremist groups (Blair et al., 2013). We could dedicate an entire chapter to listing interesting applications. Too keep it simple, we have chosen studies with which we are most familiar and that help us make the points we seek to make. We rely on readers to draw connections with the topics that are salient in their areas of expertise.

## **SURVEY EXPERIMENTS FOR CAUSAL INFERENCE**

Survey experiments for causal inference are experiments that happen to be embedded in a survey instrument. Respondents are randomly assigned to different versions of a treatment, and then they answer one or more outcome questions.<sup>5</sup> Much like in field and laboratory experiments, the researcher can identify an average treatment effect on one or more outcomes of interest (see Bowers and Leavitt, Chapter 41; Morton and Vásquez-Cortés, Chapter 51; Sinclair, Chapter 52; and Wilke and Humphries, Chapter 53, this *Handbook* for details). Because scholars in the discipline use causal inference survey experiments to illuminate real-world social and political phenomena, however, finding a non-zero treatment effect is not enough. Researchers must also convince others that their interpretations of treatment-outcome relationships are valid.

Anyone acquainted with causal inference in the social sciences might respond, ‘of course, how could it be any other way?’ In answer, scholars have identified at least three distinct challenges to valid interpretation, which, in our view, must be taken seriously and addressed head-on.

First, survey experimenters can easily overlook or not be able to incorporate factors that moderate the relationship between treatment and outcome (confounding). Second, some respondents might come to the survey experiment having already been pretreated in the very world to which the researcher is trying to infer (pretreatment contamination). Third, respondents might interpret the same treatment in a survey experiment differently due to differences in life experiences and the nature of the environments in which they live, which can be tantamount to receiving different treatments altogether (lack of information equivalence). Ignoring any of these possible complications can lead to wrong interpretations of estimated treatment-outcome relationships.

### ***Confounding***

The most obvious challenge to interpretation in causal inference survey experiments is confounding, which arises from the omission of one or more factors that moderate the relationship between treatment and outcome. This is akin to the problem of omitted variable bias in observational studies (King et al., 1995). The analogy might sound counterintuitive, as we are taught that experiments balance the distribution of both observed and unobserved covariates across groups. However, survey experiments randomize constructs that are not necessarily independent from each other outside the survey framework. Consequently, a treatment in a simple two-group design might inadvertently activate elements that correlate with the treatment in the real world, so that the researcher cannot disentangle the effect of a manipulated treatment from the confounder that is activated indirectly.

Consider the study of corruption. A recurrent debate in the literature is whether citizens sanction corrupt politicians with their votes. Since scholars cannot manipulate corruption, they use hypothetical situations in survey experiments to understand voters' reactions to corrupt politicians and hope that the findings translate to actual voting behavior.<sup>6</sup>

The simplest design presents respondents with a vignette describing a current officeholder seeking reelection. The control group receives information about the incumbent's profile only, while the treatment includes additional information about the incumbent's illicit activities. The design logic is straightforward: if voters sanction corrupt politicians in the treatment group, then, by inference, all that prevents voters from sanctioning corrupt politicians in the external world is the absence of credible information. More bluntly, when voters do not punish corrupt politicians, it is probably because they are unaware of their bad deeds.

Simple and elegant as this design is, it ignores the possibility that voters also respond to other activities the politician undertakes – and perhaps to the politician's personal characteristics as well. Any of these factors could moderate the original relationship between corruption and vote. In the extreme case, the inclusion of such other factors eliminates any initial effect between corruption and vote, which raises questions about the validity of the original interpretation (i.e. that information is the key). Thus, subsequent studies on corruption manipulate not only corruption, but also the provision of public goods, shared partisanship, and even gender (Anduiza et al., 2013; Winters and Weitz-Shapiro, 2013; Eggers et al., 2018). The accumulated evidence in these studies suggests that other factors moderate the relationship between corruption and vote.<sup>7</sup>

In almost all experimental studies, it is easy to identify several potential confounders. In the preceding example, other possible confounders include coercion, vote-buying, a politician's experience in office, and the credibility of the source (Botero et al., 2015;

Mares and Visconti, 2019; Weitz-Shapiro and Winters, 2017). All of these could simultaneously confound the relationship between corruption and vote – and in different directions. However, traditional survey experimental designs set a low limit on the number of potential confounders that can be included, which invariably raises the annoying and ever-present possibility that the experimenter's conclusion will be wrong, or at least incomplete.

Factorial survey experiments (see Auspurg and Hinz, 2015, and Sniderman, 2018, for overviews) provide one way to address the confounding problem, in that the researcher can manipulate both the explanatory variable of interest and a large number of confounders. However, including many potential confounders comes at the cost of statistical power. The researcher now faces a trade-off between accounting for all potential factors that get in the way of proper interpretation and the capacity to identify a non-zero treatment effect.

Conjoint experiments (Hainmueller et al., 2014) overcome this problem by combining clever design and technological advancements in computer-assisted surveys. In the standard conjoint design, the researcher presents respondents with multiple choice tasks between two or more hypothetical alternatives. The combinations consist of independently randomized attributes. In our earlier example, an experimenter might present respondents with two candidates for office. For each candidate, respondents see information about the candidate's level of public goods provision (low or high), party affiliation (left or right), gender (male or female), and so on. In turn, each one of these attributes is randomly assigned to take one of the values included in parenthesis. Because the exercise is hypothetical, the researcher can repeat this exercise multiple times.

Conjoint experiments, then, can incorporate an unusually large number of factors because respondents answer multiple choice tasks that completely randomize the attributes of each alternative, allowing researchers to explore a wide range of combinations before running



into power limitations. Some tasks, like voting, can be reasonably presented as a choice between two or more alternatives. Others, such as the study of immigration attitudes, where researchers ask respondents to put themselves in the role of an immigration officer to determine which individuals should get priority in the admission process, require more creativity (Hainmueller and Hopkins, 2015).

This increased leverage to address confounding comes with a caution: a seemingly impeccable logic, not empirical reality, drives the methodology. This reality opens the door to possible invalid interpretations of empirical results as they apply to the external world. One potential problem is that the quest for satisfying the logic of the methodology itself can come at the cost of realism, in that some combinations rarely, if ever, exist in the bigger world. For example, if a study randomizes occupation and education levels independently, respondents could potentially encounter a doctor with no post-secondary education (Hainmueller and Hopkins, 2015).<sup>8</sup>

Moreover, scholars can now include virtually as many factors as they can imagine, limited only by statistical power and their own discretion about which factors to include and which to exclude. Which factor has the largest effect size presumably depends heavily on the choices researchers make,<sup>9</sup> and thus misinterpretation of results once again becomes a potential problem. Researchers can be misled into thinking that a given cause is more important than others, although, in fact, any result is the product of choices.

Hainmueller and colleagues offer several valuable examples to emulate. In those examples, the use of well-established theories drives the crucial choices. Unfortunately, we already see a tendency in the work that followed the introduction of conjoint experiments to discard theoretical justification and view every factor in the research design as a treatment. This changes the purpose of the study from proper interpretation centered on one treatment of interest to a horse race to determine which factor has a larger effect size.

### ***Pretreatment Contamination***

As we already mentioned, the goal of causal inference survey experiments is to learn about attitudes and behaviors beyond the survey framework. This presents researchers with an interesting dilemma: a research question worthy of pursuit is also likely to be one where respondents encountered the treatment of interest prior to and outside the experiment. This very pretreatment, if not accounted for, can generate wrong interpretations about the effect of the treatment (Druckman and Leeper, 2012; Gaines et al., 2007).

As one of us noted in earlier work, the effect of pretreatment contamination depends on two factors: when the pretreatment occurs vis-à-vis the survey experiment and the longevity of the pretreatment effect, assuming there is one. We do not repeat the details here, except to say that, depending on the existence and endurance of pretreatments, the same static experiment can generate conclusions ranging from no effect at all to a large effect.

In short, survey experiments and the contexts to which experimenters seek to infer can and often do interact across time in highly complex ways. At the extreme, researchers cannot correctly interpret the experimental results without a thorough understanding of the contextual dynamics. However, if they are intimately familiar with the dynamics that happen in the world to which they are trying to infer, they probably can live without the experiment.

Note that attention to pretreatment moves the focus to dynamics, a shift that should resonate with most scholars. After all, the phenomena in the world that scholars study are dynamic. However, nearly all experimental designs are static, and thus they usually are incapable of addressing the effects of pretreatment. What to do?

One possibility: the researcher could simply include a separate question in the survey asking respondents if they have experienced a version of the treatment recently. This is problematic because the question can trigger different complications depending on

its placement. Including the question before treatment is problematic, in that respondents who did not experience the treatment before might be primed by it and approach the experiment as if they have been pretreated. In other words, the experimenter risks replacing pretreatment contaminations with primacy effects. Conversely, including the question after treatment can trigger a false memory.

Chong and Druckman (2013) propose two possible ways by which to identify and account for pretreatment effects: directly manipulate pretreatment and trace effects over time or find a real-world situation where some respondents have been pretreated and others have not.<sup>10</sup> In a first study, they estimate the effects of two competing frames for and against increased law enforcement. To address pretreatment, they conduct a two-wave panel study in which respondents do or do not receive treatment in the first wave (i.e. they are pretreated or not). In the second wave, they explore whether those not treated in the first wave show greater response to the second-wave treatment than those who had been pretreated. They find a big difference in second wave responses, with those not treated earlier showing greater response to the second-wave treatment. In their second study, they take advantage of a situation where some people have followed a controversy and others have not. Again, the results support the idea that pretreatment affects the experimental results.

Although the Chong-Druckman approach provides leverage on pretreatment contamination, implementing a short panel might be out of reach of many research budgets. Moreover, the field is converging in the opposite way. The norm is to perform increasingly complex one-shot studies, and, when resources permit, the preferred option is to replicate the same experiment with a different sample.

### ***Lack of Information Equivalence<sup>11</sup>***

Suppose that a researcher designs a survey experiment that satisfactorily addresses confounding and pretreatment contamination.

Can that researcher justifiably claim valid interpretation? The answer is a resounding ‘no’. In fact, the final challenge to proper interpretation that we consider, and which only recently has come to the fore, is both the most pervasive and most difficult to resolve. The problem is what Dafoe et al. (2018) call (a lack of) information equivalence. Once expressed, its logic is intuitive, even though solving the problem currently fringes on the impossible.

At the risk of oversimplifying, the idea goes as follows. Respondents routinely interpret and answer survey questions in terms of their life experiences. These experiences will vary greatly, especially when the experiment is embedded in a national survey. To the extent that people’s life experiences are sufficiently strong to influence their interpretations of treatments, the result is that, even though the experimental treatment is the same for everyone, different respondents essentially respond to different treatments. How different depends on how much the contextual considerations vary across respondents, and how much those considerations influence their responses.

To return to one of our previous examples, consider an experiment in which information about corruption primes different thoughts in the minds of a respondent from a rich and highly educated district, and a respondent from a poor district with low education levels. To the former, corruption might mean ‘committed a heinous and inexcusable white-collar crime’. To a respondent from a poor district with low education levels, where constituents depend on their officeholders for assistance, corruption might mean ‘doing a good deed’.

Note that the lack of information equivalence undermines basic tenets of both survey and experimental research. On the survey side, researchers must assume that the same question means the same thing to all respondents (King et al., 2004). On the experimental side, lack of information equivalence violates the Stable Unit Treatment Value Assumption (SUTVA), which states that all units assigned

to receive a treatment experience it in the same way (Cox, 1958).

This lack-of-information-equivalence problem cannot be easily solved without altering the scope of the study. Consider the running example in Dafoe et al. (2018). The democratic peace is a proposition in international relations suggesting that democracies never go to war with other democracies (Russett, 1993). Two alternative explanations underlie this proposition. First, democracies dislike war generally and are less likely than autocracies to go to war with anyone. Alternatively, democracies perceive other democracies as less threatening, so they only go to war less often with fellow democracies. To assess between the two explanations, Tomz and Weeks (2013) used a survey experiment that presented respondents with a hypothetical country in the process of acquiring nuclear weapons. They randomly presented the country as a democracy or dictatorship, and respondents indicate whether they favor or oppose a military intervention from their home countries.

In this study, the treatment is a country's political regime. The study deliberately avoids including explicit country labels to prevent confounding. However, the typical democracy that is suspected of developing nuclear weapons (e.g. Israel) is remarkably different from the typical dictatorship that carries the same suspicion (e.g. North Korea). Note that the limitation here is different from confounding. Even if the researcher manipulates one of the main potential confounders (e.g. economic development), the challenge to interpretation will persist. Rich democracies with nuclear power (e.g. France) still differ from poor democracies with nuclear power (e.g. Pakistan).

Whereas increasingly complex designs help with respect to the first two challenges to valid interpretation (confounding and pretreatment contamination), the verdict on whether they can help with the lack of information equivalence problem remains to be seen. The problem is more encompassing and much more difficult to overcome. The

number of possible interpretations of a treatment are countless. 'Good theory' might convince an audience depending on the application, but we currently do not see plausible solutions to the problem itself. The authors themselves could only hint at possibilities.

The lack-of-information-equivalence problem, we might note in closing, is inherent to surveys and survey responses. It is not a derivative of experiments. To overcome the lack of information equivalence problem would not only bring elation to survey experimentalists, it would bring elation to all researchers who use surveys.

### **Summary**

On the surface, conducting survey experiments is straightforward and deceptively easy. Moreover, the cost is relatively low, which makes them especially attractive to graduate students. In fact, the challenges to valid interpretations of survey experiments are many. These challenges, we emphasize, would not be apparent had earlier generations of survey experiments not existed. Increasingly, these challenges have become apparent, and the next generation of survey experimentalists presumably will be more aware of them as they create their own experimental designs. As a result, if the past is any indication, the demand for increasingly complex and sophisticated survey experiments will continue to grow.

### **SURVEY EXPERIMENTS AS A MEASUREMENT TECHNIQUE**

Scholars who use survey experiments to measure attitudes on sensitive issues, or to measure implicit attitudes that respondents themselves fail to see, also seek to design experiments that facilitate proper interpretations of the results. They, too, have encountered not-easily-identified or remedied problems that complicate the task.

Many, although not all, of the potential challenges in the study of measurement are a function of the survey context, as opposed to, as we saw earlier, the features of the larger context that can influence what respondents bring to the survey. With respect to studies that measure attitudes on sensitive issues, scholars routinely assume that eliminating respondents' perceptions of a lack of anonymity is the key to obtaining honest responses and is thus the key to researchers correctly interpreting the experimental results. Scholars who study implicit attitudes face the same challenge, plus the possibility of a lack of information equivalence, which, as we have already seen, is a potentially big challenge in causal inference studies.

### ***List Experiments and Randomized Response Techniques***

Scholars take for granted that respondents do not always answer survey questions truthfully when they are asked about sensitive issues for which there are 'right', or socially desirable, answers. To overcome this possible social desirability bias,<sup>12</sup> researchers have developed and refined two types of survey experiment: the survey list experiment and the randomized response technique. Both techniques are designed to convince individual respondents that their responses to questions about sensitive issues cannot be traced to them. There is some, albeit limited, evidence that these techniques induce less bias than direct questions (Blair et al., 2015; Lensvelt-Mulders et al., 2005b; Rosenfeld et al., 2016).

In a list experiment, the researcher randomly assigns respondents to one of two (or more) conditions. Individuals in the control condition are presented with a list of items; individuals in the treatment condition see the same list plus an additional item, which is the item of interest and the one on which the experimenter wants to ensure the respondent of anonymity. The average difference

between the treatment and control conditions represents the percentage of respondents who responded to the sensitive item in a 'socially undesirable' way (Blair and Imai, 2012).

The randomized response technique (Boruch, 1971; Warner, 1965) is one of the oldest techniques for asking sensitive survey questions.<sup>13</sup> In the most common version of a randomized response question, the respondent is directly asked a yes or no question about a sensitive topic. The respondent is also given some randomization device, like a coin or die. The respondent is told to answer the direct question when the randomization device takes on a certain value (tails) or to say 'yes' when the randomization device takes a different value (heads).<sup>14</sup> Users of the method assume that respondents will believe their anonymity is protected because the researcher cannot know whether a 'yes' resulted from agreement with the sensitive item or the randomization device. Researchers know the expected distribution of the condition, which allows an estimate of overall agreement with the sensitive item (See Lensvelt-Mulders et al., 2005a and Blair et al., 2015 for summaries).

How likely is it that respondents will perceive their answers to socially-sensitive matters as protected and thus truly anonymous? More specifically, what would it take for them to feel their answers are anonymous, especially if they already harbor suspicions? If they do not perceive the safety of autonomy, they will likely shape their responses to portray themselves in the best light possible, rather than answer honestly (Leary and Kowalski, 1990).

There are conditions under which the basic list experiment will almost surely fail to provide anonymity. Most obviously, if all or none of the items on the list anger respondents, those who seek to hide their true feelings and attitudes must answer dishonestly (Blair, 2015). Respondents might not interpret other response options as fully anonymous, either. If the treatment item is something respondents want to renounce unequivocally, they

might report a very low number to dissociate themselves from that item, on the logic that being associated with ‘three of the four [list items] may be interpreted as a 75% chance’ that the respondent holds the socially undesirable attitude (Zigerell, 2011: 544).

The most widely used randomized response technique also offers only limited anonymity to respondents. If a respondent answers ‘yes’, the answer *could* have been dictated by the randomization device, but it could also signal agreement with the sensitive item (Edgell et al., 1982; Yu et al., 2008). Thus, answering ‘yes’ is not unequivocally protected by the design. This response bias can affect respondents who do not hold the sensitive attitude just as readily as it affects respondents who do hold it. Respondents who hold the sensitive attitude might say ‘no’ when directed to be truthful, and respondents who do not hold the sensitive attitude might say ‘no’ when directed to say ‘yes’ (Edgell et al., 1982).

Knowledge of the various problems has helped researchers sharpen both survey experiments and randomize response techniques as tools for measurement. As researchers have learned more about the ways in which respondents respond to survey experiments designed for measurement, they have developed more complex and penetrating list experiments and randomized response techniques to account for them. In the process, researchers arguably have added some complexity to get closer to the right interpretations.<sup>15</sup>

For list experiments, the added complexity comes from increased attention to preparation and design. In terms of preparation, researchers pay even more attention to piloting to find control items that not only fit with the treatment item, but are negatively correlated with other control items (Glynn, 2013). Negatively correlated control items minimize the number of people who will score very high or very low on the control list, a problem that can compromise anonymity.

Variations on the list experiment have helped to isolate the effect of the treatment

item. One variation is the double list experiment (Droitcour et al., 2004; Glynn, 2013), which attempts to solve the problem of respondent interpretation by using two control lists. The treatment item is randomly selected to appear on either the first or the second control list so that some respondents see it on the first list and some respondents see it on the second. If researchers observe the same treatment effect on both lists, there is less risk that the effect depends on the choice of control items or on how respondents interpret the list. Another modification is a placebo-controlled list experiment, which uses a fourth item as a placebo on the control list to ensure that the difference between the two lists is due to the treatment item, not the presence of an extra item (Riambau and Ostwald, 2019).

Users of survey experiments have come up with variations in the randomized response techniques so as, first, to provide what respondents will view as full anonymity and, second, to keep them from viewing one response as riskier. One such variant is the crosswise model (Jann et al., 2011; Yu et al., 2008).<sup>16</sup> In the crosswise model, respondents are presented with two statements, one sensitive statement and one non-sensitive statement, for which the population mean is known. The respondent is asked to say if neither or both statements are true or if one statement is true. Unlike a typical randomized response question, where individuals who agree with the sensitive statement only occupy the ‘yes’ group, the crosswise model allows people who agree with the sensitive statement to occupy either group.<sup>17</sup>

### ***Beyond Ensuring Anonymity***

The jury is out on the effectiveness of these new techniques. They appear to provide something closer to true anonymity, so they come closer to revealing ‘the truth’ than their predecessors. However, is ensuring anonymity a *sufficient* condition to obtain honest answers

to sensitive questions? It is unlikely to be a *necessary* condition – see coercive measures like the bogus pipeline (Jones and Sigall, 1971) for techniques to obtain honest responses that ignore anonymity altogether – but unspoken in work using list experiments and randomized response techniques is the assumption that respondents will answer honestly if they perceive their answers to be anonymous.

We see several reasons why anonymity is *not* a sufficient condition to obtain honest answers to sensitive questions. First, even with anonymity, respondents have no incentive to answer honestly. If a prejudiced person is presented with a list experiment that uses a treatment item designed to measure prejudice, what incentive does that prejudiced individual have to comply with the instructions of the list experiment? In addition to anonymity, a further assumption must be made: respondents want to express their socially undesirable opinions in a way that eludes social sanctions.

Second, anonymity does not help respondents interpret the question as the researcher intended. When the purpose of a question is unclear, respondents must either increase their own cognitive efforts in order to understand the question or satisfice and provide an answer that seems reasonable, even without understanding the question. All survey questions assume that the respondent interprets the question in the way intended by researchers; techniques to ensure anonymity make that interpretation less likely by obfuscating the question's purpose.

Anonymity does not solve many other pitfalls familiar to survey questions and survey experiments. It does not help researchers to avoid question ordering effects or contamination from earlier questions in the survey; it does not reveal how respondents interpret the sensitive item and thus cannot ensure information equivalence. Who knows what other novel problems it does not address? Future research should further explicate the assumptions necessary to obtain honest answers to sensitive questions. Future research should

also reveal further limitations of techniques to measure sensitive attitudes.

One limitation is clear even without further research: these questions do not uncover implicit attitudes. Many sensitive topics appear so sensitive that individual's conscious, explicit attitudes differ from their implicit attitudes (Greenwald and Banaji, 1995). Even many non-sensitive attitudes seem to be beyond an individual's conscious awareness (Nisbett and Wilson, 1977). Techniques like list experiments and randomized response techniques purport to offer anonymity so that respondents feel comfortable revealing their unsavory conscious attitudes, but these techniques do nothing to draw out attitudes that respondents do not know they have. In the next section, we cover survey experimental techniques to reveal respondents' implicit attitudes.

### ***Priming and Implicit Attitudes***

Whereas techniques to measure explicit attitudes seek to provide respondents with anonymity, techniques to measure implicit attitudes seek to keep the respondent consciously unaware of the implicit attitude being measured. To do so, researchers use priming experiments.<sup>18</sup>

In a priming experiment, researchers expose respondents to a stimulus representing topic X in order to influence their responses to a survey question about topic Y, without the conscious knowledge of the respondents. A control group is not exposed to the stimuli representing topic X, so the difference between the treatment group and control group is due to exposure to the treatment stimuli. Priming experiments work by directing respondents' consciousness away from topic X and towards topic Y so that respondents do not consciously censor their feelings about topic X (Macrae et al., 1994; Schwarz and Clore, 1983).

The earliest priming experiments simply randomized the order in which questions were asked (McFarland, 1981). For example,

Schwarz and Bless (1991) show that a question about someone's marriage *before* a question about their general life satisfaction increases life satisfaction for people with good marriages and decreases it for people with bad marriages. The priming paradigm would later be used to measure sensitive attitudes about social groups by priming the social group and then asking respondents about a related topic. For example, Hurwitz and Peffley (1997) prime race by randomizing whether the target of law enforcement action was white or black. This allows them to determine the effect of race on judgments of crime and punishment.

While these priming experiments are extremely innovative, they are also so simple and straightforward that suspicious respondents might realize a connection between the sensitive item being primed and the non-sensitive item being solicited. And, because they measure implicit attitudes by estimating the relationship between the prime and outcomes, they also suffer from the flaws of information equivalence and confounding we discussed in the previous section.

To prevent subjects from ascertaining the goal of the study, researchers try to hide the prime amid other, ostensibly more important, information. One way to do this is with an endorsement experiment (Cohen, 2003). In an endorsement experiment, respondents are asked how much they support a policy. In the treatment condition, the policy is 'endorsed' by a group that respondents would not consciously admit to influencing their opinion. In the control condition, the policy is not endorsed by any group. The average difference in support between the endorsed and unendorsed policy represents the change in support for the policy because of the endorsement.<sup>19</sup>

Though endorsement experiments help hide the goal of the study by distracting attention away from the group and towards a substantive policy, they still suffer from lack of information equivalence. In endorsement experiments, the problem manifests itself because a group's endorsement may be used

as a heuristic to understand substantive policy details (Lupia, 1994). The basic endorsement experiment cannot differentiate bias towards the endorsing group from use of the endorsing group as an information heuristic. To ensure information equivalence, researchers utilize endorsements as part of factorial experiments that vary substantive details about the policy along with group endorsement. For example, Nicholson (2011) uses this design to show that a group's endorsement of a policy is overwhelmed by information about the social groups who are helped or harmed by the policy.

Priming experiments still suffer from several problems. First, the treatment might not prime the intended concept. Unlike list experiments or randomized response questions, the item of interest is not directly enumerated to the respondent. It is possible that the treatment activates different attitudes than the researcher intends. Second, the mental construct being primed may already be salient in the minds of all respondents (i.e. pretreated), rendering the prime impotent.<sup>20</sup> Unfortunately, methods to validate the estimates of priming experiments do not exist yet.

## Summary

Scholars want to measure concepts validly and reliably. Direct survey questions are often used to measure concepts, but direct questions fail when respondents lie or do not have conscious access to the attitude the researcher is interested in. To address the reasons that direct questions fail, scholars began measuring some concepts with survey experiments in lieu of direct questions. As scholars learned more about measuring concepts with survey experiments, they learned the pitfalls of survey experiments for measurement and further adapted their measures to account for those pitfalls. Thus, the history of measuring sensitive concepts has been to increase complexity of design for the purpose of increasing validity.

One problem plagues all measurement: how do we know that our measure is valid? For some outcomes, such as voter turnout, we can compare our measure with population estimates. But for other outcomes, such as racism or the effect that political parties have on citizens' policy preferences, no population estimate with which to validate our measure exists. We are searching for a truth we cannot know.

Despite an inability to validate most measures, we make progress by fully explicating the assumptions of each measurement strategy and then determining if the strategy fulfills them. When measurement strategies do not satisfy their own assumptions, researchers must create measures that do. For example, researchers assume anonymity is the key to obtaining honest answers about sensitive topics. Yu et al. (2008) noticed that randomized response techniques fail to satisfy this assumption and created the crosswise model to provide respondents with full anonymity.

Part of explicating assumptions is explicating the reasons as to why direct questions will not validly measure a concept of interest. Sometimes scholars will find that people are not as squeamish as scholars expect, and thus direct questions work well. Other times the techniques that provide anonymity do not reveal the attitude of interest because the barrier to measurement is respondents lacking conscious access to that attitude. We think it is important that the theoretical assumptions about the concept we are measuring match the assumptions of our measurement strategy.

## CONCLUSION

Survey experimental research has matured quickly and dramatically. In both areas of endeavor that we reviewed, researchers appear to be functioning in the best tradition of 'normal science' (Kuhn, 1962), which we applaud. The growing complexity and sophistication in research designs has led to new discoveries, and, subsequently, more

challenges, which in turn have generated even more creative and complex designs. The more we learn, it seems, the more complex and sophisticated the next generation of research must be to address the newest, and often unexpected, discoveries. No serious researcher will be surprised to know that a final stopping point is nowhere to be seen.

Increased complexity implies more moving parts, and the more moving parts there are, the less transparent the methods and empirics can become. Recent work on conjoint analysis exemplifies this statement (Hainmueller et al., 2014). The logic of leveraging pair comparisons to isolate independent effects across manipulations fringes on impeccable, and as a result, the potential to remedy some existing challenges is, at least in theory, high. Yet, knowing precisely how respondents respond to and interpret many descriptive pairs remains slightly difficult to ascertain. Likewise, recent measurement advances like the crosswise model (Yu et al., 2008) appear to offer anonymity in a way that is convincing to respondents. Yet, scholars have no way of assessing the validity of this approach on most sensitive attitudes.

We are reminded of the dictum of 'no more than three independent variables', as applied it to the growing complexity of probit and logit models (Achen, 2002). We see no reason to apply it to survey experimental research at this time, although keeping it close at hand as the two areas progress would be wise. Increasing the complexity and sophistication of survey experimental designs has thus far evolved in a logical, progressive, and helpful way. Whether a lack of transparency will begin to obfuscate the power of the designs remains to be seen.

Overall, reviewing the evolution of survey experimental research has been, for us, an eye-opener. Although we have remained conversant with the literature as it has grown, peering into the bowels of the research has increased our understanding of the challenges that await new generations of researchers. It has also increased what was our already-high respect for the work.



## Notes

- 1 We do not address external validity or inferring from an experiment to some (often undefined) world outside it. Suffice it to say that we agree with the contention that the primary purpose of an experiment should be to test theory, not to infer to some outside world (Mook, 1983). However, for the purposes of this chapter, we take the general practice of inferring from experiment to world as given.
- 2 The speed with which researchers can respond to each other underlines the crucial importance of publishing the results of well-designed survey experiments that generate null findings.
- 3 Note that this distinction is primarily a working and admittedly artificial one. The study of causal inference requires good measurement and the study of measurement cannot occur without causal inference. Our distinction echoes what is current practice in the field.
- 4 These problems apply to causal inference studies as well. We bring them up in the discussion of measurement because it is there that the potential consequences of survey contamination are most obvious.
- 5 Depending on the question of interest, the control group may receive no treatment at all, or an innocuous version of the treatment (placebo). Some survey experiments include both pure control and placebo conditions. The conventional advice is to include some form of control group to identify the direction of the treatment effect (Gaines et al., 2007).
- 6 Note that survey experiments are not the only way to study corruption. Occasionally, researchers find direct measures in observational data (Fernández-Vázquez et al., 2016; Ferraz and Finan, 2008; Golden and Picci, 2005; Olken, 2007).
- 7 This example is certainly not the only case of researchers using multiple factors in survey experiments. In the study of American politics, the practice is traceable at least back to the early studies on racial prejudice (Sniderman et al., 1991) and heuristic processing (Mondak, 1993).
- 8 The standard practice is to prevent these combinations from appearing, but that might introduce bias in the average marginal component effect. An alternative is to allow for illogical combinations with a relatively low probability (Hainmueller et al., 2014).
- 9 Although previous work provides guidelines for the number of factors that should be included (Auspurg and Hinz, 2015; Hainmueller et al., 2014), we are not aware of any developments regarding which factors to include.
- 10 Chong and Druckman (2010, 2013) pioneered dynamic studies. Even today, few have followed their footsteps.
- 11 In their original study, Dafoe et al. (2018) use the term information equivalence. We use lack of information equivalence because a lack of equivalence is the problem they address.
- 12 Other instances of response bias are due to demand effects (when the respondent learns what the researcher wants to hear and obliges) and acquiescence bias (the tendency of respondents to agree rather than disagree with statements). Survey experiments help overcome these biases by hiding the intent of the researcher. Other reasons for respondent's lying are when the respondent deliberately provides untrue responses for fun, lack of attention, and desire to get through the survey quickly. Most survey experimental techniques will not overcome these issues.
- 13 While many people do not consider randomized response models as experiments, we view them as experiments in which the researcher does not know the experimental condition of the respondent. Results can still be analyzed in an 'experimental' fashion (and often compared to results from other experiments to reduce bias) because the data-generating process is still known. The same logic applies to non-randomized response models, the close cousins of randomized response models.
- 14 This is known as the 'forced response' model, introduced by Fox and Tracy (1986). Other models use slightly different procedures.
- 15 A full discussion of all recent advances in list experiments and randomized response techniques is beyond our scope. Here we focus on design advances and omit work on statistical analysis of survey experiments and the comparison of responses to direct questions (Ahluquist, 2018; Aronow et al., 2015; Blair and Imai, 2012; Blair et al., 2015; Blair et al., 2018; Chou et al., 2017; Corstange, 2009; Rosenfeld et al., 2016).
- 16 Technically, the crosswise model is a non-randomized response model. Non-randomized response models pair a sensitive question with some nonrandom phenomenon, instead of with random phenomenon like a coin flip. The sensitive question and the non-random phenomenon are paired in such a way that the researcher cannot know if the respondent agrees with the sensitive question or the non-random phenomenon.
- 17 Lensvelt-Mulders et al. (2005a), Azfar and Murrell (2009), and Gingerich (2015) report other randomized response techniques and advances.
- 18 Other techniques are also used, such as implicit associations tests (Greenwald et al., 1998) and physiological measures (Rankin and Campbell, 1955). Survey experiments enjoy one major

advantage: they can easily be administered to respondents outside of the laboratory.

- 19 Endorsement experiments have recently been used to measure explicit attitudes towards groups that may be dangerous to support publicly. Rather than measure implicit attitudes, these 'explicit' endorsement experiments work like list experiments, where individuals can freely express their support for the sensitive group because the researcher cannot differentiate policy support from group support at an individual level. Whereas list experiments hide the respondent's opinion by pairing the sensitive item with non-sensitive control items, endorsement experiments hide the respondent's opinion by pairing the sensitive item with a policy (e.g. Blair, 2015).
- 20 The outcome being measured could also activate the prime, which would accidentally treat the control group. This happens when the outcome is too close mentally to the prime. For example, the term 'welfare' may make racial minorities salient in the minds of white respondents.

## REFERENCES

- Achen, Christopher H. 2002. Toward a New Political Methodology: Microfoundations and ART. *Annual Review of Political Science* 5(1): 423–50.
- Ahlquist, John S. 2018. List Experiment Design, Non-Strategic Respondent Error, and Item Count Technique Estimators. *Political Analysis* 26(1): 34–53.
- Anduiza, Eva, Aina Gallego, and Jordi Muñoz. 2013. Turning a Blind Eye: Experimental Evidence of Partisan Bias in Attitudes Towards Corruption. *Comparative Political Studies* 46(12): 1664–92.
- Aronow, Peter M., Alexander Coppock, Forrest W. Crawford, and Donald P. Green. 2015. Combining List Experiment and Direct Question Estimates of Sensitive Behavior Prevalence. *Journal of Survey Statistics and Methodology* 3(1): 43–66.
- Auspurg, Katrin, and Thomas Hinz. 2015. *Factorial Survey Experiments*. London: Sage.
- Azfar, Omar, and Peter Murrell. 2009. Identifying Reticent Respondents: Assessing the Quality of Survey Data on Corruption and Values. *Economic Development and Cultural Change* 57(2): 387–411.
- Blair, Graeme. 2015. Survey Methods for Sensitive Topics. *Comparative Politics Newsletter* 12: 44.
- Blair, Graeme, C. Christine Fair, Neil Malhotra, and Jacob N. Shapiro. 2013. Poverty and Support for Militant Politics: Evidence from Pakistan. *American Journal of Political Science* 57(1): 30–48.
- Blair, Graeme, and Kosuke Imai. 2012. Statistical Analysis of List Experiments. *Political Analysis* 20(1): 47–77.
- Blair, Graeme, Kosuke Imai, and Yang-Yang Zhou. 2015. Design and Analysis of the Randomized Response Technique. *Journal of the American Statistical Association* 110(511): 1304–19.
- Blair, Graeme, Winston Chou, and Kosuke Imai. 2018. List Experiments with Measurement Error. *Political Analysis*, 1–26.
- Boruch, Robert F. 1971. Assuring Confidentiality of Responses in Social Research: A Note on Strategies. *The American Sociologist*, 308–311.
- Botero, Sandra, Rodrigo Castro Cornejo, Laura Gamboa, Nara Pavao, and David W. Nickerson. 2015. Says Who? An Experiment on Allegations of Corruption and Credibility of Sources. *Political Research Quarterly* 68(3): 493–504.
- Broockman, David E., Joshua L. Kalla, and Jasjeet S. Sekhon. 2017. The Design of Field Experiments With Survey Outcomes: A Framework for Selecting More Efficient, Robust, and Ethical Designs. *Political Analysis* 25(4): 435–64.
- Campbell, Donald T. 1988. Can We Be Scientific in Applied Social Science? In Donald T. Campbell and E. Samuel Overman (eds) *Methodology and Epistemology for Social Science: Selected Papers*, 315–33. Chicago: University of Chicago Press.
- Chong, Dennis, and James N. Druckman. 2010. Dynamic Public Opinion: Communication Effects over Time. *American Political Science Review* 104(4): 663–80.
- Chong, Dennis, and James N. Druckman. 2013. Counterframing Effects. *The Journal of Politics* 75(1): 1–16.
- Chou, Winston, Kosuke Imai, and Bryn Rosenfeld. 2017. Sensitive Survey Questions with Auxiliary Information. *Sociological Methods & Research*. Available at <https://doi.org/10.1177/0049124117729711> (Accessed 20 January 2020).

- Cohen, Geoffrey L. 2003. Party Over Policy: The Dominating Impact of Group Influence on Political Beliefs. *Journal of Personality and Social Psychology* 85(5): 808–22.
- Corstange, Daniel. 2009. Sensitive Questions, Truthful Answers? Modeling the List Experiment with LISTIT. *Political Analysis* 17(1): 45–63.
- Cox, D. R. 1958. *Planning of Experiments*. Wiley Classics Library. New York: Wiley.
- Dafoe, Allan, Baobao Zhang, and Devin Caughey. 2018. Information Equivalence in Survey Experiments. *Political Analysis* 26(4): 399–416.
- Droitcour, Judith, Rachel A. Caspar, Michael L. Hubbard, Teresa L. Parsley, Wendy Visscher, and Trena M. Ezzati. 2004. The Item Count Technique as a Method of Indirect Questioning: A Review of Its Development and a Case Study Application. In Paul P. Biemer, Robert M. Groves, Lars E. Lyberg, Nancy A. Mathiowetz, and Seymour Sudman (eds) *Measurement Errors in Surveys*, 185–210. Hoboken: Wiley.
- Druckman, James N., and Thomas J. Leeper. 2012. Learning More from Political Communication Experiments: Pretreatment and Its Effects. *American Journal of Political Science* 56(4): 875–896.
- Edgell, Stephen E., Samuel Himmelfarb, and Karen L. Duchan. 1982. Validity of Forced Responses in a Randomized Response Model. *Sociological Methods & Research* 11(1): 89–100.
- Eggers, Andrew C., Nick Vivyan, and Markus Wagner. 2018. Corruption, Accountability, and Gender: Do Female Politicians Face Higher Standards in Public Life? *The Journal of Politics* 80(1): 321–26.
- Fernández-Vázquez, Pablo, Pablo Barberá, and Gonzalo Rivero. 2016. Rooting Out Corruption or Rooting for Corruption? The Heterogeneous Electoral Consequences of Scandals. *Political Science Research and Methods* 4(2): 379–97.
- Ferraz, Claudio, and Federico Finan. 2008. Exposing Corrupt Politicians: The Effects of Brazil's Public Released Audits on Electoral Outcomes. *Quarterly Journal of Economics* 123(2): 703–45.
- Findley, Michael G., Brock Laney, Daniel L. Nielson, and J. C. Sharman. 2017. External Validity in Parallel Global Field and Survey Experiments on Anonymous Incorporation. *The Journal of Politics* 79(3): 856–72.
- Fox, James, and Paul Tracy. 1986. *Randomized Response*. London: Sage.
- Gaines, Brian J., James H. Kuklinski, and Paul J. Quirk. 2007. The Logic of the Survey Experiment Reexamined. *Political Analysis* 15(1): 1–20.
- Gingerich, Daniel. 2015. Randomized Response: Foundations and New Developments. *Newsletter of the Comparative Politics Organized Section of the American Political Science Association* (The Organized Section in Comparative Politics of the American Political Science Association) 25(1): 16–27.
- Glynn, Adam N. 2013. What Can We Learn with Statistical Truth Serum? *Public Opinion Quarterly* 77(S1): 159–72.
- Golden, Miriam A., and Lucio Picci. 2005. Proposal for a New Measure of Corruption, Illustrated with Italian Data. *Economics and Politics* 17(1): 37–75.
- Gonzalez-Ocantos, Ezequiel, Chad Kiewiet De Jonge, Carlos Meléndez, Javier Osorio, and David W. Nickerson. 2012. Vote Buying and Social Desirability Bias: Experimental Evidence from Nicaragua. *American Journal of Political Science* 56(1): 202–217.
- Greenwald, Anthony G., Debbie E. McGhee, and Jordan L. K. Schwartz. 1998. Measuring Individual Differences in Implicit Cognition: The Implicit Association Test. *Journal of Personality and Social Psychology* 74(6): 1464–80.
- Greenwald, Anthony G., and Mahzarin R. Banaji. 1995. Implicit Social Cognition: Attitudes, Self-Esteem, and Stereotypes. *Psychological Review* 102(1): 4.
- Hainmueller, Jens, and Daniel J. Hopkins. 2015. The Hidden American Immigration Consensus: A Conjoint Analysis of Attitudes toward Immigrants. *American Journal of Political Science* 59(3): 529–48.
- Hainmueller, Jens, Daniel J. Hopkins, and Teppei Yamamoto. 2014. Causal Inference in Conjoint Analysis: Understanding Multidimensional Choices via Stated Preference Experiments. *Political Analysis* 22(1): 1–30.
- Hurwitz, Jon, and Mark Peffley. 1997. Public Perceptions of Race and Crime: The Role of

- Racial Stereotypes. *American Journal of Political Science* 41(2): 375–401.
- Jann, Ben, Julia Jerke, and Ivar Krumpal. 2011. Asking Sensitive Questions Using the Cross-wise Model: An Experimental Survey Measuring Plagiarism. *Public Opinion Quarterly* 76(1): 32–49.
- Jones, Edward E., and Harold Sigall. 1971. The Bogus Pipeline: A New Paradigm for Measuring Affect and Attitude. *Psychological Bulletin* 76(5): 349–64.
- King, Gary, Christopher J. L. Murray, Joshua A. Salomon, and Ajay Tandon. 2004. Enhancing the Validity and Cross-Cultural Comparability of Measurement in Survey Research. *American Political Science Review* 98(1): 191–207.
- King, Gary, Robert O. Keohane, and Sidney Verba. 1995. *Designing Social Inquiry: Scientific Inference in Qualitative Research*. Princeton: Princeton University Press.
- Kuhn, Thomas S. 1962. *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.
- Leary, Mark R., and Robin M. Kowalski. 1990. Impression Management: A Literature Review and Two-Component Model. *Psychological Bulletin* 107(1): 34.
- Lensvelt-Mulders, Gerty J. L. M., Joop J. Hox, and Peter G. M. Van Der Heijden. 2005a. How to Improve the Efficiency of Randomised Response Designs. *Quality and Quantity* 39(3): 253–265.
- Lensvelt-Mulders, Gerty J. L. M., Joop J. Hox, Peter G. M. Van der Heijden, and Cora JM Maas. 2005b. Meta-Analysis of Randomized Response Research: Thirty-Five Years of Validation. *Sociological Methods & Research* 33(3): 319–348.
- Lupia, Arthur. 1994. Shortcuts versus Encyclopedias: Information and Voting Behavior in California Insurance Reform Elections. *American Political Science Review* 88(1): 63–76.
- Macrae, C. Neil, Galen V. Bodenhausen, Alan B. Milne, and Jolanda Jetten. 1994. Out of Mind but Back in Sight: Stereotypes on the Rebound. *Journal of Personality and Social Psychology* 67(5): 808.
- Mares, Isabela, and Giancarlo Visconti. 2019. Voting for the Lesser Evil: Evidence from a Conjoint Experiment in Romania. *Political Science Research and Methods*, 1–14. Available at <https://doi.org/10.1017/psrm.2019.12> (Accessed 20 January 2020)
- McFarland, Sam G. 1981. Effects of Question Order on Survey Responses'. *Public Opinion Quarterly* 45(2): 208.
- Mondak, Jeffery J. 1993. Source Cues and Policy Approval: The Cognitive Dynamics of Public Support for the Reagan Agenda. *American Journal of Political Science* 37(1): 186.
- Mook, Douglas G. 1983. In Defense of External Invalidity. *American Psychologist* 38(4): 379.
- Mutz, Diana C. 2011. *Population-Based Survey Experiments*. Princeton: Princeton University Press.
- Nicholson, Stephen P. 2011. Dominating Cues and the Limits of Elite Influence. *The Journal of Politics* 73(4): 1165–77.
- Nisbett, Richard E., and Timothy D. Wilson. 1977. Telling More than We Can Know: Verbal Reports on Mental Processes. *Psychological Review* 84(3): 231.
- Olken, Benjamin A. 2007. Monitoring Corruption: Evidence from a Field Experiment in Indonesia. *Journal of Political Economy* 115(2): 200–249.
- Rankin, Robert E., and Donald T. Campbell. 1955. Galvanic Skin Response to Negro and White Experimenters. *The Journal of Abnormal and Social Psychology* 51(1): 30.
- Riambau, Guillem, and Kai Ostwald. 2019. Placebo Statements in List Experiments: Evidence from a Face-to-Face Survey in Singapore. Working Paper. <http://guillemriambau.com/Placebo%20Statements%20in%20List%20Experiments.pdf> (Accessed 20 January 2020).
- Rosenbaum, Paul R. 1999. Choice as an Alternative to Control in Observational Studies. *Statistical Science* 14(3): 259–304.
- Rosenfeld, Bryn, Kosuke Imai, and Jacob N. Shapiro. 2016. An Empirical Validation Study of Popular Survey Methodologies for Sensitive Questions. *American Journal of Political Science* 60(3): 783–802.
- Russett, Bruce M. 1993. *Grasping the Democratic Peace: Principles for a Post-Cold War World*. Princeton: Princeton University Press.
- Schwarz, Norbert, and Gerald L. Clore. 1983. Mood, Misattribution, and Judgments of Well-Being: Informative and Directive Functions of Affective States. *Journal of Personality and Social Psychology* 45(3): 513.

- Schwarz, Norbert, and Herbert Bless. 1991. Constructing Reality and Its Alternatives: An Inclusion/Exclusion Model of Assimilation and Contrast Effects in Social Judgment.. In Leonard L. Martin and Abraham Tesser (eds) *The Construction of Social Judgments*, 217–245, New Jersey: Lawrence Erlbaum Associates..
- Sniderman, Paul M. 2018. Some Advances in the Design of Survey Experiments. *Annual Review of Political Science* 21(1): 259–75.
- Sniderman, Paul M., Thomas Piazza, Philip E. Tetlock, and Ann Kendrick. 1991. The New Racism. *American Journal of Political Science* 35(2): 423.
- Tomz, Michael R., and Jessica L. P. Weeks. 2013. Public Opinion and the Democratic Peace. *American Political Science Review* 107(4): 849–65.
- Warner, Stanley L. 1965. Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias. *Journal of the American Statistical Association* 60(309): 63–69.
- Weitz-Shapiro, Rebecca, and Matthew S. Winters. 2017. Can Citizens Discern? Information Credibility, Political Sophistication, and the Punishment of Corruption in Brazil. *Journal of Politics* 79(1): 60–74.
- Winters, Matthew S., and Rebecca Weitz-Shapiro. 2013. Lacking Information or Condoning Corruption: When Do Voters Support Corrupt Politicians? *Comparative Politics* 45(4): 418–36.
- Yu, Jun-Wu, Guo-Liang Tian, and Man-Lai Tang. 2008. Two New Models for Survey Sampling with Sensitive Characteristic: Design and Analysis. *Metrika* 67(3): 251.
- Zigerell, Lawrence J. 2011. You Wouldn't like Me When I'm Angry: List Experiment Misreporting. *Social Science Quarterly* 92(2): 552–562.



# Deep Learning for Political Science

Kakia Chatsiou and Slava Jankin Mikhaylov

## INTRODUCTION

Political science, and social science in general, have traditionally been using computational methods to study areas such as voting behavior, policy making, international conflict, and international development. More recently, increasingly available quantities of data are being combined with improved algorithms and affordable computational resources to predict, learn, and discover new insights from data that is large in volume and variety. New developments in the areas of machine learning, deep learning, natural language processing (NLP), and, more generally, artificial intelligence (AI) are opening up new opportunities for testing theories and evaluating the impact of interventions and programs in a more dynamic and effective way. Applications using large volumes of structured and unstructured data are becoming common in government and industry, and increasingly also in social science research.

This chapter offers an introduction to such methods drawing examples from political science. Focusing on the areas where the strengths of the methods coincide with challenges in these fields, the chapter first presents an introduction to AI and its core technology – machine learning, with its rapidly developing subfield of deep learning. The discussion of deep neural networks is illustrated with the NLP tasks that are relevant to political science. The latest advances in deep learning methods for NLP are also reviewed, together with their potential for improving information extraction and pattern recognition from political science texts.

We conclude by reflecting on issues of algorithmic bias – often overlooked in political science research. We also discuss the issues of fairness, accountability, and transparency in machine learning, which are being addressed at the academic and public policy levels.

## AI: MACHINE LEARNING AND NLP

The European Commission (2019) defines AI as ‘systems that display intelligent behaviour by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals’. As a scientific discipline, AI includes several techniques like machine learning (with deep learning and reinforcement learning as specific examples), machine reasoning, and robotics (European Commission, 2019). However, much of what is discussed as AI in the public sphere is machine learning, which is an ‘algorithmic field that blends ideas from statistics, computer science and many other disciplines [...] to design algorithms that process data, make predictions, and help make decisions’ (Jordan, 2019).

Machine learning has a history of successful deployment in both industry and academia, going back several decades. Deep learning has more recently made great progress in such applications as speech and language understanding, computer vision, and event and behavior prediction (Goodfellow et al., 2016). These rapid technological advances and the promise of automation and human-intelligence augmentation (Jordan, 2019) reignited debates on AI’s impact on jobs and markets (Brynjolfsson et al., 2018; Samothrakis, 2018; Schlogl and Sumner, 2018) and the need for AI governance (Aletras et al., 2016; Benjamins et al., 2005).

Machine learning (and deep learning as its subfield) is defined as the ‘field of study that gives computers the ability to learn without being explicitly programmed’ (Samuel, 1959). In this context, ‘learning’ can be viewed as the use of statistical techniques to enable computer systems to progressively improve their performance on a specific task using data without being explicitly programmed (Goldberg and Holland, 1988). To be able to learn how to perform a task and become better at it, a machine should:

- be provided with a set of example information (inputs) and the desired outputs. The goal is then to learn a general rule that can take us from

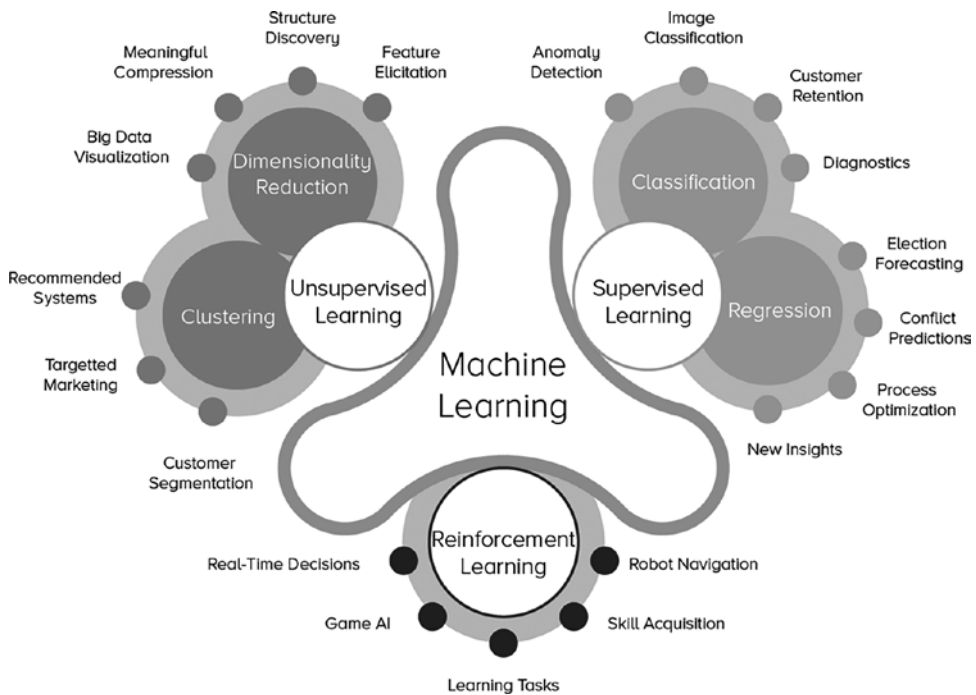
the inputs to the outputs. This type of learning is called *Supervised Learning*. This works well even in cases when the input information is not available in full;

- be provided with an incomplete set of example information to learn from, where some of the target outputs are missing. This type of learning is called *Semi-supervised Learning*. When example information is available in one domain and we want to apply the knowledge to another domain with no available example information, this is called *Transfer Learning*;
- obtain training labels for a small number of instances while at the same time optimize which elements it needs to learn labels for. This is called *Active Learning*, and, in some cases, it can be implemented interactively in order to ask a human user for information on how best to label different elements;
- be asked to find structure in the input without having any labels provided in advance (as input). This type of learning is called *Unsupervised Learning* and can be used both for discovering hidden patterns in the data as well as learning features or parameters from the data; and
- be given information not about the structure of the data itself but rather about whether it has learned something correctly or incorrectly, in the form of rewards and punishments. This is called *Reinforcement Learning* and is the type of learning best performed in dynamic environments such as when driving a vehicle or playing a game against an opponent (Bishop, 2006).

Figure 55.1 summarizes different types of learning and how they relate to their subtasks.

One of the most fruitful areas of machine learning applications in political science relates to work that treats text as data. Such quantitative text analysis could involve tasks such as *classification, clustering, dimensionality reduction, structured prediction or learning*.

Assigning a category to a group of documents or other elements (‘classification’) is useful when, for example, there is a need to understand audience sentiment from social media or customer reviews or sort party manifestos into predefined categories on the ideological spectrum. Spam filtering is an example of classification from our



**Figure 55.1** Machine learning and related tasks

contemporary daily life, where the inputs are email (or other) messages and the classes are ‘spam’ and ‘not spam’. The task involves a dataset containing text documents with labels, which is then used to train a classifier aiming to automatically classify the text documents into one or more predefined categories. Inputs are divided into two or more classes, and the algorithm assigns unseen inputs to one or more (multi-label classification) of these classes. This is typically tackled via supervised learning. In political science work, such models have been used, for example, to understand US Supreme Court decisions (Evans et al., 2007), party affiliation (Yu et al., 2008), and in measuring polarization (Peterson and Spirling, 2018).

Separating elements into groups (‘clustering’) is similar to classification, only the groups are not known beforehand, hence this task usually involves unsupervised learning. Sanders et al. (2017) and Preoțiu-Pietro et al. (2017) are examples of the potential use

of clustering to better understand political ideologies and parliamentary topics.

Reducing the complexity of data (‘Dimensionality Reduction’) involves simplifying inputs by mapping them into a lower-dimensional space. Principal-components analysis and related methods like correspondence analysis have been used to analyze preferences for foreign aid (Baker, 2015) and the ideological mapping of candidates and campaign contributors (Bonica, 2014). Topic modeling is a related problem, where multiple documents are reduced to a smaller set of underlying themes or topics. Feature extraction is a type of dimensionality reduction task and can be accomplished using either semi-supervised or unsupervised learning. Selection and extraction of text features from documents or words is essential for text mining and information retrieval, where learning is done by seeking to reduce the dimension of the learning set into a set of features (Nguyen et al., 2015; Uysal, 2016).



'Structured prediction or structured (output) learning' involves performing structured predictions and is an umbrella term for supervised machine learning techniques that involve predicting structured objects, rather than scalar discrete or real values (BakIr, 2007). In Lafferty et al. (2001), for example, the issue of translating a natural-language sentence into a syntactic representation such as a parse tree can be seen as a structured-prediction problem in which the structured-output domain is the set of all possible parse trees.

Table 55.1 summarizes some of these techniques providing examples from political science.

These political text-as-data applications are related to the broader field of NLP, which is concerned with the interactions between computers and human or natural languages (rather than formal languages). After the 1980s and alongside the developments in machine learning and advances in hardware and technology, NLP has mostly evolved around the use of statistical models to automatically identify patterns and structures in language, through the analysis of large sets of annotated texts or corpora. In addition to document classification and dimensionality-reduction applications in political science, leveraging the latest developments in machine learning and deep learning methods, the NLP field has made significant progress on several additional tasks:

- *Extracting text from an image.* Such a task usually involves a form of *Optical Character Recognition*, which can help with determining the corresponding text characters from an image of printed or handwritten text.
- *Identifying boundaries and segment text into smaller units (for example from documents to characters).* Examples of such tasks include morphological segmentation, word segmentation, and sentence-boundary disambiguation.
- *Morphological segmentation* is the field of separating words into individual morphemes and identifying the class of the morphemes is an essential step of text preprocessing before textual data can be used as an input in some machine learning algorithms. Some such tasks can be quite challenging to perform automatically, sometimes depending on morphological complexity (i.e., the internal structure of words) of the language being considered.
- *Word segmentation or tokenization* makes possible the separation of continuous text into separate words.
- *Sentence-boundary disambiguation* helps identify where a sentence starts and where it ends. This is not as simple as identifying where a period or other punctuation mark is, since not all punctuation signals the end of a sentence (consider abbreviations, for example) and not all sentences have punctuation.
- *Assigning meaning to units. Part-of-speech tagging,* involves automatically determining and assigning a part of speech (e.g., a verb or a noun) to a word is usually the first step to looking at word context and meaning. Of course, many

**Table 55.1 Overview of machine learning methods and examples from political science**

<i>Method</i>	<i>Type of learning</i>	<i>Examples</i>
Classification	Supervised	<ul style="list-style-type: none"> <li>• understand audience sentiment from social media</li> <li>• sort party manifestos into predefined categories on the ideological spectrum</li> <li>• understand US Supreme Court decisions (Evans et al., 2007)</li> <li>• extract party affiliation (Yu et al., 2008)</li> <li>• measure polarization (Peterson and Spirling, 2018)</li> </ul>
Clustering	Unsupervised	<ul style="list-style-type: none"> <li>• understand political ideologies and parliamentary topics (Preoțiu-Pietro et al., 2017; Sanders et al., 2017)</li> </ul>
Dimensionality reduction, e.g., topic modeling, feature extraction	Semi-supervised Unsupervised	<ul style="list-style-type: none"> <li>• preferences for foreign aid (Baker, 2015)</li> <li>• ideological mapping of candidates and campaign contributors (Bonica, 2014)</li> <li>• extraction of text features from documents (Uysal, 2016; Nguyen et al., 2015)</li> </ul>

words have more than one meaning or could be assigned different parts of speech, which can prove challenging for NLP, as it needs to select the meaning which makes more sense in the current context. With the emergence of deep learning methods, word embeddings have been used to capture semantic properties of words and their context (see the next section for a more detailed presentation).

- *Extracting information from the text and synthesizing it.* NLP tasks such as *Named Entity Recognition*, *Sentiment Analysis*, *Machine Translation*, and *Automated Text Summarization* build on the above tasks in order to identify and extract specific content from texts and synthesize it to generate new insights or content.
- *Machine Translation* studies ways to automate the translation between languages. Deep learning methods are improving the accuracy of algorithms for this task (Nallapati et al., 2016). This leads to scaling-up opportunities in comparative politics research (de Vries et al., 2018).
- *Named Entity Recognition* helps determine the elements in a text that are proper names (such as people or places) and what type of elements they are (e.g., person, location, organization).
- *Sentiment Analysis* is the automatic extraction of opinions or subjective information from a set of documents or reviews, to determine ‘polarity’ about specific ideas. For example, scholars have used *Sentiment Analysis* to identify trends of public opinion in social media (Ceron et al., 2014; Proksch et al., 2015).
- *Automated Text Summarization* is a common dimensionality-reduction task in machine learning and NLP. It involves producing a readable, coherent, and fluent summary of a longer text, which should include the main points outlined in the document. *Extractive summarization* involves techniques such as identifying key words from the source document and combining them into a continuous text to make a summary. *Abstractive summarization* involves automatically paraphrasing or shortening parts of the original text.

With the deep learning methods being extremely data hungry, we believe that a primary area where the field will benefit from the latest technology is in the text-as-data or broader NLP domain. In what follows, we outline several deep learning models that

have made recent advances in NLP possible and highlight how they can be used in political science research.

## DEEP LEARNING NLP FOR POLITICAL ANALYSIS

### *Understanding ‘Learning’*

To define *deep learning* and understand the difference between deep learning and other machine learning approaches, first we need some idea of what machine learning algorithms *do*. As mentioned above, the field of machine learning is concerned with the question of how to construct computer programs that automatically improve with experience.

But what does *learning* mean in this context?

A computer program is said to learn from experience  $E$  with respect to some class of tasks  $T$  and performance measure  $P$  if its performance at tasks in  $T$ , as measured by  $P$ , improves with experience  $E$ . (Mitchell, 1997; our emphasis)

This type of learning that particularly pertains to NLP regardless of the type of learning (supervised, unsupervised, active, etc.) is very much based on a ‘bag-of-words’ approach that only considers one dimension of the text, without taking onboard any of the contextual information – a rather ‘shallow’ type of learning.

Deep learning, on the other hand, offers the potential to combine multiple *layers* of representation of information, sometimes grouped in a hierarchical way.

### *Understanding ‘Deep’*

Deep learning is a type of machine learning (representation learning) that enables a machine to automatically learn the patterns needed to perform regression or classification when provided with raw data. The approach puts an emphasis on learning

successive *layers* of increasingly meaningful representations. It involves multiple levels of representation. Deng (2014: 199–200) defines deep learning as a class of machine learning algorithms that

- use a cascade of multiple layers of non-linear processing units for feature extraction and transformation, and each successive layer uses the output from the previous layer as input;
- learn in supervised (e.g., classification) and/or unsupervised (e.g., pattern analysis) manners; and
- learn multiple levels of representations that correspond to different levels of abstraction – the levels form a hierarchy of concepts.

In deep learning, each level learns to transform its input data into a slightly more abstract and composite representation. In an image-recognition application, the raw input may be a matrix of pixels, the first representational layer may abstract the pixels and encode edges, the second layer may compose and encode the arrangements of edges, the third layer may encode eyes and a nose, and the fourth layer may recognize that the image contains a face (for more information about

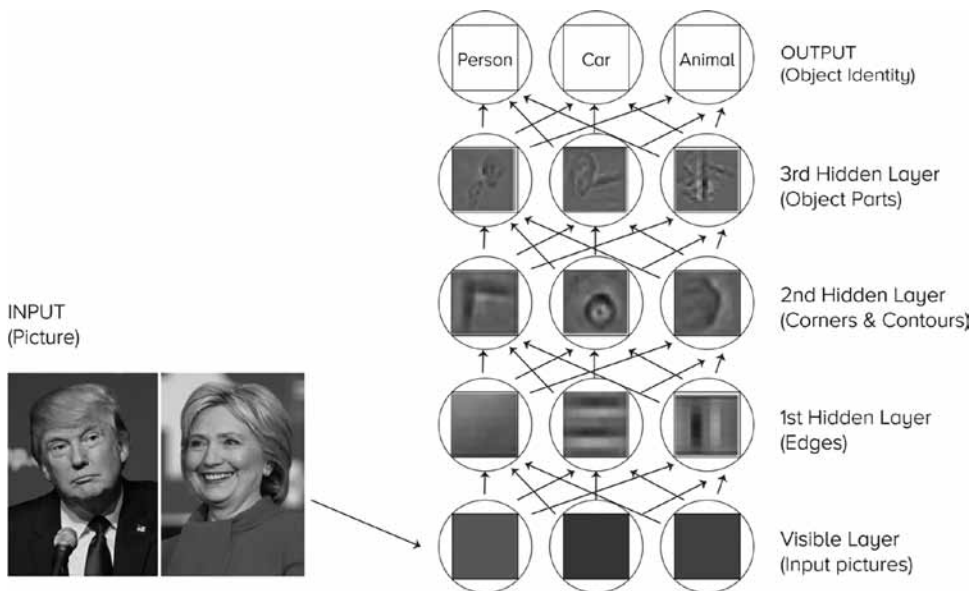
feature visualizations from computer-vision deep neural networks, see Olah et al., 2017 and Zhang and Zhu, 2018). Importantly, a deep learning process can learn which features to optimally place in which level *on its own*. Figure 55.2 shows how a deep learning hierarchy of complex concepts can be built from simpler concepts.

We will next discuss the application of deep learning algorithms in generating insights from images and text data.

### Working with Image Data

Convolutional neural networks (CNNs) are a category of artificial neural networks that have proven very effective when trying to classify or detect features in images. CNNs have been very successful at identifying objects, faces, and traffic signs in images and are currently advancing computer vision in robotics and self-driving vehicles.

CNNs have been trained on satellite imagery to map and estimate poverty, where data on economic livelihoods are scarce and



**Figure 55.2 Building a hierarchy of complex concepts out of simpler concepts**

where outcomes cannot be studied via other data. Jean et al. (2016) combine satellite imagery with survey data from five African countries (Nigeria, Tanzania, Uganda, Malawi, and Rwanda) to train a CNN to identify image features that can explain up to 75% of the variation in the local-level economic outcomes by estimating consumption expenditure. Figure 55.3 shows four different convolutional filters used for extracting these features, which identify (from left to right) features corresponding to urban areas, non-urban areas, water, and roads. Babenko et al. (2017) focus on an urban subsample of satellite images in Mexico (using images from Digital Globe and Planet) identifying rural and urban ‘pockets’ of poverty that are inaccessible and changing frequently – areas that are unlikely to integrate without the support of the necessary policy measures (Figure 55.4).

CNNs have also been used to map informal settlements (‘slums’) in developing countries, using high- and low-resolution satellite imagery (Helber et al., 2018), to help international aid organizations to provide effective social and economic aid.

But how do they work?

Analogous to how children learn to recognize a cat from a dog, we need to ‘show’ an



**Figure 55.3 Visualization of features. Four different convolutional filters (which identify, from left to right, features corresponding to urban areas, non-urban areas, water, and roads) in the convolutional-neural-network model used for extracting**

Source: Jean et al. (2016).

algorithm millions of pictures (‘input’) of a dog before it can reliably make generalizations and predictions for images it has never seen before. However, machines do not ‘see’ in the same way we do – their ‘language’ consists of numbers. One way around this is to represent every image as multi-dimensional arrays of numbers, and CNNs offer a way to move from an image to a set of vectors.



**Figure 55.4 Examples of Digital Globe (left) and Planet (right, Michoacán) imagery**

Source: Babenko et al. (2017).

The main building block of CNN is the *convolutional layer, filter, or kernel*. Convolution is a mathematical operation that allows us to condense information by combining two functions into one. Take the very simple, pixelated representation of a black and white heart in Figure 55.5 element (a) for example. If each cell is a pixel, then we could represent black pixels with value 1 and white pixels with value 0 (see Figure 55.5, element (b)) – this is the ‘input’.

Using a filter, as in Figure 55.5 element (c), with predefined black and white pixels, we can now perform a convolution and create a ‘feature map’ (Figure 55.6, element (d))

by layering the filter on top of the input and sliding it for each row. At every step, we perform element-wise matrix multiplication and sum the result, which goes into the feature map – represented in the black background in Figure 55.6.

We then slide the filter over the next position and perform the same multiplication (see Figure 55.7).

We repeat until the ‘input’ is reduced from a 5x5 matrix to a 3x3 feature map, as in Figure 55.8 element (c). The example above is a two-dimensional convolution using a 3x3 filter – in reality, these convolutions are performed in three dimensions (width, height,

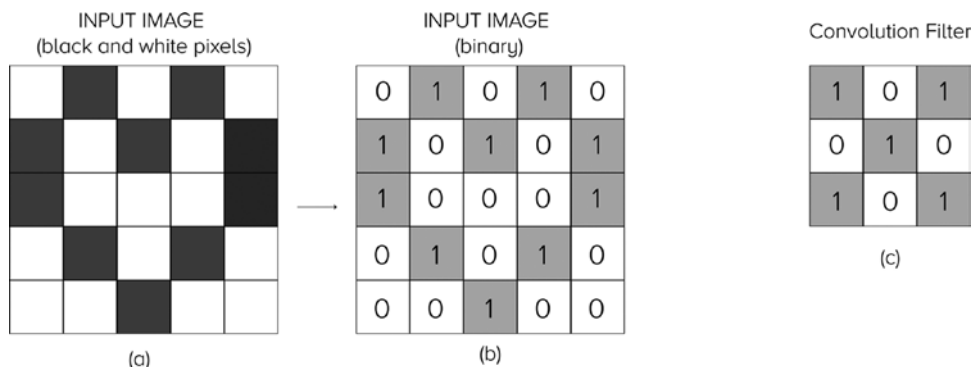


Figure 55.5 Convolution of a black and white image of a heart – the essential elements

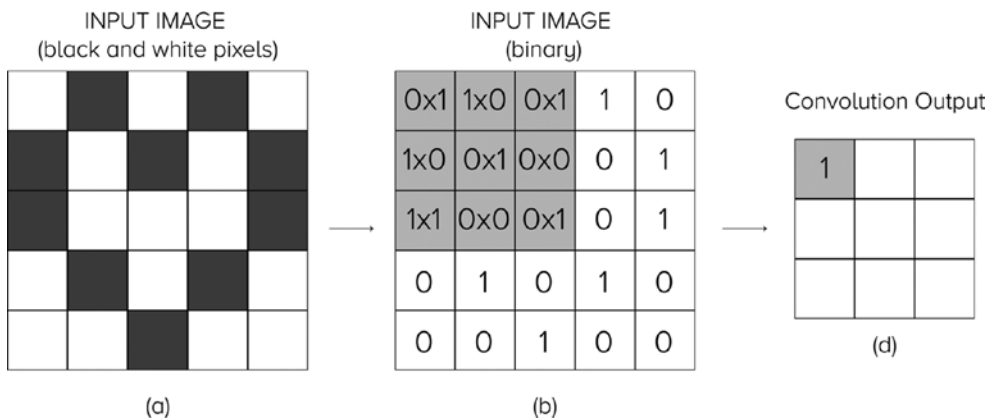
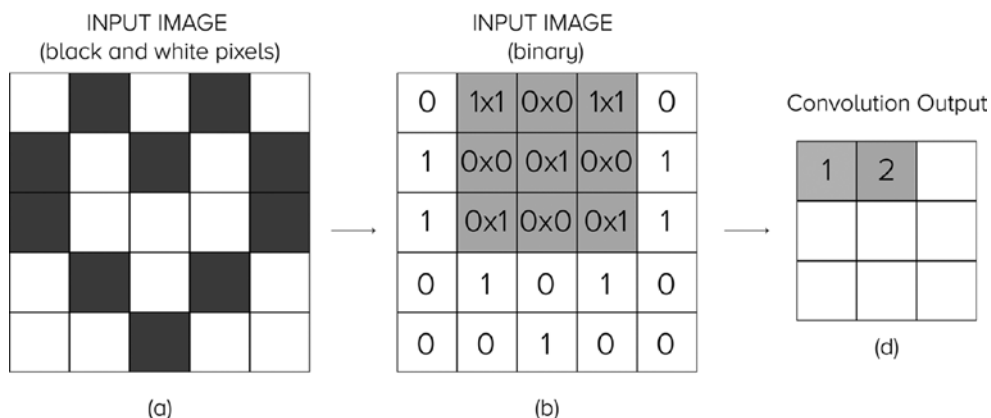
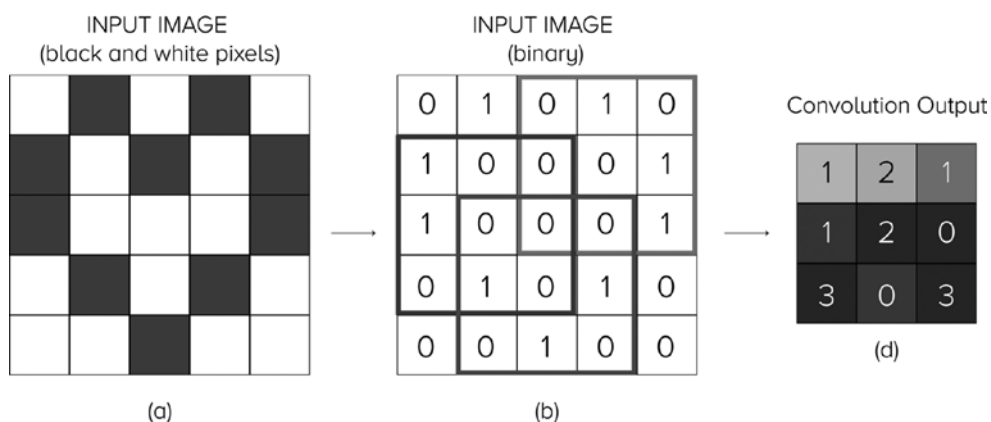


Figure 55.6 Convolution of a black and white image of a heart – step one



**Figure 55.7 Convolution of a black and white image of a heart – step two**



**Figure 55.8 Convolution of a black and white image of a heart – step three**

and RGB color channel) with the filter being 3D as well. Multiple convolutions take place on an input, each using a different filter with a distinct feature map as the output. After a convolution operation, we usually perform *pooling* (usually *max pooling*, i.e., taking the max value in the pooling window) to reduce the dimensionality and reduce the number of parameters (see Figure 55.9).

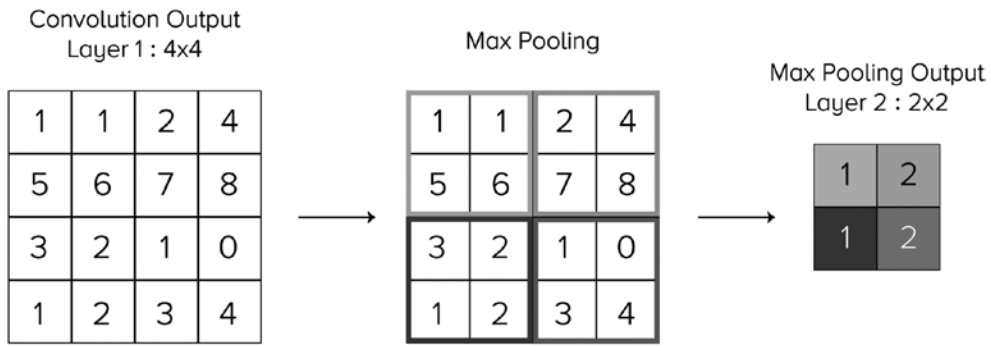
This is crucial when dealing with the volume of data that is fed to the algorithm, as it both speeds training time and helps avoid overfitting of the algorithm.

CNNs seem to suit the task of image classification, as they can help us predict

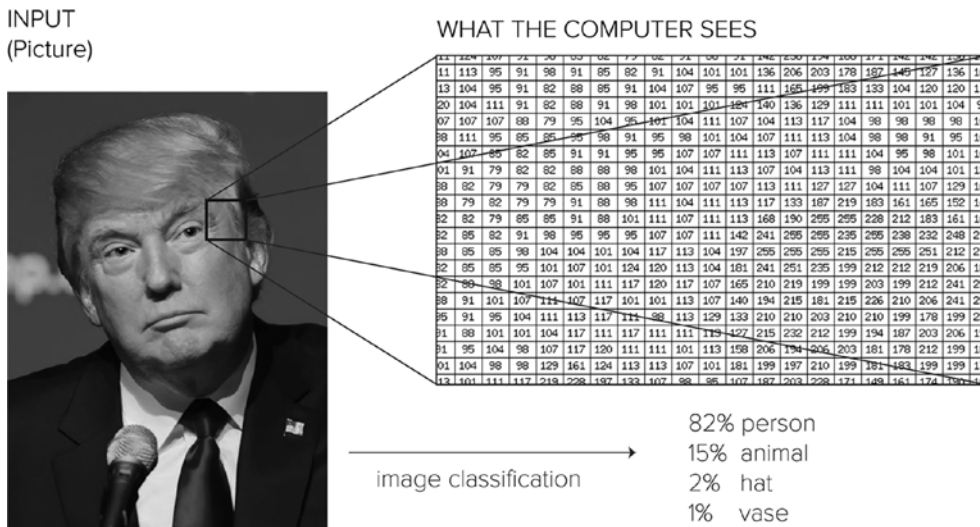
a distribution over specific labels (as in Figure 55.10) to indicate confidence of prediction for a given image. But what about text data?

### **Working with Text Data**

The study of political discourse using text as data has a long tradition in political science. Political texts have long been used as an important form of social practice that contributes to the construction of social identities and relations (Fairclough, 1989, 1992; Laclau and Mouffe, 1985). Text as a representation



**Figure 55.9** Max pooling operation – using a 2x2 input window and stride 2. Each 2x2 outline color maps onto a max-pooling single output on the 2x2 layer



**Figure 55.10** The task in image classification is to predict a single label (or a distribution over labels as shown here – percentages indicating level of confidence) for a given input image. Images are three-dimensional arrays of integers from 0 to 255, of size width x height x 3. The 3 represents the three color channels (red, green, blue)

of discourses has been studied systematically to derive information about actors and combine them with additional resources such as surveys and observations, as well as knowledge and reflective understanding of the context by scholars, yet not in a reproducible and quantifiable way (see Blommaert and Bulcaen, 2000, for a review).

Over the past two decades, scholars have sought to extract information such as

policy and ideology positions and gauge citizen political engagement by treating words as data in a more consistent way. Since some of the earliest implementations of text-scaling methods such as *Wordscores* (Laver et al., 2003) and *Wordfish* (Slapin and Proksch, 2008) to estimate party positions from texts and the increasing availability of annotated political corpora, the availability and complexity of quantitative text-analysis methods

have increased dramatically (Barberá, 2015; Grimmer and Stewart, 2013; Herzog and Benoit, 2015; Lauderdale and Herzog, 2016). Most of these methods tend to involve a ‘bag-of-words’ approach to determine relevance and cluster documents or their parts in groups (see also Laver, 2014). Such approaches assume that each document can be represented by a multiset (‘bag’) of its words, that ignores word order and grammar. Word frequencies in the document are then used to classify the document into a category. Some methods like Wordscores employ a version of the Naive Bayes classifier (Benoit and Nulty, 2013) in a supervised learning setting by leveraging pre-labeled training data, whereas others, like WordFish, are based on a Poisson distribution of word frequencies, with ideological positions estimated using an expectation-maximization algorithm (Proksch and Slapin, 2009; Slapin and Proksch, 2008).

What these approaches do not capture, though, is the linguistic and semiological context, i.e., the information provided by the words around the target elements. Such a context would allow for a better representation of that context and offer a richer understanding of word relationships in a political text. One way to do that is by using *word embeddings*, a set of methods to model language, combining concepts from NLP and graph theory.

### *Representing words in context: word embeddings*

Word embeddings are a set of language modeling and dimensionality-reduction techniques, where words or phrases from a document are mapped to vectors or numbers. They usually involve a mathematical embedding from a space with a single dimension for each word to a continuous vector space with a reduced dimension. The underlying idea is that ‘[y]ou shall know a word by the company it keeps’ (Firth, 1957: 11), and it has evolved from ideas in structuralist linguistics and ordinary language philosophy, as expressed in the work of Zellig Harris, John Firth, Ludwig Wittgenstein, and vector-space models for information

retrieval in the late 1960s to the 1980s. In the 2000s, Bengio et al. (2006) and Holmes and Jain (2006) provided a series of papers on the ‘Neural Probabilistic Language Models’ in order to address the issues of dimensionality of word representations in contexts, by facilitating learning of a ‘distributed representation of words’. The method developed gradually and really took off after 2010, partly due to major advances in the quality of vectors and the training speeds of the models.

There are many variations of word-embedding implementations, and many research groups have created similar but slightly different types of word embeddings that can be used in the deep learning pipelines. Popular implementations include Google’s Word2Vec (Mikolov et al., 2013), Stanford University’s GloVe (Pennington et al., 2014), Facebook’s fastText (Bojanowski et al., 2016) and Allen Institute for AI’s ELMo (Peters et al., 2018). For a recent discussion of word embeddings in a political science context, see Spirling and Rodriguez (2019).

Now that we have a mechanism to turn text into dense vectors (very much like we did with the image of the heart in the previous section), let’s see how CNNs can be applied to NLP tasks for political texts.

### *CNNs for text analysis*

CNNs have recently been applied to various NLP tasks with very good results in accuracy and precision (Johnson and Zhang, 2014; Kalchbrenner et al., 2014; Kim, 2014).

Instead of image pixels, each row of the matrix corresponds to one token (usually a word, but it could also be a character; see Jacovi et al., 2018 and Zhang et al., 2015) or rather a vector that represents a word. These vectors are typically *word embeddings* such as Word2Vec or GloVe (see previous section). Kim (2014) describes the general approach of using CNNs for NLP, assuming a single layer of networks and pretrained static word vectors on very large corpora (Word2Vec vectors from Google, trained on 100 billion tokens from Google News). Sentences are mapped to embedding

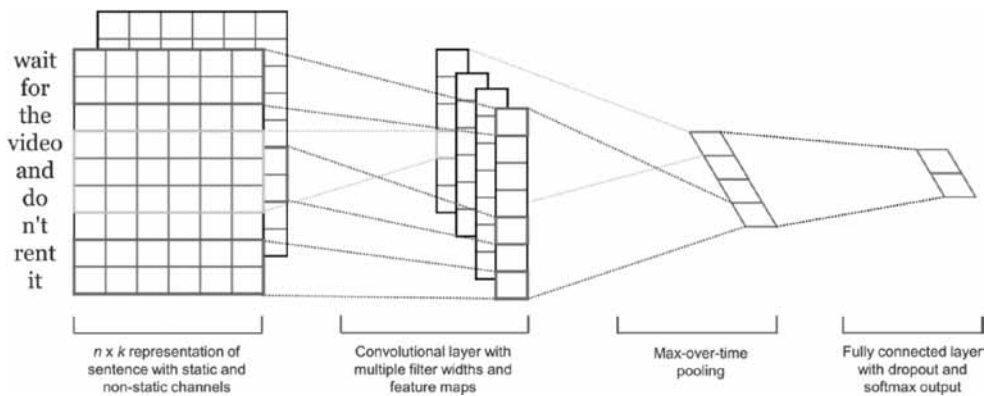


vectors and are available as a matrix input to the model. Convolutions are performed across the input word-wise using differently sized kernels, such as two or three words at a time. The resulting feature maps are then processed using a max pooling layer to condense or summarize the extracted features. Figure 55.11 shows a single-layer CNN architecture for sentence classification from Kim (2014).

Figure 55.12 shows how a CNN would work for a sentence-classification task adapted from Zhang and Wallace (2015). Assuming the sentence we wanted to classify was Michelle Obama’s ‘*When they go low, we go high*’, this would generate a 7×4 sentence matrix, with three filter region sizes: 2, 3, and 4, each of which has two filters for each region size. Every filter performs convolution on the sentence matrix and generates (variable-length) feature maps. Then, 1-max pooling is performed over each map, i.e., the largest number from each feature map is recorded. Thus, a univariate feature vector is generated from all six maps, and these six features are concatenated to form a feature vector for the penultimate layer. The final *softmax* layer then receives this feature vector as input and uses it to classify the sentence; here, we assume binary classification and hence depict two possible output states.

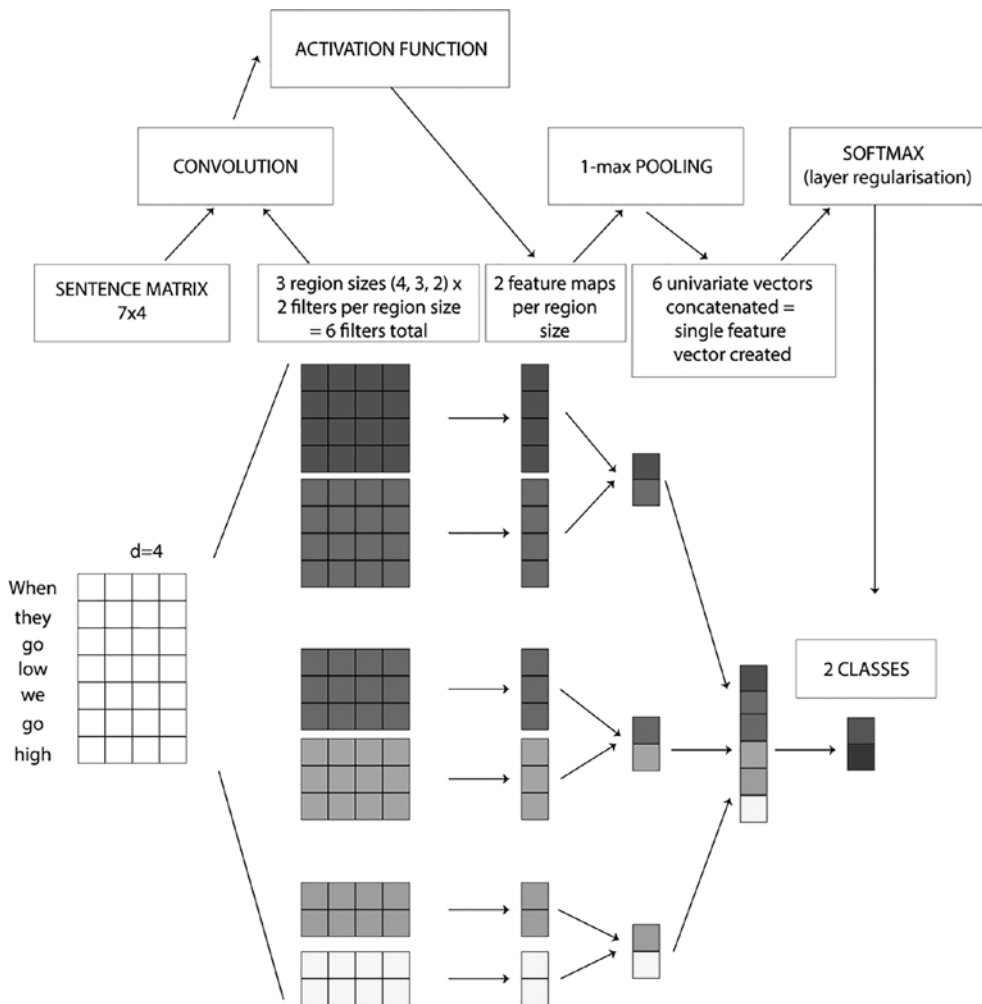
Despite CNNs being a little unintuitive in their language implementation, they perform really well on tasks like text classification. They are very fast, as convolutions are highly parallelizable, form an integral part of computer graphics, and are implemented on graphical processing units (GPUs). They also work much better compared to other ‘bag-of-words’ approaches such as n-grams, as they can learn representations automatically without the need to represent the whole vocabulary (whereas in the case of n-grams, for example, if we had a large vocabulary, computing anything beyond tri-grams would become quite expensive in terms of computational power), with architectures as deep as 29 layers performing sufficiently well (Zhang et al., 2015).

CNNs have been successfully deployed for NLP tasks such as automatic summarization, fake news detection, and text classification. Narayan et al. (2018), for example, apply CNNs to automatically summarize a real-world, large-scale dataset of online articles from the British Broadcasting Corporation (BBC). They demonstrate experimentally that this architecture captures long-range relationships in a document and recognizes related content, outperforming other state-of-the-art abstractive approaches when evaluated automatically and by humans.



**Figure 55.11** Illustration of a single-layer convolutional neural network architecture for sentence classification

Source: Kim (2014).



**Figure 55.12 Illustration of a convolutional neural network for sentence classification. The quote is attributed to Michelle Obama, later used by Hillary Clinton's campaign in the 2016 US presidential election**

Source: adapted from Zhang and Wallace (2015).

Yamshchikov and Rezagholi (2018) develop a model of binary text classifiers based on CNNs, which helps them label statements in the political programs of the Democratic and Republican parties in the United States, whereas Bilbao-Jayo and Almeida (2018) propose a new approach to automate the analysis of texts in the Manifestos Project, to allow for a quicker and more streamlined classification of such types of political texts.

The Manifesto Project (Lehmann et al., 2018) includes data on parties' policy positions, derived from content analysis of parties' electoral manifestos. It covers over 1,000 parties from 1945 until today in over 50 countries on five continents. The corpus includes manually annotated election manifestos using the Manifesto Project coding scheme, which is widely used in comparative politics research. Bilbao-Jayo and Almeida (2018) use

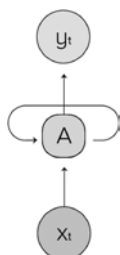
multi-scale CNNs with word embeddings and two types of context data as extra features, like the previous sentence in the manifesto and the political party. Their model achieves reasonably high performance of the classifier across several languages of the Manifesto Project.

Another type of neural network that has shown good performance in NLP tasks are recurrent neural networks (RNNs) and, in particular, a variation of that algorithm, the long short-term memory (LSTM) RNNs.

### *LSTM RNNs for text analysis*

As you read this paragraph, you understand each word based on your understanding of previous words – those right before this word, words expressed in the paragraphs and sections above, as well as words that you might have read in the previous chapters of this *Handbook* (or even words that you have read in other books and articles).

Every time we read a new word, we do not forget what we read before – our understanding has some degree of *persistence*. Unfortunately, CNNs cannot reason about previous steps in the learning process to inform later ones. RNNs overcome this issue because they permit loops, thus allowing for the information in the neural network to persist. A simple RNN is a class of artificial neural networks where connections between nodes form a directed graph along a sequence, incorporating previous knowledge (see Figure 55.13, adapted from Olah, 2015).



**Figure 55.13** A simple RNN network with a feedback loop. A simple RNN, A, looks at some input,  $x_t$ , and outputs a value,  $y_t$

Source: adapted from Olah (2015).

A sequence of RNN blocks can be regarded as multiple copies of the same network, linked to one another like a chain, each passing an input to its future self (Figure 55.14). This enables it to display dynamic temporal behavior for a time sequence and make these networks work really robustly with sequence data such as text, time-series data, videos, and even DNA sequences.

This suits textual data, which for the most part is sequence or list data, and which has been applied with success to NLP tasks such as speech recognition, language modeling, translation, and image captioning (Ba et al., 2014; Gregor et al., 2015). However, simple RNNs are not well suited for remembering information that is not close to the current node they are in (also called long-distance dependencies), a problem detailed in Bengio et al. (1994).

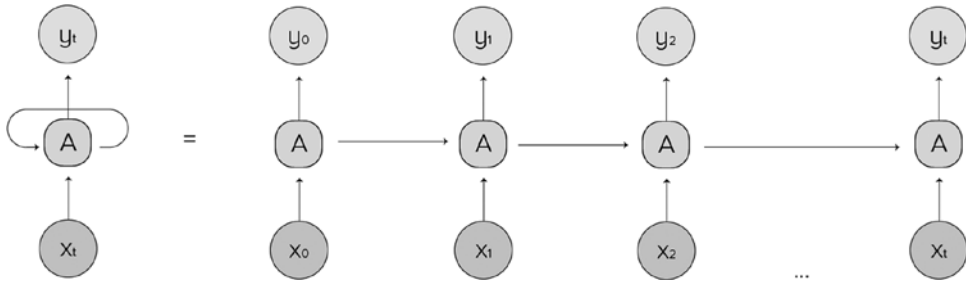
LSTM neural networks (Hochreiter and Schmidhuber, 1997) provide a solution to this issue. LSTMs also have the RNN chain-like structure, but the repeating module has a different structure. Instead of having a single neural network layer, there are four, all interacting in a special way. Figure 55.15 shows the repeating module in a standard RNN with a single layer (A1) and an LSTM with four interacting layers (A2). The LSTM has the advantage of incorporating context from both the input ( $\times$ ) and the previous knowledge (represented with dashed lines in A2) and also feed the augmented knowledge to the next iteration.

Standard LSTMs (like those in Figure 55.15) are *unidirectional* – in other words, they preserve information from the past inputs that have already passed through the different iterations of the hidden layers of the neural network. Let us now consider the following word sequence:

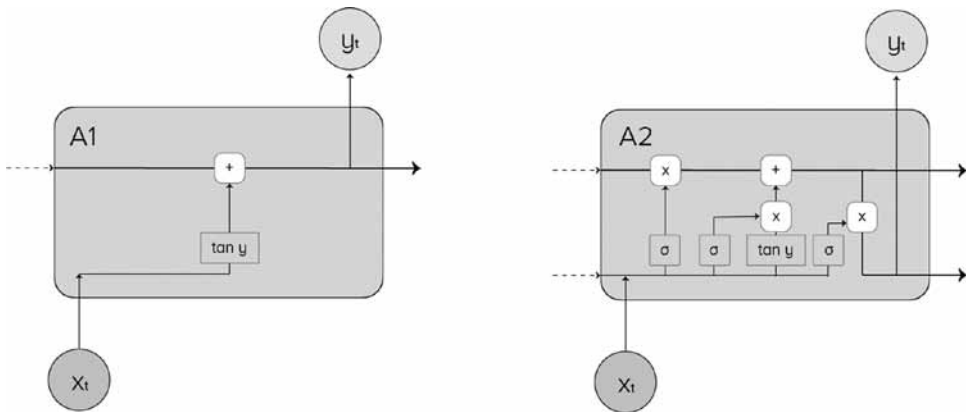
'Let's make ...'

There are a lot of possibilities for what word sequences could follow. All the sentences below are possible:

'Let's make some cake!'



**Figure 55.14** A sequence of simple RNNs



**Figure 55.15** The repeating module in a standard RNN with a single layer (A1) and an LSTM with four interacting layers (A2). The LSTM has the advantage of incorporating context from both the input ( $x$ ) and the previous knowledge (represented with dashed lines in A2), as well as feeding the augmented knowledge to the next iteration

‘Let’s make fun of Bob!’

‘Let’s make my friend see some sense, because I think she is making a huge mistake!’

What if you knew that the words that followed the first word sequence were actually these?

‘Let’s make ... great again!’

Now the range of options is narrower, and it is easy to predict that the next word is probably a noun phrase such as ‘America’ or ‘this business’.

A unidirectional LSTM will only be able to consider past input (‘let’s make’). If you wish to see the future, you would need to use a

*bidirectional* LSTM, which will run the input in two ways: one from the past to the future and one from the future to the past. When running backwards, it preserves information from the future, and by combining this knowledge with the past, it provides improved and more contextualized predictions.

Both types of LSTMs have been used to detect fake news and propaganda discourse in traditional and social media text, where the problem of detecting bots – automated social media accounts governed by software but disguised as human users – has strong societal and political implications.

Kudugunta and Ferrara (2018) propose a deep neural network based on contextual LSTM architecture that exploits both content

and metadata to detect bots at the tweet level. Their proposed technique is based on synthetic minority oversampling to generate a large labeled dataset suitable for deep nets training, from a minimal amount of labeled data (roughly 3,000 examples of sophisticated Twitter bots). The proposed model can, from the first tweet, achieve high classification accuracy (> 96%) in separating bots from humans.

Event detection using neural-network algorithms on tweets describing an event is another area of application of particular interest to media agencies and policy makers. Iyyer et al. (2014) assume that an individual's words often reveal their political ideology, and they use RNNs to identify the political position demonstrated at the sentence level, reporting that their model outperforms 'bag of words' or wordlists models in both the training and a newly annotated dataset. Makino et al. (2018), for example, propose a method to input and concatenate character and word sequences in Japanese tweets by using CNNs and reporting an improved accuracy score, whereas Rao and Spasojevic (2016) apply word embeddings and LSTM to text classification problems, where the classification criteria are decided by the context of the application. They show that using LSTMs with word embeddings vastly outperforms traditional techniques, particularly in the domain of text classification of social media messages' political leaning. The research reports an accuracy of classification of 87.57%, something that has been used in practice to help company agents provide customer support by prioritizing which messages to respond to.

Other scholars have used hybrid neural-network approaches to work with text, by combining aspects of the CNN and RNN algorithms. Ajao et al. (2018), for example, propose a framework that detects and classifies fake news messages from Twitter posts, using such a hybrid of CNNs and LSTM RNNs, an approach that allows them to identify relevant features associated with fake news stories without previous knowledge of

the domain. Singh et al. (2018) use a combination of the CNN, LSTM, and bidirectional LSTM to detect (overt and covert) aggression and hate speech on Facebook and social media comments, where the rise of user-generated content in social media coupled with almost non-existent moderation in many such systems has seen aggressive content rise.

Hybrid neural-network approaches also perform well in the task of automatic identification and verification of political claims. The task assumes that given a debate or political speech, we can produce a ranked list of all of the sentences based on their worthiness for fact checking – potential uses of this would be to predict which claims in a debate should be prioritized for fact-checking. As outlined in Atanasova et al. (2018), of a total of seven models compared, the most successful approaches used by the participants relied on recurrent and multi-layer neural networks, as well as combinations of distributional representations, matching claims' vocabulary against lexicons, and measures of syntactic dependency.

### ***Working with Multimodal Data***

With the resurgence of deep learning for modeling data, the parallel progress in fields of computer vision and NLP, as well as with the increasing availability of text/image datasets, there has been a growing interest in using multimodal data that combines text with images. The popularity of crowd-sourcing tools for generating new, rich datasets combining visual and language content has been another important factor favoring multimodal input approaches.

Ramisa et al. (2018), for example, have compiled a large-scale dataset of news articles with rich metadata. The dataset, *BreakingNews*, consists of approximately 100,000 news articles collected over 2014, illustrated with one to three images and their corresponding captions. Each article is enriched with other data like related images from Google Images, tags,



**Figure 55.16** The BreakingNews dataset. The dataset contains a variety of news-related information including the text of the article, captions, related images, part-of-speech tagging, GPS coordinates, semantic topics list, or results of sentiment analysis for about 100,000 news articles. The figure shows two sample images. Such a volume of heterogeneous data makes BreakingNews a good benchmark for several tasks exploring the relation between text and images

Source: Ramisa et al. (2018).

shallow and deep linguistic features (e.g., parts of speech, semantic topics, or outcomes of a sentiment analyzer), GPS latitude/longitude coordinates, and reader comments. The dataset is an excellent benchmark for taking joint vision and language developments a step further. Figure 55.16 illustrates the different components of the Ramisa et al. (2018) *BreakingNews* corpus, which contains a variety of news-related information for about 100K news articles. The figure shows two sample images. Such a volume of heterogeneous data makes *BreakingNews* a good benchmark for several tasks exploring the relation between text and images.

The paper used CNN for source detection, geolocation prediction, and article illustration, and a mixed LSTM/CNNs model for caption generation. Overall results were very promising, especially for the tasks of source detection, article illustration, and geolocation. The automatic caption-generation task, however, demonstrated sensitivity to loosely related text and images.

Ajao et al. (2018) also fed mixed data inputs (text and images) to CNNs in order to detect fake news in political-debate speech, and they noted that except for the usual patterns in what would be considered misinformation, there

also exists some hidden patterns in the words and images that can be captured with a set of latent features extracted via the multiple convolutional layers in the model. They put forward the TI-CNN (text and image information based convolutional neural network) model, whereby explicit and latent features can be projected into a unified feature space, with the TI-CNN able to be trained with both the text and image information simultaneously.

## Recent Developments

Deep neural networks have revolutionized the field of NLP. Furthermore, deep learning in NLP is undergoing an ‘ImageNet’ moment. In a paradigm shift, instead of using word embeddings as initializations of the first layer of the networks, we are now moving to pretraining the entire models that capture hierarchical representations and bring us closer to solving complex language-understanding tasks. When the ImageNet challenge AlexNet (Krizhevsky et al., 2012) solution showed a dramatically improved performance of deep learning models compared to traditional competitors, it arguably spurred the whole deep learning research wave. Over the last 18 months, pretrained

language models have blown out of the water previous state-of-the-art results across many NLP tasks. These advances can be characterized within the broader framework of transfer learning, where the weights learned in state-of-the-art models can be used to initialize models for different datasets, and this ‘fine-tuning’ achieves superior performance even with as little as one positive example per category (Ruder et al., 2019).

One of the assumptions of standard word embeddings like Word2Vec is that the meaning of the word is relatively stable across sentences. An alternative is to develop contextualized embeddings as part of the language models. Embeddings from language models (ELMo) (Peters et al., 2018), universal language model fine-tuning (ULMFiT) (Howard and Ruder, 2018), and generative pretraining transformer (OpenAI GPT) (Radford et al., 2018) were initial extremely successful pre-trained language models.

More recently GPT2 (Radford et al. 2019) extended the previous GPT model and was used to generate realistic-sounding artificial text. Bullock and Luengo-Oroz (2019) used the pretrained GPT2 model to generate fake but natural-sounding speeches in the United Nations General Debate (see Baturo et al., 2017, for more details about the data and a substantive example). Bidirectional encoder representations from transformers (BERT) (Devlin et al., 2019) extended GPT through bi-directional training and dramatically improved performance on various metrics. While BERT was the reigning champion for several months, it may have recently been overtaken by XLNet (Yang et al., 2019), which outperforms BERT on about 20 NLP tasks.

In parallel with the advances in transfer learning, we are also further understanding what we are learning with the deep neural networks. Liu et al. (2019) show that RNNs (and LSTMs in particular) pick up general linguistic properties, with the lowest layers representing morphology and being the most transferable between tasks, middle layers representing syntax, and the highest layers

representing task-specific semantics. Large pretrained language models do not exhibit the same monotonic increase in task specificity, with the middle layers being the most transferable. Tenney et al. (2019) focus on BERT and show that the model represents the steps of the traditional NLP pipeline, with the parts-of-speech tagging followed by parsing, named-entity recognition, semantic roles, and, finally, coreference. Furthermore, the model adjusts the pipeline dynamically, taking into account complex interactions between different levels of hierarchical information.

Detailed discussion of the above models is beyond the scope of this chapter. Instead, we want to emphasize the pace of development in NLP research, which is leveraging pretrained language models for downstream tasks. Instead of downloading pretrained word embeddings like Word2Vec or GloVe as discussed earlier in the chapter, we are now in a position to download pretrained language models and fine-tune them to a specific task.

## CONCLUSION

It is appealing to think of machine learning algorithms as objective, unbiased actors that are beyond the influence of human prejudices. It is also appealing to think of empirical research in political science that utilizes machine learning algorithms as being sufficiently removed from any potential bias. Unfortunately, this is rarely the case.

Algorithms are designed by humans and learn by observing patterns in the data that very often represent biased human behavior. It is no surprise that algorithms tend to adopt and, in some occasions, perpetuate and reinforce the experiences and predispositions of the humans that have constructed them and those of society as a whole; this is also known as *algorithmic bias*. Although machine learning has been transformative in many fields, it has received criticism in the areas of causal inference, algorithmic bias, and data privacy.

This is forming into a distinct area of social science research, focusing on the lack of (suitable) training data, difficulties of data access and data sharing, data bias and data provenance, privacy preserving data usage, and inadequate tasks, tools, and evaluation settings (Danks and London, 2017).

The quality of insights delivered by algorithms crucially depends on data quality and data provenance. In particular, in each case, we need to effectively query very distinct (heterogeneous) data sources before we can extract and transform them for input into the data models. Common aspects of data quality that may affect the robustness of insights include consistency, integrity, accuracy, and completeness. How image or textual data is preprocessed may affect how data is interpreted and may also lead to biases. For example, dataset biases in computer vision can lead to feature representation flaws where CNNs, despite high accuracy, learn from unreliable co-appearing contexts (Zhang et al., 2018).

The consequences of biased algorithms can be quite real and severe. In 2016, an investigative study by ProPublica (Angwin et al., 2016) provided evidence that a risk-assessment machine learning algorithm used by US courts wrongly flagged non-white defendants at almost twice the rate of white defendants. More recently, Wang and Kosinski (2018) showed how deep neural networks can outperform humans in detecting sexual orientation. Apart from the ethical issues of the study, the ease of deployment of such ‘AI Gaydar’ raises issues of people’s privacy and safety.

The issues of algorithmic bias are also highlighted in the Wellcome Trust Report (Matthew Fenech et al., 2018) with a focus on how AI has been used for health research. The report identifies, among other ethical, social, and political challenges, issues around implications of algorithmic transparency and explainability on health, the difference between an algorithmic decision and a human decision, and what makes algorithms, and the entities that create them, trustworthy. The report highlights the importance of

stakeholders across the public- and private-sector organizations collaborating in the development of AI technology, and it raises awareness of the need for AI to be regulated.

Such algorithmic-bias issues may seem to be removed from everyday political science research. However, various methodological approaches discussed earlier in this chapter are not bias free. Word embeddings have been shown to carry societal biases that are encoded in human language (Garg et al., 2018). These range from biased analogies (Bolukbasi et al., 2016; Manzini et al., 2019; Nissim et al., 2019) to bias in language ID (Blodgett and O’Connor, 2017), natural-language inference (Rudinger et al., 2017), coreference resolution (Rudinger et al., 2018), and automated essay scoring (Amorim et al., 2018).

There are corresponding efforts to reduce algorithmic bias in deep neural-network applications, for example, through postprocessing (Bolukbasi et al., 2016) or directly modeling the problem (Zhao et al., 2018). However, the bias still remains encoded implicitly (Gonen and Goldberg, 2019), and transparency and awareness about the problem may be better as a research and deployment strategy (Caliskan et al., 2017; Dwork et al., 2012; Gonen and Goldberg, 2019).

There are legitimate concerns about algorithmic bias and discrimination, algorithmic accountability and transparency, and general ‘black box’ perception of deep neural-network models (Knight, 2017; Mayernik, 2017). In order to address these issues, scholars (Fiesler and Proferes, 2018; Mittelstadt et al., 2016; Olhede and Wolfe, 2018; Prates et al., 2018), AI technologists, international organizations (European Group on Ethics in Science and New Technologies (EGE), 2018), and national governments (House of Lords Select Committee, 2018) have been recently advocating for a more ‘ethical’ and ‘beneficial’ AI that will be programmed to have humans’ interests at heart and could never hurt anyone.

Kusner et al. (2017), for example, provide an ethical framework for machine



decision-making, whereby a ‘decision is considered fair towards an individual if it is the same in both the actual world and a “counterfactual” world, where the individual would belong to a different demographic group’. In addition, it is vital to think about who is being excluded from AI systems and what is missing from the datasets that drive machine learning algorithms. Often, these blind spots tend to produce disparate impacts on vulnerable and marginalized groups. This leads to the invisibility of these communities and their needs because there are not enough feedback loops for individuals to give their input. While the collection of even more personal data might make algorithmic models better, it would also increase the threats to privacy.

Russell et al. (2015) present relevant questions to be considered: what are the power dynamics between different industry and research groups? Will the interests of the research community change with greater state funding? Will government intervention encourage AI research to become less transparent and accountable? What organizational principles and institutional mechanisms exist to best promote beneficial AI? What would international cooperation look like in the research, regulation, and use of AI? Will transnational efforts to regulate AI fall to the same collective-action problems that have undermined global efforts to address climate change?

To ensure that future iterations of the ethical principles are adopted widely around the world, further research will be needed to investigate long-standing political questions such as collective action, power, and governance, as well as the global governance of AI, to name a few.

## REFERENCES

Ajao, O., Bhowmik, D. and Zargari, S. (2018) Fake News Identification on Twitter with Hybrid CNN and RNN Models. In: Proceedings of the 9th International Conference on Social Media and Society – SMSociety,

- Copenhagen, Denmark, pp. 226–230. New York: ACM.
- Aletras, N., Tsarapatsanis, D., Preotiuc-Pietro, D. and Lampos, V. (2016) Predicting judicial decisions of the European Court of Human Rights: A natural language processing perspective. *PeerJ Computer Science* 2: e93.
- Amorim, E., Cançado, M. and Veloso, A. (2018) Automated Essay Scoring in the Presence of Biased Ratings. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1 (Long Papers), New Orleans, United States: Association for Computational Linguistics, pp. 229–237. Available at: <https://doi.org/10.18653/v1/N18-1021> (accessed 17 December 2018).
- Angwin, J., Larson, J., Mattu, S. and Kirchner, L. (2016) Machine Bias. *ProPublica*, 23 May, 2016. Available at: <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing> (accessed 17 December 2018).
- Atanasova, P., Barron-Cedeno, A., Elsayed, T., Suwaileh, R., Zaghouni, W., Kyuchukov, S., Da San Martino, G. and Nakov, P. (2018) Overview of the CLEF-2018 CheckThat! Lab on Automatic Identification and Verification of Political Claims. Task 1: Check-Worthiness. arXiv:1808.05542 [cs]. Available at: <http://arxiv.org/abs/1808.05542> (accessed 19 December 2018).
- Ba, J., Mnih, V. and Kavukcuoglu, K. (2014) Multiple Object Recognition with Visual Attention. arXiv:1412.7755 [cs]. Available at: <http://arxiv.org/abs/1412.7755> (accessed 19 December 2018).
- Babenko, B., Hersh, J., Newhouse, D., Ramakrishnan, A. and Swartz, T. (2017) Poverty Mapping Using Convolutional Neural Networks Trained on High and Medium Resolution Satellite Images, With an Application in Mexico. arXiv:1711.06323 [cs, stat]. Available at: <http://arxiv.org/abs/1711.06323> (accessed 18 December 2018).
- Baker, A. (2015) Race, paternalism, and foreign aid: Evidence from US public opinion. *American Political Science Review* 109(1): 93–109.
- BakIr, G. (ed.) (2007) Predicting Structured Data. *Advances in Neural Information Processing Systems*. Cambridge, MA: MIT Press.

- Barberá, P. (2015) Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data. *Political Analysis* 23(1): 76–91.
- Baturo, A., Dasandi, N. and Mikhaylov, S. J. (2017) Understanding state preferences with text as data: Introducing the un general debate corpus. *Research & Politics* 4(2): 1–9. Available at <https://journals.sagepub.com/doi/pdf/10.1177/2053168017712821> (accessed 19 December 2019)
- Bengio, Y., Simard, P. and Frasconi, P. (1994) Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks* 5(2): 157–166.
- Bengio, Y., Schwenk, H., Senécal, J.-S., Morin, F. and Gauvin, J.-L. (2006) Neural Probabilistic Language Models. In: Holmes, D. E. and Jain, L. C. (eds), *Innovations in machine learning*. Berlin/Heidelberg: Springer-Verlag, pp. 137–186.
- Benjamins, V. R., Selic, B., Casanovas, P., Breuker, J. and Gangemi, A. (2005) *Law and the Semantic Web: Legal Ontologies, Methodologies, Legal Information Retrieval, and Applications*. Berlin Heidelberg: Springer.
- Benoit, K. and Nulty, P. (2013) Classification methods for scaling latent political traits. In: *Presentation at the Annual Meeting of the Midwest Political Science Association*, Chicago, United States, pp. 11–13.
- Bilbao-Jayo, A. and Almeida, A. (2018) Automatic political discourse analysis with multi-scale convolutional neural networks and contextual data. *International Journal of Distributed Sensor Networks* 14(11): 1–11.
- Bishop, C. M. (2006) *Pattern Recognition and Machine Learning*. Information Science and Statistics. New York: Springer.
- Blodgett, S. L. & O'Connor, B. (2017) Racial Disparity in Natural Language Processing: A Case Study of Social Media African-American English. In: *2017 Workshop on Fairness, Accountability, and Transparency in Machine Learning*, Nova Scotia, Canada. arXiv preprint arXiv:1707.00061: 1–4.
- Blommaert, J. and Bulcaen, C. (2000) Critical discourse analysis. *Annual Review of Anthropology* 29: 447–466.
- Bojanowski, P., Grave, E., Joulin, A. and Mikolov, T. (2016) Enriching Word Vectors with Subword Information. arXiv:1607.04606 [cs]. Available at: <http://arxiv.org/abs/1607.04606> (accessed 19 December 2018).
- Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V. and Kalai, A. T. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In: Lee, D. D., von Luxburg, R. Garnett, M. Sugiyama, and Guyon, I, (eds). *Advances in Neural Information Processing Systems 29: 30th Annual Conference on Neural Information Processing Systems 2016: Barcelona, Spain, 5–10 December 2016*, Red Hook, NY: Curran Associates, Inc.: 4349–4357.
- Bonica, A. (2014) Mapping the ideological marketplace. *American Journal of Political Science* 58(2): 367–386.
- Brynjolfsson, E., Mitchell, T. and Rock, D. (2018) What Can Machines Learn, and What Does It Mean for Occupations and the Economy? In: *AEA Papers and Proceedings*, Nashville, TN: American Economic Association, 108: 43–47.
- Bullock, J. and Luengo-Oroz, M. (2019) Automated Speech Generation from UN General Assembly Statements: Mapping Risks in AI Generated Texts. In: *The 2019 International Conference on Machine Learning AI for Social Good Workshop*, Long Beach, United States: 1–5. Available at <http://arxiv.org/abs/1906.01946v1> (accessed 15 October 2019).
- Caliskan, A., Bryson, J. J. and Narayanan, A. (2017) Semantics derived automatically from language corpora contain human-like biases. *Science* 356(6334): 183–186.
- Ceron, A., Curini, L., Iacus, S. M. and Porro, G. (2014) Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France. *New Media & Society* 16(2): 340–358.
- Danks, D. and London, A. J. (2017) Algorithmic bias in autonomous systems. In: *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17)*, Melbourne Australia. Red Hook, NY: Curran Associates, Inc: 4691–4697.
- de Vries, E., Schoonvelde, M. and Schumacher, G. (2018) No longer lost in translation: Evidence that Google Translate works for comparative bag-of-words text applications. *Political Analysis* 26(4): 417–430.

- Deng, L. (2014) Deep learning: Methods and applications. *Foundations and Trends® in Signal Processing* 7(3–4): 197–387.
- Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2019) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol 1 (Long and Short papers), Minneapolis, United States: Association for Computational Linguistics: 4171 – 4186.
- Dwork, C., Hardt, M., Pitassi, T., Reingold, O. and Zemel, R. (2012) Fairness through awareness. In: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, Cambridge, United States: ACM. pp. 214–226.
- European Commission (2019) A Definition of AI: Main Capabilities and Scientific Disciplines. *High-Level Expert Group on Artificial Intelligence*, 8 April. Available at <https://web.archive.org/web/20191014134019/https://ec.europa.eu/digital-single-market/en/news/definition-artificial-intelligence-main-capabilities-and-scientific-disciplines> (accessed 14 October 2019).
- European Group on Ethics in Science and New Technologies (EGE) (2018) Statement on Artificial Intelligence, Robotics and ‘Autonomous’ Systems. EU report. Available at: [https://web.archive.org/web/20191014135156/http://ec.europa.eu/research/ege/pdf/ege\\_ai\\_statement\\_2018.pdf](https://web.archive.org/web/20191014135156/http://ec.europa.eu/research/ege/pdf/ege_ai_statement_2018.pdf) (accessed 14 October 2019).
- Evans, M., McIntosh, W., Lin, J. and Cates, C. L. (2007) Recounting the courts? Applying automated content analysis to enhance empirical legal research. *Journal of Empirical Legal Studies* 4(4): 1007–1039.
- Fairclough, N. (1989) *Language and Power*. London; New York: Longman.
- Fairclough, N. (1992) Discourse and text: Linguistic and intertextual analysis within discourse analysis. *Discourse & Society* 3(2): 193–217.
- Fenech, M., Strukelj, N. and Buston, O. (2018) *Ethical, Social and Political Challenges of Artificial Intelligence in Health*. London: Wellcome Trust and Future Advocacy. Available at: <https://wellcome.ac.uk/sites/default/files/ai-in-health-ethical-social-political-challenges.pdf> (accessed 21 December 2018).
- Fiesler, C. and Proferes, N. (2018) ‘Participant’ Perceptions of Twitter Research Ethics. *Social Media + Society* 4(1): 1–14.
- Firth, J. (1957) A synopsis of linguistic theory 1930–1955. In: *Studies in Linguistic Analysis*. Oxford: Philological Society.
- Garg, N., Schiebinger, L., Jurafsky, D. and Zou, J. (2018) Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16): E3635–E3644.
- Goldberg, D. E. and Holland, J. H. (1988) Genetic algorithms and machine learning. *Machine Learning* 3(2): 95–99.
- Gonen, H. and Goldberg, Y. (2019) Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. arXiv preprint arXiv:1903.03862.
- Goodfellow, I., Bengio, Y. and Courville, A. (2016) *Deep Learning*. Cambridge, MA, London: MIT Press.
- Gregor, K., Danihelka, I., Graves, A., Jimenez Rezende, D. and Wierstra, D. (2015) DRAW: A Recurrent Neural Network For Image Generation. arXiv:1502.04623 [cs]. Available at: <http://arxiv.org/abs/1502.04623> (accessed 19 December 2018).
- Grimmer, J. and Stewart, B. M. (2013) Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis* 21(3): 267–297.
- Helber, P., Gram-Hansen, B., Varatharajan, I., Azam, F., Coca-Castro, A., Kopackova, V. and Bilinski, P. (2018) Mapping Informal Settlements in Developing Countries with Multi-resolution, Multi-spectral Data. arXiv:1812.00812 [cs, stat]. Available at: <http://arxiv.org/abs/1812.00812> (accessed 18 December 2018).
- Herzog, A. and Benoit, K. (2015) The most unkindest cuts: Speaker selection and expressed government dissent during economic crisis. *The Journal of Politics* 77(4): 1157–1175.
- Hochreiter, S. and Schmidhuber, J. (1997) Long short-term memory. *Neural Computation* 9(8): 1735–80.
- Holmes, D. E. and Jain, L. C. (eds) (2006) *Innovations in Machine Learning: Theory and Applications*. Studies in Fuzziness and Soft Computing 194. Berlin: Springer.

- House of Lords Select Committee (2018) AI in the UK: ready, willing and able? *House of Lords* 36. Available at: <https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf> (accessed 14 October 2019).
- Howard, J. and Ruder, S. (2018) Universal Language Model Fine-tuning for Text Classification. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers), Melbourne, Australia: ACL, pp. 328–339.
- Iyyer, M., Enns, P., Boyd-Graber, J. L. and Resnik, P. (2014) Political Ideology Detection Using Recursive Neural Networks. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore, United States: ACL, pp. 1113–1122.
- Jacovi, A., Shalom, O. S. and Goldberg, Y. (2018) Understanding Convolutional Neural Networks for Text Classification. arXiv:1809.08037 [cs]. Available at: <http://arxiv.org/abs/1809.08037> (accessed 19 December 2018).
- Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B. and Ermon, S. (2016) Combining satellite imagery and machine learning to predict poverty. *Science* 353(6301): 790–794.
- Johnson, R. and Zhang, T. (2014) Effective Use of Word Order for Text Categorization with Convolutional Neural Networks. arXiv:1412.1058 [cs, stat]. Available at: <http://arxiv.org/abs/1412.1058> (accessed 25 October 2018).
- Jordan, M. I. (2019) Artificial intelligence – The revolution hasn't happened yet. *Harvard Data Science Review* (1). Available at <https://doi.org/10.1162/99608f92.f06c6e61> (accessed 14 October 2019).
- Kalchbrenner, N., Grefenstette, E. and Blunsom, P. (2014) A Convolutional Neural Network for Modelling Sentences. arXiv:1404.2188 [cs]. Available at: <http://arxiv.org/abs/1404.2188> (accessed 25 October 2018).
- Kim, Y. (2014) Convolutional Neural Networks for Sentence Classification. arXiv:1408.5882 [cs]. Available at: <http://arxiv.org/abs/1408.5882> (accessed 25 October 2018).
- Knight, W. (2017, April) The dark secret at the heart of AI: No one really knows how the most advanced algorithms do what they do that could be a problem. MIT Technology Review 120(2). Available at: <https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai/> (accessed 14 October 2019).
- Krizhevsky, A., Sutskever, I. and Hinton, G. E. (2012) ImageNet Classification with Deep Convolutional Neural Networks. In: Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K.Q. (eds). Advances in Neural Information Processing Systems 25: Neural Information Processing Systems 2012, Red Hook, NY: Curran Associates, Inc, pp. 1097–1105.
- Kudugunta, S. and Ferrara, E. (2018) Deep neural networks for bot detection. *Information Sciences* 467: 312–322.
- Kusner, M. J., Loftus, J. R., Russell, C. and Silva, R. (2017) Counterfactual Fairness. arXiv:1703.06856 [cs, stat]. Available at: <http://arxiv.org/abs/1703.06856> (accessed 18 December 2018).
- Laclau, E. and Mouffe, C. (1985) Hegemony and Socialist Strategy: Towards a Radical Democratic Politics, 1st ed.: Radical Thinkers. London; New York: Verso.
- Lafferty, J. D., McCallum, A. and Pereira, F. C. N. (2001) Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In: Proceedings of the Eighteenth International Conference on Machine Learning, San Francisco, United States: Morgan Kaufmann Publishers Inc., pp. 282–289.
- Lauderdale, B. E. and Herzog, A. (2016) Measuring political positions from legislative speech. *Political Analysis* 24(3): 374–394.
- Laver, M. (2014) Measuring policy positions in political space. *Annual Review of Political Science* 17(1): 207–223.
- Laver, M., Benoit, K. and Garry, J. (2003) Extracting policy positions from political texts using words as data. *American Political Science Review* 97(2): 311–331.
- Lehmann, P., Werner, K., Lewandowski, J., Matthieß, T., Merz, N., Regel, S. and Werner, A. (2018) Manifesto Corpus. Version: 2018-01. Berlin: WZB Berlin Social Science Center. Available at: <https://manifesto-project.wzb.eu> (accessed 14 October 2019)
- Liu, N. F., Gardner, M., Belinkov, Y., Peters, M. and Smith, N. A. (2019) Linguistic Knowledge and Transferability of Contextual Representations. In: NAACL 2019, Minneapolis, United States. arXiv preprint arXiv:1903.08855.

- Available at <https://arxiv.org/abs/1903.08855> (accessed 14 October 2019).
- Makino, K., Takei, Y., Miyazaki, T. and Goto, J. (2018) Classification of Tweets about Reported Events using Neural Networks. In: Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text, Brussels, Belgium: Association for Computational Linguistics, pp. 153–163. Available at: <http://aclweb.org/anthology/W18-6121> (accessed 14 October 2019).
- Manzini, T., Lim, Y. C., Tsvetkov, Y. and Black, A. W. (2019) Black is to criminal as Caucasian is to police: Detecting and removing multiclass bias in word embeddings. In: NAACL 2019, Minneapolis, United States, pp. 1–5. arXiv preprint arXiv:1904.04047. Available at <https://arxiv.org/abs/1904.04047> (accessed 14 October 2019).
- Mayernik, M. S. (2017) Open data: Accountability and transparency. *Big Data & Society* 4(2), pp. 1–5. Available at <https://journals.sagepub.com/doi/pdf/10.1177/2053951717718853> (accessed 14 October 2019).
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. and Dean, J. (2013) Distributed Representations of Words and Phrases and their Compositionality. arXiv:1310.4546 [cs, stat]. Available at: <http://arxiv.org/abs/1310.4546> (accessed 19 December 2018).
- Mitchell, T. M. (1997) *Machine Learning*. New York: McGraw-Hill.
- Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S. and Floridi, L. (2016) The ethics of algorithms: Mapping the debate. *Big Data & Society* 3(2), pp. 1–21. Available at <https://doi.org/10.1177/2053951716679679> (accessed 14 October 2019).
- Nallapati, R., Zhou, B., dos Santos, C. N., Gulcehre, C. and Xiang, B. (2016) Abstractive Text Summarization Using Sequence-to-Sequence RNNs and Beyond. arXiv:1602.06023 [cs]. Available at: <http://arxiv.org/abs/1602.06023> (accessed 25 October 2018).
- Narayan, S., Cohen, S. B. and Lapata, M. (2018) Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization. arXiv:1808.08745 [cs]. Available at: <http://arxiv.org/abs/1808.08745> (accessed 19 December 2018).
- Nguyen, V.-A., Boyd-Graber, J., Resnik, P. and Miler, K. (2015) Tea Party in the House: A Hierarchical Ideal Point Topic Model and its Application to Republican Legislators in the 112th Congress. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (vol. 1: Long Papers), Beijing, China: Association for Computational Linguistics, pp. 1438–1448.
- Nissim, M., van Noord, R. and van der Goot, R. (2019) Fair is Better than Sensational: Man is to Doctor as Woman is to Doctor. arXiv preprint arXiv:1905.09866. Available at <https://arxiv.org/abs/1905.09866> (accessed 14 October 2019).
- Olah, C. (2015) Understanding LSTM Networks. Available at: <http://colah.github.io/posts/2015-08-Understanding-LSTMs/> (accessed 21 December 2018).
- Olah, C., Mordvintsev, A. and Schubert, L. (2017) Feature Visualization. *Distill*. Available at <https://distill.pub/2017/feature-visualization> (accessed 14 October 2019).
- Olhede, S. C. and Wolfe, P. J. (2018) The Growing Ubiquity of Algorithms in Society: Implications, Impacts and Innovations. In: *Philosophical Transactions of the Royal Society, Series A: Mathematical, Physical, and Engineering Sciences* 376(2128), pp. 1–16. Available at <https://royalsocietypublishing.org/doi/pdf/10.1098/rsta.2017.0364> (accessed 14 October 2019).
- Pennington, J., Socher, R. and Manning, C. D. (2014) Glove: Global Vectors for Word Representation. In: EMNLP, Doha, Qatar.
- Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. and Zettlemoyer, L. (2018). Deep Contextualized Word Representations. In: NAACL, New Orleans, United States.
- Peterson, A. and Spirling, A. (2018) Classification accuracy as a substantive quantity of interest: Measuring polarization in Westminster systems. *Political Analysis* 26(1): 120–128.
- Prates, M., Avelar, P. and Lamb, L. C. (2018) On Quantifying and Understanding the Role of Ethics in AI Research: A Historical Account of Flagship Conferences and Journals. arXiv:1809.08328 [cs]: 188–173. DOI: 10.29007/74gj.

- Preotjuc-Pietro, D., Liu, Y., Hopkins, D. and Ungar, L. (2017) Beyond Binary Labels: Political Ideology Prediction of Twitter Users. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (vol. 1: Long Papers), Vancouver, Canada, pp. 729–740.
- Proksch, S.-O. and Slapin, J. B. (2009) How to avoid pitfalls in statistical analysis of political texts: The case of Germany. *German Politics* 18(3): 323–344.
- Proksch, S.-O., Lowe, W. and Soroka, S. (2015) Multilingual sentiment analysis: A new approach to measuring conflict in parliamentary speeches. *Legislative Studies Quarterly* 44(1): 97–131.
- Radford, A., Narasimhan, K., Salimans, T. and Sutskever, I. (2018) Improving language understanding by generative pre-training. OpenAI.
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. and Sutskever, I. (2019) Language models are unsupervised multitask learners. OpenAI Blog 1(8).
- Ramisa, A., Yan, F., Moreno-Noguer, F. and Mikolajczyk, K. (2018) BreakingNews: Article annotation by image and text processing. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40(5): 1072–1085.
- Rao, A. and Spasojevic, N. (2016) Actionable and Political Text Classification using Word Embeddings and LSTM. Available at: <https://arxiv.org/abs/1607.02501> (accessed 29 November 2018).
- Ruder, S., Peters, M., Swayamdipta, S. and Wolf T. (2019) Transfer Learning in Natural Language Processing. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, pp. 15–18.
- Rudinger, R., May, C. and Van Durme, B. (2017) Social Bias in Elicited Natural Language Inferences. In: Proceedings of the First ACL Workshop on Ethics in Natural Language Processing, pp. 74–79.
- Rudinger, R., Naradowsky, J., Leonard, B. and Van Durme, B. (2018) Gender Bias in Coreference Resolution. arXiv preprint arXiv:1804.09301.
- Russell, S., Dewey, D. and Tegmark, M. (2015) Research priorities for robust and beneficial artificial intelligence. *AI Magazine* 36(4): 105–114.
- Samothrakis, S. (2018) Viewpoint: Artificial Intelligence and Labour. arXiv:1803.06563 [cs]. Available at: <https://www.ijcai.org/proceedings/2018/0803.pdf> (accessed 12 December 2018).
- Samuel, A. L. (1959) Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development* 3(3): 210–229.
- Sanders, J., Lisi, G. and Schonhardt-Bailey, C. (2017) Themes and topics in parliamentary oversight hearings: a new direction in textual data analysis. *Statistics, Politics and Policy* 8(2): 153–194.
- Schlogl, L. and Sumner, A. (2018) The Rise of the Robot Reserve Army: Automation and the Future of Economic Development, Work, and Wages in Developing Countries. ID 3208816, SSRN Scholarly Paper, 2 July. Rochester, NY: Social Science Research Network. Available at: <https://papers.ssrn.com/abstract=3208816> (accessed 19 December 2018).
- Singh, V., Varshney, A., Akhtar, S. S. Vijay, D. and Shrivastava, M. (2018) Aggression Detection on Social Media Text Using Deep Neural Networks. In: Proceedings of the 2nd Workshop on Abusive Language Online (ALW2), Brussels, Belgium, pp. 43–50. Association for Computational Linguistics. Available at: <http://aclweb.org/anthology/W18-5106>.
- Slapin, J. B. and Proksch, S.-O. (2008) A scaling model for estimating time-series party positions from texts. *American Journal of Political Science* 52(3): 705–722.
- Spirling, A. and Rodriguez, P. L. (2019) Word Embeddings: What Works, What Doesn't, and How to Tell the Difference for Applied Research. NYU manuscript. <https://github.com/ArthurSpirling/EmbeddingsPaper>
- Tenney, I., Das, D. and Pavlick, E. (2019) BERT Rediscovers the Classical NLP Pipeline. ACL 2019.
- Uysal, A. K. (2016) An improved global feature selection scheme for text classification. *Expert Systems with Applications* 43: 82–92.
- Wang, Y. and Kosinski, M. (2018) Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *Journal of Personality and Social Psychology*, 114(2): 246.
- Yamshchikov, I. P. and Rezagholi, S. (2018) Elephants, Donkeys, and Colonel Blotto. In:

- Proceedings of the 3rd International Conference on Complexity, Future Information Systems and Risk, Madeira, Portugal, pp. 113–119.
- Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. and Le, Q. V. (2019) XLNet: Generalized Autoregressive Pretraining for Language Understanding. arXiv 1906.08237.
- Yu, B., Kaufmann, S. and Diermeier, D. (2008) Classifying party affiliation from political speech. *Journal of Information Technology & Politics* 5(1): 33–48.
- Zhang, X., Zhao, J. and LeCun, Y. (2015) Character-level Convolutional Networks for Text Classification. arXiv:1509.01626 [cs]. Available at: <http://arxiv.org/abs/1509.01626> (accessed 25 October 2018).
- Zhang, Y. and Wallace, B. (2015) A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification. Available at: <https://arxiv.org/abs/1510.03820> (accessed 25 October 2018).
- Zhang, Q. and Zhu, S. (2018) Visual interpretability for deep learning: A survey. *Frontiers of Information Technology and Electronic Engineering* 19(1): 27–39.
- Zhang, Q., Wang, W. and Zhu, S. C. (2018) Examining CNN Representations with Respect to Dataset Bias. In: 32nd AAAI Conference on Artificial Intelligence, New Orleans, United States.
- Zhao, J., Zhou, Y. Li, Z., Wang, W. and Chang, K.-W. 2018. Learning Gender-Neutral Word Embeddings. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, pp. 4847–4853, Association for Computational Linguistics.

# Machine Learning in Political Science: Supervised Learning Models

Santiago Olivella and Kelsey Shoub

## INTRODUCTION

The set of theoretical and computational approaches used under the rubric of ‘machine learning’ (ML) is so diverse that it is easy to think of the field as a kind of catch-all, interdisciplinary exercise at the intersection between statistics, computer science and other affiliated disciplines. Such a perspective, however, would obscure the fact that these seemingly disparate approaches share a common goal: to improve a computer’s performance on a given (typically predictive) task by identifying empirical relationships, patterns, and trends in data that rely on minimal distributional and functional form assumptions on the part of analysts (Hastie et al., 2009; Jordan and Mitchell, 2015). This goal, which is what *learning* is typically taken to mean in this context, happens to be shared by a number of disciplines – including, of course, Political Science and International Relations.

Indeed, our discipline’s approach to the field of ML has resulted not only in careful applications of tools devised in other domains, but also in the development of tools designed to address problems of specific disciplinary interest – such as the fast estimation of ideal points that easily scales to millions of subjects (Imai et al., 2016), the structural analysis of text-as-data with large corpora (Roberts et al., 2014), and the discovery of subgroup causal effects in contexts with relatively few unknown confounders in high-dimensional datasets (Ratkovic and Tingley, 2017).

The goal of this chapter is twofold: to present some of the most commonly used tools in the field of predictive ML, and to illustrate how these tools have been adopted (and often adapted) to answer questions of interest in Political Science. We focus on *supervised models*, whereas other contributions to the *Handbook* (e.g. Chatsiou and Mikhaylov, Chapter 55, this *Handbook* on Deep Learning for Political Science) offer insights into other approaches within ML and computational



social sciences. Supervised models are those where both an outcome of interest *and* relevant predictors are available for a set observations and can be used to learn the patterns of their association, so as to better predict instances in which the outcome is *not* observed.

We begin the chapter with a simple taxonomy of ML tools using the supervised/unsupervised distinction and offer a brief overview of some of the most important concepts in the field. We then discuss the main ideas behind three major supervised learning approaches, namely tree-based methods, kernel-based approaches, and support vector machines. Although we provide enough mathematical detail to understand the formal underpinnings of each approach, we emphasize the intuitions behind these models (rather than their theoretical foundations) to make the materials accessible to a wider audience. Finally, we conclude the chapter with some thoughts on the promise and perils of ML modeling and offer some suggestions for additional readings.

## TYPES OF MACHINE LEARNING MODELS AND RELEVANT CONCEPTS

As is the case in many other social sciences, quantitative data analysis in our disciplines has typically relied on some variant of the linear model. The modal empirical exercise defines a specific functional relationship between the conditional expectation of an outcome of interest and a limited set of predictors suggested by theory, tradition, or a combination of both. Although useful in its parsimony and intelligibility, this common approach has a number of limitations.

For instance, classical regression models require positing a particular specification for the conditional expectation function at the onset of analysis. Even if a few different specifications are considered, choosing among them after estimation results have been observed can at best offer proof that there is

*one* way in which the world is consistent with proposed hypotheses – not a very strong test of a hypothesis. In the absence of real competition from the myriad alternative plausible models, the typical approach offers few meaningful chances for the researcher to be proved wrong (Ho et al., 2007). Similarly, classical regression models are unable to handle situations in which the number of included predictors is far larger than the number of available observations, as is common in text-as-data applications discussed elsewhere in this volume. By increasing modeling flexibility while actively avoiding tracking sample idiosyncrasies too closely, ML models offer viable steps towards resolving these and many other issues raised by classical approaches.

Despite being unified by the same goals and general flexibility, there is a daunting variety of modeling approaches that fall under the purview of ML. Different taxonomies of models exist based on a variety of criteria. For instance, we can distinguish between *generative models* (which define a probability model for the joint distribution of outcomes and predictors) and *discriminative models* (which either define a probability model for the conditional distribution of outcomes given predictors or otherwise propose no distributional assumptions whatsoever) (Bishop, 2016; Ng and Jordan, 2002). Examples of the former include the *latent Dirichlet allocation* and the *hidden Markov model*, while examples of the latter include models like *support vector machines* and *regression trees* (as well as less exotic models, like the logistic regression).

Perhaps the most commonly used dimension along which ML models tend to be classified, however, pertains to the relevance of a target outcome in the learning task. If the goal is to learn structure in a set of inputs (or features) without appealing to a target outcome, then we are said to face an *unsupervised learning* (or descriptive) problem (Murphy and Bach, 2012). Examples of such tasks include problems such as the organization of unstructured text into topics, and the

estimation of latent traits (such as ideology) based on observed behaviors (such as roll-call votes or decisions to follow actors embedded in a social network).

On the other hand, when both features and outcomes are observed for some set of observations – typically called the *training set* – which are then used to infer the outcome values of unlabeled observations, we are said to have a *supervised learning* problem. Examples of such tasks include classical problems such as email spam detection, hand-written digit recognition, and object-detection in images. They also include the problems of learning a response surface defined over an input space or (equivalently) of understanding how an outcome is related to inputs in potentially complex ways (Hastie et al., 2009). Although many interesting ML developments occur in the realm of unsupervised learning (as evidenced, for instance, by the discussions in Egerod and Klemmensen, Chapter 27, and Bouchat, Chapter 28, of this *Handbook*), we focus on supervised learning techniques in this chapter, as they encompass the kinds of predictive tasks most commonly associated with ML.

Within supervised learning models, we can further distinguish between models based on the measurement type of the target outcome. If the outcome is continuous, learning tasks are referred to as *regression problems*. In turn, categorical outcomes (regardless of whether they have two or more categories) give rise to *classification problems* (Bishop, 2016; Murphy and Bach, 2012).<sup>1</sup> Regardless of whether they are regression or classification problems, however, supervised learning models share the same goal: learn the potentially complicated relationships that relate (combinations of) features  $x$  to the outcome of interest  $y$  *in general*, using information available in the set of observations for which the pair  $(x, y)$  is fully observed.

The *in general* qualification is an important one, as it is typically easy to learn even complicated relationships in-sample – that is, relationships that are conditional on the training set. The goal, however, is to learn relationships for which expected *generalization error*

(i.e. the error that can be expected to ensue when learned relationships are evaluated *out-of-sample*, on a random *test set* of observations not involved in the learning process) is low (Hastie et al., 2009). In fact, while it is always possible to reduce *training error* (i.e. error as computed using the training sample) by making models arbitrarily complex, such flexibility typically results in high expected generalization error, as models start to *overfit* their training data (i.e. they start to pick up on idiosyncratic relationships that are conditional on the set of observations used to train the models).<sup>2</sup>

Accordingly, and since the ultimate goal of supervised learning is to find generalizable patterns of association, models are typically subject to some form of *regularization* – typically in the form of a constraint that pushes the model towards parsimony – and are selected based on their ability to generate good out-of-samples predictions. Clearly, it is impossible to evaluate a model's performance on the universe of unsampled test instances, so an approximate measure of performance must be devised. Although several approaches are viable, none is more commonly used than *cross-validation (CV)* – the exercise of further splitting the training data into a training set and a validation set (used to evaluate predictive accuracy, but omitted from the learning phase). To further minimize issues related to bad draws, multiple such splits are typically conducted.

The most popular approach to cross-validation, *k-fold CV*, partitions the training data into  $k$  subsets, estimates the model leaving each subset aside, evaluates predictive accuracy on the held-out set of observations, and approximates the generalization error of the specific model by taking the average of these held-out errors (Hastie et al., 2009). When models require the definition of so-called *tuning parameters* (i.e. ancillary parameters that govern the model's behavior, such as the number of topics in a topic model), multiple values are evaluated using this iterative estimation process, allowing researchers to base their value definitions on a data-driven procedure.

Overall, the typical workflow in ML involves: first (and crucially) representing raw data using quantitative features (e.g. transforming unstructured texts into a document feature matrix); second, choosing an appropriate model for the learning task at hand, tuning the model’s parameters using some form of cross-validation; and third, evaluating the model’s out-of-sample predictive accuracy (i.e. its expected generalization error) using an entirely separate test set.

To give a flavor of what some of these decisions look like, we now turn to a discussion of two of the most commonly used sets of models in supervised learning within Political Science: tree-based approaches and support vector machines.

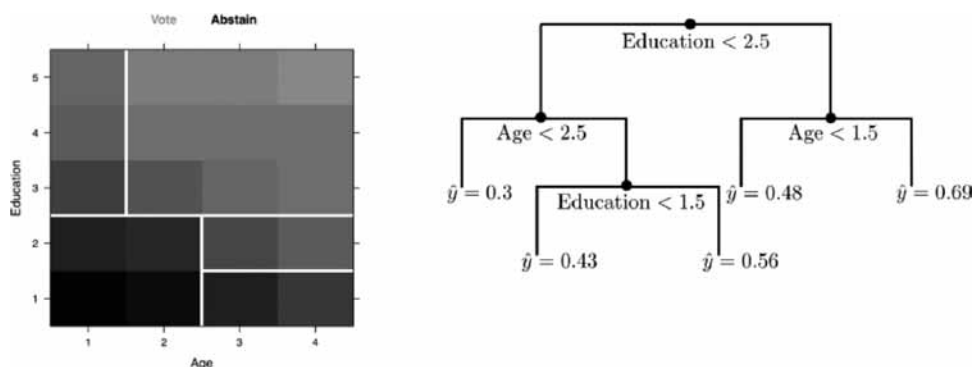
### EXAMPLES OF SUPERVISED LEARNING MODELS

#### *Tree-Based Approaches: CART, Random Forests, and Tree Boosting*

For a single-tree model, the goal is to partition the space of predictor features (i.e. the set of all unique combinations of predictor

values) into contiguous regions within which prediction is easier, thus improving overall predictive accuracy by sorting observations into their respective bins. Intuitively, this goal is best achieved if the regions are defined by their degree of homogeneity with respect to the outcome of interest, so that region-specific predicted values are as close to the target as possible.

To clarify these abstract ideas, consider the task of predicting turnout based on age and education levels. The left panel of Figure 56.1 shows our feature space (defined by unique values of the two predictor features) as well as a hypothetical target turnout distribution, with darker shades indicating higher turnout rates. In turn, the center panel shows a stratification of all observations into regions for which a simple, constant prediction model (e.g. using the average turnout rate for all observations within the region) offers a good approximation of the underlying turnout distribution. When we restrict regions to form non-overlapping ‘boxes’ that are aligned with the coordinate axes of the feature space (i.e. such that their edges are parallel to said axes), the partition can be represented by a recursive binary tree – hence giving these models their name. The right panel of Figure 56.1 shows



**Figure 56.1** Left panel: feature space comprised of ages and education levels, partitioned into regions that are relatively homogeneous with respect to probability of turnout (indicated with lighter shades of gray for every combination of feature values). Right panel: binary-tree representation of feature-space partition depicted on the left panel. Values at terminal nodes indicate predicted turnout probability in each corresponding region

the recursive binary tree that corresponds to the partition depicted in the left panel and is the result of fitting a *classification and regression tree* (or CART, see Breiman et al., 1984) model to the hypothetical data.

More formally, we can define the predictive task as one in which we need to learn parameters  $\Theta$  in a function

$$f(\mathbf{x}_i | \Theta) = T(\mathbf{x}_i | \Theta) = \sum_{b=1}^B \hat{y}_b \mathbf{1}(x_i \in R_b)$$

where  $R_b$  denotes the  $b_{th}$  region (out of  $B$  total regions) in the partition of the feature space,  $\hat{y}_b$  is the constant prediction for all observations in that region (typically the average or modal category for observations in the region), and  $\mathbf{1}(\cdot)$  is the indicator function. The  $\Theta$  parameters control both the variables for creating each split in the recursive tree, as well as the value at which each split occurs.

Finding the best  $\Theta$  for any given loss function  $L(y_i, f(\mathbf{x}_i))$  is prohibitively costly, as it requires solving an extremely hard discrete optimization problem. Instead, a simple greedy heuristic for finding a local optimum consists of sequentially picking a predictor-value pair  $p, v$  among the set of all predictors  $P$  that minimizes prediction loss among the induced regions. This approach, known as *recursive binary splitting*, forms the tree through a sequence of locally optimal recursive binary splits, starting with the entire feature space at its root<sup>3</sup> and ending when a given stopping criterion is reached (for instance, when fewer than a pre-specified number of observations would fall into a new region).

Although the final model has many advantages (e.g. it is easy to interpret and visualize, it can capture complex interactions between predictors when grown to be deep enough, it performs a kind of automatic feature selection by choosing locally predictive variables, and it can be easily adjusted to accommodate values that are missing<sup>4</sup> or that are measured in different scales), it is also affected by serious limitations. First, growing a single, deep tree using binary recursive splitting can result in a grossly overfit model. In turn, and as an

example of the common bias-variance trade-off, this high level of in-sample predictive accuracy usually comes at the expense of high variability in prediction, as single trees grown recursively can often times yield wildly different predictions as a result of small changes in the training set. Other limitations of common implementations of single-tree models include their practical inability to accommodate additive relationships, the lack of natural measures of prediction uncertainty, and the fact that variable choice in recursive binary splitting is biased towards choosing variables with many potential splitting points.

While there are strategies designed to ameliorate some of these issues,<sup>5</sup> the most common approach consists of building ensembles of trees. Formally, a tree-ensemble with  $M$  trees – grown in a slightly different way to induce variety and learn relationships using a ‘wisdom-of-crowds’ approach – can be defined by

$$f(\mathbf{x}_i | \Theta) = \frac{1}{M} \sum_m^M T(\mathbf{x}_i | \Theta_m)$$

with the goal of learning optimal parameters

$$\tilde{\Theta} = \arg \min_{\Theta} \sum_i L(y_i, f(\mathbf{x}_i | \Theta)) \tag{1}$$

with strategies used to induce differences across trees resulting in a variety of flavors of tree-ensemble methods. Although there is a cost to be paid in terms of interpretability – for example, making a prediction is no longer as easy as ‘dropping’ an observation down the binary tree, and following it until we hit a terminal region with a corresponding predicted value – the gains in predictive accuracy and reduction of generalization error offer substantial advantages. We now present two of the most commonly used ensemble models (viz. random forests and gradient boosting machines).

### *Bagging and Random Forests*

The first approach to building an ensemble of trees consists of bootstrapping samples from the training set, fitting a single tree to each

resampled set, and then aggregating predictions made by each such tree by, for instance, taking their average. Adequately named *bagging* (a portmanteau for *bootstrap aggregating*), this approach can help accommodate non-interactive relationships (by building trees with branches that are functions of a single predictor, for instance) as well as non-additive relationships. However, the approach remains unsuccessful in terms of reducing estimator variance (and thus improving overall generalization error) unless an additional step is taken to de-correlate the trees that form the ensemble.

A simple way to reduce variance by way of reducing correlation among trees is to restrict the choice of each splitting variable to a random subset of predictors  $p \subset P$ , so that each bagged tree provides a different ‘perspective’ on the prediction problem. In fact, it can be shown that correlation  $\rho(x)$  declines as the relative size of  $p$  (i.e. the number of predictors used at any given splitting point) decreases (Hastie et al., 2009). This model, known as a *random forest* (RF) (Breiman, 2001), is one of the most popular ML approaches used in Political Science, with simple to use implementations available through open-source software (e.g. `randomForest` in **R**) that require cross-validation of only a handful of parameters (including the ensemble size  $M$  and the size of the random subset of predictors at each splitting point,  $|p|$ ). And because trees can be grown in parallel, the approach has the potential to scale relatively well in the presence of large ensembles and datasets.

In addition to improving predictive accuracy by reducing variance and maintaining each tree’s low bias properties, RFs increase the set of functional associations the model can represent and performs the same kind of on-the-fly feature selection performed by single-tree approaches. The model is also capable of providing estimates of generalization error by generating *out-of-bag* (i.e. out-of-sample) predictions for observations *not* in the bootstrap sample for each tree. Finally,

measures of uncertainty can be readily produced using bias-corrected versions of the infinitesimal jackknife (Efron and Hastie, 2016; Wager et al., 2014) at no additional computational costs.

### Gradient Tree-Boosting

An alternative approach to building ensembles of trees is known as *boosting* (Freund and Schapire, 1997; Schapire, 1990). Boosting is an example of *forward stagewise additive modeling* – a procedure that tackles a complex optimization problem involving a sum of basis functions (such as Equation 1 *sequentially*, in  $M$  steps. At each step, the parameters involved in each basis function (in this case, trees) fit up until that point are left unchanged, thus reducing the complexity of the optimization problem by solving for the optimal parameters of a single basis function at each stage. As such, the program in Equation 1 becomes

$$\tilde{\Theta}_m = \operatorname{argmin}_{\Theta_m} \sum_i L(y_i, f_{m-1}(\mathbf{x}_i) + \nu T_m(\mathbf{x}_i | \Theta_m)),$$

where  $0 < \nu \leq 1$  is a step size that controls the learning rate, and  $f_{m-1}(\mathbf{x}_i)$  is the prediction of the tree ensemble fit in the first  $m - 1$  stages. What distinguishes boosting from other stagewise approaches is its focus on repeatedly modified versions of the training set, effectively transformed to gradually shift focus, at each stage, to observations that have been poorly fit by the ensemble up until that point.

A general formulation of this approach is given by *gradient boosting machines* (GBMs), which fits a tree  $T_m(\mathbf{x}_i | \Theta_m)$  to the negative gradient of the loss function with respect to  $f(\mathbf{x}_i)$ , evaluated at  $f_{m-1}(\mathbf{x}_i)$  at each stage of the sequential ensemble construction. In doing so, the model progressively concentrates on some measure of difference between the predictions of the  $(m - 1)$ th step and the observed values – which is precisely the information contained in the negative

gradient. For example, with (half) squared error loss, the  $i_{th}$  component of the negative gradient is given by

$$-\frac{\partial}{\partial f(\mathbf{x}_i)} \frac{1}{2} [y_i - f(\mathbf{x}_i)]^2 \Bigg|_{f_{m-1}(\mathbf{x}_i)} = y_i - f_{m-1}(\mathbf{x}_i)$$

which is the model's residual up to the  $(m - 1)_{th}$  step. Other differentiable loss functions make it possible to use gradient tree-boosting for a variety of classification and regression problems. The AdaBoost.M1 algorithm (Freund and Schapire, 1997), for instance, is a commonly used implementation of gradient boosting for binary classification using an exponential loss function and predictions in the set  $\{-1, 1\}$ .

Although not parallelizable like RFs (as they must be learned sequentially), GBMs are typically very fast to estimate. Tuning them requires defining values for the number of trees in the ensemble  $M$ , for the step size  $v$ , and for the tree depth of each ensemble member. In practice, trees are usually grown on training sets to be 'weak' learners (i.e. shallow trees, with depth defined primarily by the theoretical order of anticipated interactions among predictors),  $v$  is set to some small number (e.g.  $v = 0.001$ ), and  $M$  is chosen by  $K$ -fold cross-validation from a fine sequence (e.g.  $M \in \{2, 3, \dots, 5000\}$ ). Overall, gradient tree-boosted models have been found to have excellent predictive performance, prompting some to call them 'the best off-the-shelf classifiers in the world' at one point in time (Breiman, 1998; Hastie et al., 2009). Good implementations in **R** include the `gbm` package (which can be cross-validated using `caret`) and `h2o`. Performance can usually be enhanced even further by incorporating some of the ideas behind RFs, such as using data subsampling at each stage of the ensemble creation, resulting in a variant of GBM called *stochastic gradient boosting* (Friedman, 2002), also implemented in `gbm` in **R**.

### *An Application to Small-Group Preference Estimation*

In addition to their use in forecasting tasks (Muchlinski et al., 2016; Kaufman et al., 2018), tree-based models have commonly been used in the social and political sciences to study interactive effects and other types of conditional associations. In the study of causal relationships, for instance, tree-ensembles have been used to identify heterogeneous treatment effects (Green and Kern, 2012; Imai and Strauss, 2011; Wager and Athey, 2017). Their ability to identify complex functional forms without the need for researcher-defined specifications makes tree-based models ideal for another task: estimating preferences among small target populations using post-stratification of estimates obtained from non-representative samples.

In a survey of potential applications of tree-based models within Political Science, Montgomery and Olivella (2018) are able to reproduce and efficiently scale-up the exercise conducted by Ghitza and Gelman (2013). In their study, Ghitza and Gelman aim to estimate vote intentions of small groups of voters during the 2008 presidential election, as well as their likelihood of turning out to vote. The groups, defined by intersections of geographic and socio-demographic characteristics (e.g. high-school educated Latino women between 18 and 25 who live in North Carolina), were assumed to have preferences that depended on the *combination* of these characteristics, thus requiring models that allowed for 'deep interactions' on their right-hand sides.

The two-step approach they propose – which involves a predictive stage and a post-stratification stage, known as multilevel regression and post-stratification, or MRP – uses a random-intercepts model to model preferences as a function of these interactions and post-stratifies predictions based on this model using highly granular frequency counts (e.g. census tables at the block level)

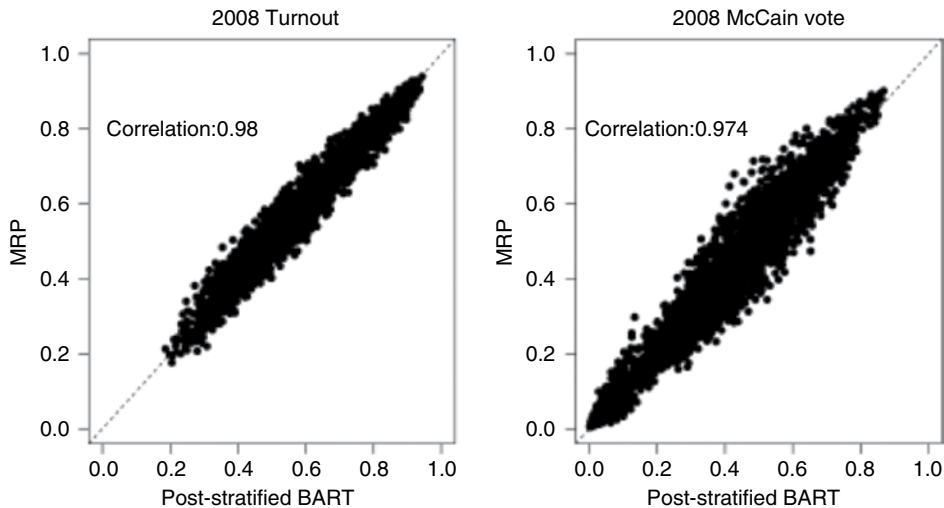
that are then aggregated to whatever level is desired. Although interested in combinations of a large number of socio-demographic characteristics, Ghitza and Gelman's exercise is restricted by the computational limits imposed by the estimation of a large number of random effects involved in fully interactive specification of the first stage model.

To address this, Montgomery and Olivella replace the first stage model with a tree-based ensemble (viz. Bayesian Additive Regression Trees, or BART), which effectively only incorporates interactive effects when the training data support them, thus reducing the computational cost of the estimation without sacrificing potentially relevant model complexity. The results of both exercises are virtually identical when estimated using the same set of four predictors (viz. state, ethnicity, income, and age), as evidenced by Figure 56.2 (which presents the set of post-stratified MRP estimates vs. the post-stratified tree-based estimates).

Contrary to the multilevel model, however, the tree-based approach can easily incorporate the full set of available predictors (a set of eight discrete variables, with a total set of 163,200 fully interacted categories) with very little additional computational overhead.

### **Kernel Methods: Support Vector Machines**

Yet another approach to the supervised learning problem is offered by support vector machines (SVMs), which have been most successfully used to solve classification problems – particularly when the number of predictive features is much larger than the number of observations  $n$ , or  $p \gg n$ . Examples of classes could be party affiliation, vote choice, sentiment, or whether or not a respondent received the experimental treatment. Like many tree-based models,



**Figure 56.2** Left panel: post-stratified predictions of turnout during the 2008 presidential election in the United States at low levels of demographic aggregation. Predictions produced by a multilevel model along the y-axis, and predictions along the x-axis produced by a Bayesian additive regression tree model. Right panel: post-stratified predictions of vote intention for McCain during the 2008 presidential election in the United States at low levels of demographic aggregation. Predictions produced by a multilevel model along the y-axis and predictions along the x-axis produced by a Bayesian additive regression tree model

SVMs are ‘one of the “must-have” tools in any machine learning toolbox’ (Efron and Hastie, 2016: 387).

Intuitively, an SVM finds a hyperplane that maximizes the distance between it and the nearest instances of each class (the support vectors), thus producing the ‘cleanest’ possible sorting of observations. This distance to the nearest instances, called the margin, generates a kind of buffer between types of observations; the optimization objective behind SVMs is to maximize the width of this buffer. To make these ideas more concrete, take an example where a researcher wants to predict whether instances are members of class A or class B. To do so, the researcher has two variables:  $X_1$  and  $X_2$ . When fitting an SVM to predict which class each observation belongs to, the SVM finds the maximum-margin (hyper-)plane that separates these two classes while maximizing the distance between the (hyper-)plane and the nearest class instances.

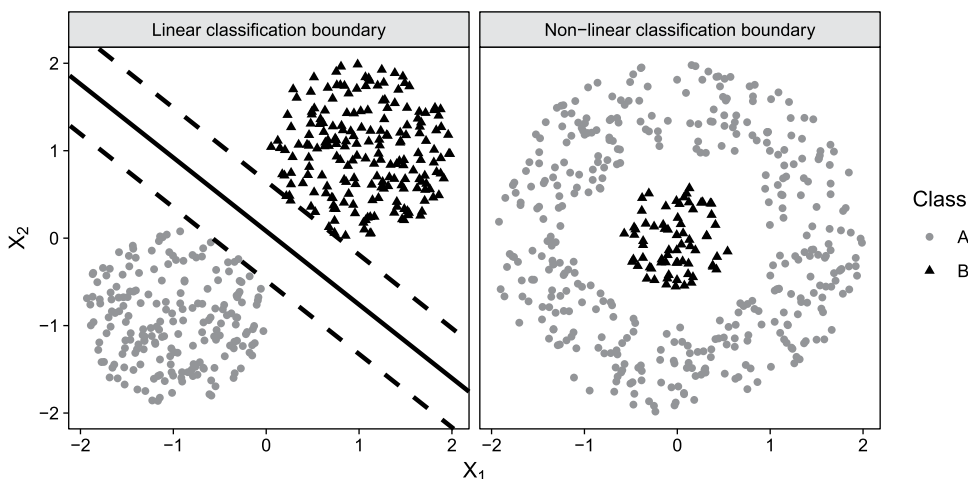
Graphically, this is depicted on the left panel of Figure 56.3. In the figure, instances belonging to class A are shown as light-gray circles, and instances belonging to class B are shown as dark-gray triangles. The

maximum-margin hyperplane is the solid line separating the two groups, and the margin (which touches the support vectors) is depicted using dashed lines equidistant from the hyperplane. As can be seen, all of the circles lie below the hyperplane (a line, in this case), while all of the triangles lie above the hyperplane: classes are perfectly linearly separable. Fitting an SVM (or, in this simple case, a *maximum margin classifier*) amounts to finding this separating line. Note that an important implication of this goal is that only those instances near the class boundary play a big role its definition, while those that remain far away from the boundary have little effect on its location and direction.

More formally, assume that a given data set consists of  $N$  instances represented by a set of features. In Figure 56.3, there are 40 instances, and the features used are the variables  $X_1$  and  $X_2$ , where  $x_{1i} \in \mathbb{R}$  and  $x_{2i} \in \mathbb{R}$ . Additionally, the target classes can be either  $-1$ , or  $y_i \in \{-1, 1\}$ . A hyperplane is defined by:

$$x : f(x) = x^T \beta + \beta_0 = 0 \tag{2}$$

where  $\beta$  is a unit-length vector. Such a plane is separating if all instances of a class lie



**Figure 56.3** Left panel: classifying by a Maximum-Margin Hyperplane: a linearly separable problem, with the corresponding separating line (solid line) and classification margins (dashed lines). Right panel: non-linearly separable classification problem of Class A (outer light gray triangles) and Class B (inner dark gray circles) instances



above it, and all instances of the other class lie below it (or, equivalently, if  $y_i \times f(x_i)$  for all training instances). Thus, a classification rule based on the separating plane would be  $G(x) = \text{sign}[f(x)]$ . Finally, note that (given  $\|\beta\| = 1$ ) the distance between any observation  $i$  and the separating plane is given by  $y_i(x_i^T \beta + \beta_0)$ . Thus, and for a margin  $M$  and set of support vectors  $\mathcal{S}$ , finding the maximum margin classifier amounts to finding

$$\underset{\beta, \beta_0}{\text{argmax}} M \text{ subject to } y_i(x_i^T \beta + \beta_0) \geq M \quad \forall i$$

To do so, the relevant Lagrangian dual objective function is maximized, and the solution for  $\beta$  is found to be:

$$\hat{\beta} = \sum_{i=1}^N \alpha_i y_i x_i = \sum_{i \in \mathcal{S}} \alpha_i y_i x_i, \quad (3)$$

where  $\alpha_i$  are the Lagrange multipliers, and the second summation highlights the fact that only those observations in the support set affect the separating hyperplane.

### The Soft-Margin Classifier

So far, we have assumed that a hyperplane can perfectly separate instances across classes. When this is not the case, we must relax the constraint imposed on the distances between points and the hyperplane, and allow for a certain amount of slack. This slack will allow for instances to be within the margin, or even to cross the (quasi-)separating hyperplane.

Thus, although the objective of the optimization stays the same (i.e. maximize the margin), the *soft-margin* constraint is given by  $y_i(x_i^T \beta + \beta_0) \geq M(1 - \epsilon_i)$ , where the  $\epsilon_i$  are the slack terms, themselves constrained so that  $\sum_i \epsilon_i \leq C$ . The dual objective is now maximized subject to  $0 \leq \alpha_i \leq C$  and  $\sum_{i=1}^N \alpha_i y_i = 0$ . As a result,  $C$  becomes a tuning parameter.

Specifically, the margin around the identified hyperplane(s) is larger for smaller values of  $C$  (e.g.  $C = 0.01$ ) and is smaller for larger values of  $C$  (e.g.  $C = 1000$ ). Larger values of  $C$  thus result in greater focus of attention on the points located very close to the decision boundary, while smaller values involve

data points farther away. It is these points (which effectively have Lagrange multipliers greater than zero) that now become the support vectors.<sup>6</sup>

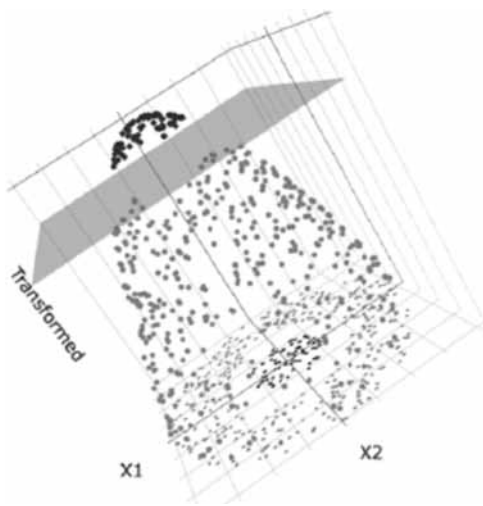
Typically, the value for  $C$  is defined through cross-validation. When identifying *linear* boundaries, however, the results tend not to be too sensitive to the specification of  $C$  (Hastie et al., 2009). The choice matters considerably more when linear boundaries no longer generate optimal classifiers, as smaller values of  $C$  can result in severe over-fitting. The generalization of these ideas to non-linear decision boundaries is what is typically called a *support vector machine*.

### Support Vector Machines and Kernels

It is often the case that a non-linear classification boundary is needed in order to correctly classify instances. For instance, consider the right panel of Figure 56.3. Although the two classes are easily recognized as occupying different regions of feature space, no hyperplane across it would result in a good separation. The optimal decision boundary, which in this case corresponds to a circle, is not linear.

In the classical regression context, dealing with non-linear associations involved incorporating non-linear functions of predictors (e.g. higher order polynomials, log-transformations) into the specification of the conditional expectation of the outcome. SVMs adopt a similar strategy: instead of operating on the space defined by the original set of predictors (where no linear boundary can correctly separate classes of the target outcome), they operate on a transformed space of higher dimensions *in which linear separability becomes possible*.

Consider, once again, the classification problem illustrated on the right panel of Figure 56.3 (now reproduced as a tilted projection at the bottom of Figure 56.4). Suppose we add a third feature equal to the negative sum of squares of the original predictors. This results in the 3D scatterplot depicted



**Figure 56.4** Illustration of kernel trick: tilted projection of instances depicted in right panel of Figure 56.3 onto space with an additional dimension  $z = -(X_1^2 + X_2^2)$  (original two-dimensional representation can be seen at the bottom). Observations of Class B (in dark gray) now cluster at the top of a conical surface. SVM learns the gray plane that cuts across this conical surface, resulting in perfect separation of Class A instances (light gray, below plane) and Class B instances (dark gray, above plane)

on Figure 56.4, slightly tilted to improve visibility. In this new, three-dimensional feature space, observations are now arrayed on a conical surface, with instances of class B rising to its apex. It is now easy to define a plane, depicted in gray in Figure 56.4, that cuts the top of this cone and separates instances of the two classes. The projection of this separating plane back onto the original two-dimensional

space generates the circular decision boundary we needed. Once again, the SVM’s goal is to learn this separating plane.

What is remarkable about actual implementations of SVMs is that there is no need to explicitly define what these additional dimensions are, provided they can be expressed as functions of the pair-wise dot products  $\langle x_i, x_j \rangle$  of the original features. Specifically, SVMs rely on kernels  $K(x_i, x_j)$  – bivariate, symmetric, and positive-definite functions that define measures of proximity or similarity between observations, and which operate on the space of original features. In practice, different kernels support different kinds of implicit added features. This approach, called the *kernel trick*, avoids explicit re-mapping onto higher-dimensional spaces. Like boosting, kernelization can be applied to many different types of learners (including GLMs; see, for instance, Hainmueller and Hazlett, 2014).

In the SVM literature, popular kernels include the polynomial kernel, the radial basis function (RBF) kernel, and the sigmoid kernel. Table 56.1 provides a summary of the kernels. All of these (and many others) are implemented in the `svm` function included in `R` package `e1071`. The use of different kernels may result in different predictions and different weights being placed on the features. Additionally, values taken by different (hyper)parameters (such  $d$  in the Polynomial kernel,  $c$  in the RBF, or  $\kappa_1$  and  $\kappa_2$  in the Sigmoid kernel) must be defined by the researcher. Once again, in practice, cross-validation is a good strategy for choosing these values.

One interesting aspect of both the RBF kernel and the sigmoid kernel is that each is

**Table 56.1** Popular SVM kernels

Kernel	Form
Linear	$K(x, x') = \langle x, x' \rangle$
$d$ th degree polynomial	$K(x, x') = (1 + \langle x, x' \rangle)^d$
Radial basis function	$K(x, x') = \exp(-\ x - x'\ ^2/c)$
Sigmoid (neural network)	$K(x, x') = \tanh(\kappa_1 \langle x, x' \rangle + \kappa_2)$

essentially a type of neural network. In fact, earlier editions of Hastie et al. (2009) refer to the sigmoid kernel as the *neural network kernel*. The first is a type of neural network actually called an RBF network. The second is a standard neural network fit with a multi-layer perceptron and one hidden layer.

Incorporating the kernel-transformed features into the SVM optimization problem is not difficult. In general, Equation 2 can be reformulated using kernels. Specifically, the function used to define the hyperplane can be rewritten as:

$$f(x) = \sum_{i \in S} \alpha_i y_i K(x, x_i) + \beta_0. \quad (4)$$

with everything else (including soft constraints and prediction rules) staying as before.

Although we have discussed SVMs in the context of binary classification, there are extensions to multinomial classification problems and even regression-like tasks, where they are typically called *support vector regressions* (Cortes and Vapnik, 1995; Grundler and Krieger, 2015; Hastie, Tibshirani and Friedman, 2009; Witten et al., 2016).

It is also worth noting that while SVM's are typically very fast and accurate (and scale particularly well in terms of the dimensionality of the original feature space), they have been shown to be very closely related to regularized logistic regression models (the so called *Loss + Penalty* representation of SVMs) (Hastie et al., 2009), and they tend to provide similar results – particularly when classes are well separated.

### *An Application in the Study of Criticism in Congress*

One common application of support vector machines is to learn characteristics of interest, effectively using the SVM as a measurement model. Examples of this are scattered throughout computer science and are becoming more common within political science. A good example is the recent study by the Pew Research Center's Data Labs team on partisan conflict and Congressional outreach

(Messing, VanKessel, Hughes, Blum, and Judd 2017).

In light of heightened political hostility within and outside the walls of Congress, they questioned who actively criticizes the other party, who 'goes negative', and how the public responds to both of these strategies. To answer these questions, they looked at 108,235 Facebook posts and 94,521 press releases made by members of Congress between January 1, 2015 and April 30, 2016 (during the 114th Congress). Within the posts and releases, they needed to identify three characteristics of the text: whether the member of Congress in that document criticized someone or some group, whether they expressed disagreement, and whether they discussed bipartisanship.

To code every post by hand for each of these characteristics would have been prohibitively time intensive. As an alternative approach, the authors resorted to a *semi-supervised* approach. Specifically, they trained SVMs to identify the concepts of interest on a small hand-coded subset of the data, and then used the trained model to classify the unlabeled documents. First, they randomly sampled a subset of both the press releases and Facebook posts to code by hand.<sup>7</sup> The team at Pew then used Mechanical Turk to garner human-coded labels. Using these codes, they then re-trained an SVM for each concept of interest. To ensure that the final SVM used to classify the remaining data performed as well as possible, they compared penalty levels<sup>8</sup> and different kernels<sup>9</sup> with five-fold cross validation to identify which combination resulted in the best performing algorithm out-of-sample. Finally, using the best performing model for each characteristic, they then applied labels to the remaining documents.<sup>10</sup>

The authors find that the party leadership and ideological extremists on both sides of the aisle are more likely to be critical of the other party and vocally disagree. Conversely, moderates in competitive districts are more likely to talk about bipartisanship. Additionally, they find that the public tends

to like, share, and comment on critical posts more than other posts on Facebook.

## CONCLUSION: PROMISE AND PERILS OF MACHINE LEARNING

ML approaches and methods provide a wide variety of new and exciting tools to political scientists that are especially useful in the face of high-dimensional data sets. As political scientists seek answers to increasingly complex questions and using increasingly complex data (e.g. text or images), such techniques and methods will be increasingly relevant. However, such a promise does not come without challenges and pitfalls. Political scientists face three general challenges when seeking to adopt and use ML techniques and other methods developed by computer scientists.

The first is derived from the fact that, for a long time, those developing ML approaches sought only to *predict* outcomes rather than *understand* the phenomena of interest. This is problematic for at least two reasons. First, many ML algorithms, methods, and models rely on black box techniques. As a result, interpreting the learned relationship between inputs and outcomes becomes much more difficult than with classic techniques used within Political Science. For example, while SVMs are fantastic tools for accurate classification, they offer few ways of determining which features are most predictive. This implies that their use for theory evaluation is extremely limited. Moreover, while *partial dependence plots* (i.e. plots of marginal predicted outcomes as a function of features of interest) (Friedman, 2001) and *variable importance* measures (e.g. weighted averages of error reduction induced by variable selection; Breiman, 1984) can provide a sense of how the target changes as a function of given predictors and which features are most relevant, they remain underutilized and misunderstood.

Second, and partly as a result of the complex problems they tackle, ML approaches tend not to be robust to technical decisions made by researchers. While ML models take away researcher degrees of freedom and can help prevent common issues related to *p*-hacking or specification-related forking-paths (Gelman and Loken, 2013), these models come with their own set of potentially consequential decisions: how is data pre-processed to extract relevant features; how are initial values of parameters chosen; and what set of criteria are used to evaluate goodness-of-fit? While cross-validation techniques can help justify and evaluate some of these choices, the added computational costs of these safe-guards can make this process prohibitively time-intensive, which results in many researchers choosing to use defaults as a path of least resistance. Indeed, recent work has shown that these kinds of technical decisions can have important substantive consequences – consequences that may be hard for the discipline to identify (Alvarez, 2016; Denny and Spirling, 2018).

A byproduct of this second major challenge is a false sense of complacency induced by seemingly technical choices. It is naïve to think that data can ‘speak for itself’ in a way that is untarnished by human biases. All data has a history – a distinctly *human* one at that. Contemporary sources of the massive online data-sets typically studied using the tools of ML, for instance, may reflect the commercial intents of those designing social platforms, rather than the actual preferences and typical behaviors of their users (Ruths and Pfeffer, 2014). Similarly, implicit and explicit human biases are embedded in data that exist already, and predictive models designed to learn patterns and trends will reproduce (and magnify) biases that went into the generation of data to begin with (Dressel and Farid, 2018).

Finally, it is important to remember that no amount of modeling sophistication or data size can, in its own right, correct the issues that plague observational studies. Selection on unobservables, interference, treatment

heterogeneity – all of these issues will continue to pose threats to valid causal inferences. Unless description is the only goal of the learning exercise, researchers will need to justify the kinds of inferential jumps needed to move from prediction to explanation (Grimmer, 2015). Of course, this challenge is perhaps the discipline's greatest opportunity to continue offering meaningful contributions to the field of ML: given our history and scholarly interests, we are uniquely positioned to generate analytic strategies at the intersection of sophisticated computation and careful attention to issues of causal identification with observational data.

## Notes

- 1 Unsupervised models can be similarly classified based on the nature of the learning output: tasks that result in continuous outputs are typically called *dimensionality reduction* problems (and typically involve some variant of the factor-analytic model), whereas tasks that result in discrete outputs are typically called *clustering* problems.
- 2 This is also an example of the so-called *bias-variance trade-off*, which suggests that there is an optimal level of bias that can be achieved such that only a small price is paid in terms of variance of the predictor. Generalization error, which can be thought of as a function of both bias *and* variance (in addition to fundamental uncertainty) of the estimator implied by the model, is therefore at a minimum when this balance is achieved (Hastie et al., 2009).
- 3 Non-binary splits can always be represented as a sequence of binary splits, and, as a result, most tree-based algorithms rely on the latter.
- 4 Typically, missing predictors in the test-set are handled using *surrogate splits* – splits based on non-missing variables that result in similar reductions in loss as the original splitting feature.
- 5 Cost-complexity pruning, for instance, was originally proposed by Breiman et al. (1984) to reduce the likelihood of overfitting and consists of going over the internal nodes of a deeply grown tree and sequentially collapsing those that reduce loss the least. It has been shown that this is equivalent to finding a sub-tree that minimizes a penalized loss-term, where the penalty term  $\lambda B$  is linear on the number of terminal nodes  $B$ , and the choice of the strength of penalization  $\lambda$  is typically chosen via cross-validation. Yet another alternative is given by *conditional inference trees*, proposed by Hothorn et al. (2006), which choose splitting variables and values using a non-parametric significance test of association between predictors and the outcome of interest, thus reducing feature selection bias and focusing on statistically predictive variables.
- 6 Incidentally, this is yet another manifestation of the bias-variance tradeoff: larger  $C$  values result in a low-bias/high-variance estimator, whereas smaller  $C$  values (and consequently a higher number of support vectors) results in low-variance/high bias estimators.
- 7 Because a sample of documents from every member of Congress would be needed and the use of each type of speech occurs in a small proportion of the text (approximately 10% or less of the time for each), random sampling would likely miss much of the picture. Instead, they drew a weighted random sample of documents.
- 8 They tested  $\gamma$  at 1, 10, 100, 1,000, and 10,000.
- 9 They tested the linear and RBF kernels.
- 10 For a longer description of the process, see the methods note in the original report.

## REFERENCES

- Alvarez, R Michael (Ed). 2016. *Computational social science*. Cambridge: Cambridge University Press.
- Bishop, CM 2016. *Pattern recognition and machine learning*. New York: Information Science and Statistics Springer. URL: <https://books.google.com/books?id=kOXDtAEACAAJ>
- Breiman, Leo. 2001. Random forests. *Machine learning* 45(1): 5–32.
- Breiman, Leo et al. 1998. Arcing classifier (with discussion and a rejoinder by the author). *The annals of statistics* 26(3): 801–849.
- Breiman, Leo, Jerome H Friedman, Richard A Olshen and Charles J Stone. 1984. *Classification and regression trees*. London: Chapman and Hall/CRC.
- Cortes, Corinna and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning* 20(3): 273–297.
- Denny, Matthew J and Arthur Spirling. 2018. Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. *Political analysis* 26(2): 168–189.
- Dressel, Julia and Hany Farid. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science advances* 4(1): eaao5580.

- Efron, Bradley and Trevor Hastie. 2016. *Computer age statistical inference*. Cambridge: Cambridge University Press.
- Freund, Yoav and Robert E Schapire. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences* 55(1): 119–139.
- Friedman, Jerome H. 2001. Greedy function approximation: A gradient boosting machine. *Annals of statistics* 29(5): 1189–1232.
- Friedman, Jerome H. 2002. Stochastic gradient boosting. *Computational statistics & data analysis* 38(4): 367–378.
- Gelman, Andrew and Eric Loken. 2013. The garden of forking paths: Why multiple comparisons can be a problem, even when there is no ‘fishing expedition’ or ‘p-hacking’ and the research hypothesis was posited ahead of time. Unpublished manuscript. *Department of Statistics, Columbia University*.
- Ghitza, Yair and Andrew Gelman. 2013. Deep interactions with MRP: Election turnout and voting patterns among small electoral subgroups. *American journal of political science* 57(3): 762–776.
- Green, Donald P and Holger L Kern. 2012. Modeling heterogeneous treatment effects in survey experiments with Bayesian additive regression trees. *Public opinion quarterly* 76(3): 491–511.
- Grimmer, Justin. 2015. We are all social scientists now: How big data, machine learning, and causal inference work together. *PS: Political science & politics* 48(1): 80–83.
- Gründler, Klaus and Tommy Krieger. 2015. *Using support vector machines for measuring democracy*. Technical Report Discussion Paper Series, Chair of Economic Order and Social Policy, Universität Würzburg.
- Hainmueller, Jens and Chad Hazlett. 2014. Kernel regularized least squares: Reducing misspecification bias with a flexible and interpretable machine learning approach. *Political analysis* 22(2): 143–168.
- Hastie, Trevor, Robert Tibshirani and Jerome Friedman. 2009. *The elements of statistical learning*. Vol. 1. New York: Springer.
- Ho, Daniel E, Kosuke Imai, Gary King and Elizabeth A Stuart. 2007. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political analysis* 15(3): 199–236.
- Hothorn, Torsten, Kurt Hornik and Achim Zeileis. 2006. Unbiased recursive partitioning: A conditional inference framework. *Journal of computational and graphical statistics* 15(3): 651–674.
- Imai, Kosuke and Aaron Strauss. 2011. Estimation of heterogeneous treatment effects from randomized experiments, with application to the optimal planning of the get-out-the-vote campaign. *Political analysis* 19(1): 1–19.
- Imai, Kosuke, James Lo and Jonathan Olmsted. 2016. Fast estimation of ideal points with massive data. *American political science review* 110(4): 631–656.
- Jordan, Michael I and Tom M Mitchell. 2015. Machine learning: Trends, perspectives, and prospects. *Science* 349(6245): 255–260.
- Kaufman, Aaron, Peter Kraft and Maya Sen. 2018. Improving supreme court forecasting using boosted decision trees. URL: <http://j.mp/2nRjto6>.
- Messing, Solomon, Patrick VanKessel, Adam Hughes, Rachel Blum and Nick Judd. 2017. Partisan Conflict and Congressional Outreach. *Pew Research Center Report*.
- Montgomery, Jacob M and Santiago Olivella. 2018. Tree-based models for political science data. *American Journal of Political Science* 62(3): 729–744.
- Muchlinski, David, David Siroky, Jingrui He and Matthew Kocher. 2016. Comparing random forest with logistic regression for predicting class-imbalanced civil war onset data. *Political analysis* 24(1): 87–103.
- Murphy, KP and F Bach. 2012. *Machine learning: A probabilistic perspective*. Adaptive Computation and Machine Learning Series. Cambridge, MA: MIT Press. URL: <https://books.google.com/books?id=RC43AgAAQBAJ>
- Ng, Andrew Y and Michael I Jordan. 2002. On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. In *Advances in neural information processing systems (conference proceedings)*, pp. 841–848.
- Ratkovic, Marc and Dustin Tingley. 2017. Sparse estimation and uncertainty with application to subgroup analysis. *Political analysis* 25(1): 1–40.
- Roberts, Margaret E, Brandon M Stewart, Dustin Tingley, Christopher Lucas, Jetson

- Leder-Luis, Shana Kushner Gadarian, Bethany Albertson and David G Rand. 2014. Structural topic models for open-ended survey responses. *American journal of political science* 58(4): 1064–1082.
- Ruths, Derek and Jürgen Pfeffer. 2014. Social media for large studies of behavior. *Science* 346(6213): 1063–1064.
- Schapire, Robert E. 1990. The strength of weak learnability. *Machine learning* 5(2): 197–227.
- Wager, Stefan and Susan Athey. 2018 Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American statistical association* 113(523): 1228–1242.
- Wager, Stefan, Trevor Hastie and Bradley Efron. 2014. Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. *The journal of machine learning research* 15(1): 1625–1651.
- Witten, Ian H, Eibe Frank, Mark A Hall and Christopher J Pal. 2016. *Data mining: Practical machine learning tools and techniques*. Burlington, MA: Morgan Kaufmann.

PART VI

# Qualitative and 'Mixed' Methods





*This page intentionally left blank*



# Set Theoretic Methods

Adrian Duşa

## INTRODUCTION

There are various ways to analyze social phenomena. The traditional, qualitative, and quantitative approaches involve specialized languages and seemingly incompatible methods – but such phenomena can also be framed in terms of set relations, as it is often the case in common, everyday language. For instance, poverty research can either employ quantitative, nationally representative samples, or they can use case studies to unfold particular, exemplar life stories that are usually obscured by numbers, or it can be framed in a set theoretical perspective, as recently demonstrated by Ragin and Fiss (2017).

Ragin and Fiss studied the relation between poverty and various configurational patterns that include race, class, and test scores and found that white people are mainly characterized by multiple advantages that protect them from poverty, while there are configurations of disadvantages that are mainly prevalent in black people. These disadvantages do not necessarily

lead to poverty, with an important exception: when they combine with black women.

Being black, being a woman, and having children, along with a configuration of disadvantages, are factors that are more than sufficient to explain poverty. This approach is less concerned about the relative effects of each independent variable included in the model, but rather about identifying membership in a particular set (in the current example, of disadvantaged black women). It is a set relational perspective, more precisely with a focus on set intersections to explain social phenomena.

This chapter begins with a short background of set theory and the different types of sets that are used in the social sciences. It presents the most important set operations that are commonly used in the mathematical framework behind a set theoretical methodology, and it shows how to formulate hypotheses using sets and exemplifies how to calculate set membership scores via the different calibration methods. Finally, it presents important concepts related to necessity

and sufficiency and ends with a discussion about how to apply set theory in Qualitative Comparative Analysis (QCA).

### SHORT BACKGROUND OF SET THEORY

Formally initiated by philosopher and mathematician Georg Cantor at the end of the 19th century (Dauben, 1979), classical set theory became part of the standard foundation of modern mathematics, well suited for the treatment of numbers (Pinter, 2014). Elementary mathematics is embedded with notions such as the set of real numbers, or the set of natural numbers, and formal demonstrations almost always employ sets and their elements as inherent, prerequisite properties of a mathematical problem.

It is nowadays called the naive set theory (Halmos, 1974), and was later extended to other versions, but the basic properties prevailed. A set can be defined as a collection of objects that share a common property. If an element  $x$  is a member of a set  $A$ , it is written as  $x \in A$ , and if it is not a member of that set, it is written as  $x \notin A$ . This is the very essence of what is called binary crisp sets, where objects are either in or out of a set.

For any object, it can be answered with ‘yes’ if it is inside the set and ‘no’ if it is not. There are only two possible truth values in this version: 1 (true) and 0 (false) – a country is either in, or outside the EU, a law is either passed or not passed, an event either happens or does not happen, etc. It has certain roots into Leibniz’s binary mathematics from the beginning of the 18th century (Aiton, 1985), later formalized into a special system of logics and mathematics called Boolean algebra, to honor George Boole’s work in the mid 19th century.

In formal notation, a membership function can be defined to attribute these two values:

$$\mu_A(x) = \begin{cases} 0 & \text{if } x \notin A \\ 1 & \text{if } x \in A \end{cases}$$

It is perhaps worth mentioning that the work of all these people was influenced by the Aristotelian logic, a bivalent system based on three principles (laws of thought): the principle of identity, the principle of non-contradiction, and the principle of excluded middle. A single truth value could be assigned for any proposition (either true, or false), but this was only possible for past events. No truth value could be assigned to a proposition referring to the future, since a future event has not yet happened. Future events can be treated deterministically (what is going to be, is going to be) or influenced by peoples’ free will (we decide what is going to be), leading to a paradox formulated by Aristotle himself.

A solution to this problem was proposed by the Polish mathematician Łukasiewicz (1970), who created a system of logic at the beginning of the 20th century that extends the classical bivalent philosophy. His system (denoted by  $\mathbb{L}_3$ ) presents not just two but three truth values:

$$\mu_A(x) = \begin{cases} 0 & \text{false} \\ \frac{1}{2} & \text{undetermined (neither} \\ & \text{false nor true, or partially true)} \\ 1 & \text{true} \end{cases}$$

Łukasiewicz’s system (using a finite number of values) was eventually generalized to multivalent systems with  $n = v - 1$  values, obtained through a uniform division of the interval  $[0, 1]$ :

$$\left\{ 0 = \frac{0}{n}, \frac{1}{n}, \frac{2}{n}, \dots, \frac{n}{n} = 1 \right\}$$

While some phenomena are inherently bivalent (an element is either in, or out of a set), there are situations where two values are unable to describe the whole picture. A social problem is not necessarily solved or unsolved but can be more or less dealt with. A country is not simply rich or poor, but it can be more or less included in the set of rich countries. There is a certain degree of uncertainty regarding the truth value, which was modeled in the middle of the 20th

century by another great mathematician who laid out the foundations of the fuzzy sets (Zadeh, 1965). These types of sets have a continuous (infinite) number of membership values, in the interval bounded by 0 (completely out of the set) to 1 (completely in the set).

**SET OPERATIONS**

Set operations are mathematical transformations that reflect the logical relations between sets to reflect various configurations that involve intersections, unions, and/or negations. The simplest way to think about these operations is an analogy using basic mathematical algebra: addition, subtraction, multiplication, and division are all very simple – but they are essential operations to build upon. In set theory, there are essentially three main operations that are used extensively in the set theoretical research and comparative analysis: set intersection, set union, and set negation.

These operations perform differently for crisp and fuzzy sets, but the fuzzy version is more general and can be applied to crisp situations as well.

**Set Intersection (Logical AND)**

In the crisp version, the goal of this operation is to find the common elements of two sets. A truth value is involved, that is, it is assigned a ‘true’ value if the element is common and a ‘false’ if otherwise. Out of the four possible combinations of true/false values in Table 57.1 for the membership in the two sets, only one is assigned a ‘true’ value for the intersection, where both individual values are true.

This is called a ‘conjunction’, meaning the logical AND expression is true only when both sets are (conjunctively) true. It is usually denoted using the ‘∩’ or multiplication ‘.’ signs.

The fuzzy version of the set intersection formula is obtained by calculating the minimum between two (or more) values:

$$A \cap B = \min(A, B) \tag{1}$$

As the minimum between 0 (false) and any other truth value is always 0, this formula holds for the data from Table 57.1, where a minimum equal to 1 (true) is obtained only when both values are equal to 1.

**Set Union (Logical OR)**

The counterpart of the set intersection is the set union, used to form larger and larger sets by pulling together all elements from all sets. In the crisp sets version, the result of the union operation is true if the element is part of at least one of the sets. Contrary to set intersection, the only possible way to have a false truth value is the situation where an object is not an element of any of the (two) sets:

The union of two sets is called a ‘disjunction’ and it is usually denoted with the ‘∪’ or ‘+’ signs, and the later should not be confused with the arithmetic addition.

The fuzzy version of this operation is exactly the opposite of the set intersection, by calculating the maximum between two (or more) values:

$$A \cup B = \max(A, B) \tag{2}$$

**Table 57.1 Set intersection for crisp sets**

A		B		
0	AND	0	=	0
0	AND	1	=	0
1	AND	0	=	0
1	AND	1	=	1

**Table 57.2 Set union for crisp sets**

A		B		
0	OR	0	=	0
0	OR	1	=	1
1	OR	0	=	1
1	OR	1	=	1

**Table 57.3 Set negation for crisp sets**

		A	
NOT	0	=	1
NOT	1	=	0

### Set Negation

Set negation is a fundamental operation in set theory, consisting of finding the complement of a set,  $A$ , from a universe,  $U$  (which is a different set of its own, formed by the elements of  $U$  that are not in  $A$ ). It is many times denoted with the ' $\sim$ ' or ' $\neg$ ' signs, and sometimes (especially in programming) with the exclamation sign '!'.

Negating multivalued crisp sets involves taking all elements that are not equal to a specific value. It is still a binary crisp operation, first by coercing the multivalued set into a binary crisp one and then by negating the resulting values.

Negation is a unary operation, and its fuzzy version is a simple subtraction from 1:

$$\sim A = 1 - A \quad (3)$$

The importance of set negation will be revealed later, especially when comparing the quantitative methods with the set theoretical ones, to reveal a certain asymmetry that is specific to sets, with a methodological effort to explain both the presence and the negation (its absence) of a certain phenomenon.

## FORMULATING HYPOTHESES USING SETS

There are multiple ways to conceptualize, measure, and hypothesize social and political phenomena. Previous chapters from this book present several such approaches, from the quantitative types centered on variables to qualitative methods focused on cases. The quantitative approach relies on very precise statistical properties stemming from large samples, and it describes the net effect of

each independent variable on the outcome (the dependent variable), controlling for all other variables in the model. It is a relatively straightforward, albeit with specialized statistical language that is extensively used in quantitative research, however, it is not the most common language to formulate scientific hypotheses.

Hsieh (1980), Novák (1991), Arfi (2010), and even Zadeh (1983) himself have shown how the set theory, and especially the fuzzy sets, can be related to the natural language. Moreover, and contrary to most common expectations, scientific hypotheses do not usually mention the specific net effects of various independent variables, instead they seem very compatible with the language of sets, much like the natural language.

For instance, hypothesizing that democratic countries do not go to war with each other (Babst, 1964) can be naturally translated into sets. The elements are countries, and there are two sets involved: the set of democratic countries, and the set of countries that do not go to war with each other. It is the type of hypothesis that can be best expressed in terms of sufficiency and subset relation, but for the moment, it should suffice to state that it is a concomitant membership of the two sets: those countries that are included in the set of democratic countries are also included in the set of countries that do not go to war with each other.

The same type of language can be applied to another common type of hypothesis in an if-then statement, for instance: 'if a student passes the final exam, then he or she graduates'. Here, too, it is about two sets: the set of students who pass the final exam, and the set of students who graduate, membership in the first guarantees membership in the second.

It seems natural to specify such hypotheses in terms of set language, both in fuzzy sets form (more or less democratic countries) and even binary crisp form (either graduate, or not). Scientific thinking, at least in the social and political sciences, is a constant interplay between abstractization and exact

measurement: we first start by specifying the (pre)conditions that make a certain outcome possible, and only then do we measure the exact values for each such condition or variable.

A statement such as: ‘welfare is education and health’ does not mention any specific values of the education, or of the health, that produce the welfare. This is but one among many possible causal recipes (in the vein of the welfare typologies contributed by Esping-Andersen, 1990) where only the ingredients (education and health) are mentioned without specifying the exact net effects that are needed to produce welfare. It is entirely possible to assign precise mathematical numbers to sets (more exactly, to set membership scores), which is the topic of the next section, but formulating hypotheses is more a matter of specifying abstract concepts (similar to sets) and less about exact values for each.

Using a different perspective on the relation between fuzzy sets and natural language, George Lakoff rejects the notion that natural language can be perfectly mapped over the set theory (Ramzipoor, 2014). He also criticizes Charles Ragin’s approach that assigns membership scores (presented in the next section about set calibration), based on his expertise combining linguistics and cognitive science. More recently, Mendel and Korjani (2018) propose a new method using the Type-2 fuzzy sets.

The whole debate is extremely interesting, for social science concepts have a dual nature stemming from both linguistics and theoretical corpus, but it is by now evident that set theory is well established in social and political research. Conceptual thinking has a long tradition in sociology, with Max Weber’s ideal types being similar to set theoretic concepts that play a central role in comparative analysis. In fact, the whole process of concept formation is embedded with the language of set theory (Mahoney, 1980; Goertz, 2006b; Schneider and Wagemann, 2012).

Despite the predominance of the quantitative methods in the social and political

sciences, there are situations where statistical analyses are impossible (mainly due to a very small number of cases) and, most importantly, where the use of set theory is actually more appropriate, for both formulating and testing theories.

## SET CALIBRATION

In the natural sciences, assigning membership scores to sets is a straightforward procedure. Objects have physical properties that can be measured and transformed into such membership scores. In the social and political sciences, the situation is much more complex. These sciences deal with highly complex phenomena that can only be conceptualized at a very abstract level. They do not exist in the physical reality and do not have visible properties to measure directly.

Concepts are very abstract things, and their measurement is even more complex: it depends on theory, which determines their definition which, in turn, has a direct effect over their operationalization which has an influence on constructing the research instrument – only then can some measurements be collected.

Each of these stages require highly specialized training involving years (sometimes a lifetime) of practice before mastering the activity. Theoreticians are rare, or at least those who have a real impact over the research praxis of the entire academic community. Most researchers follow a handful of theories that attempt to explain the social and political reality. Each such theory should be ideally reflected into a clear definition of the abstract concept.

Based on the definition, the process of operationalization is yet another very complex step towards obtaining some kind of numerical measurements about the concept. It is based on the idea that, given the impossibility of directly measuring the concept, researchers can only resort to measuring its effect over the observable reality. For instance, we cannot tell

how altruistic a person is unless we observe how the person behaves in certain situations related to altruism. There are multiple ways for a person to manifest this abstract concept, and the operationalization is a process that transforms a definition into measurable indicators, usually via some other abstract dimensions and subdimensions of the concept.

Finally, obtaining numerical scores based on the indicators from the operationalization phase is yet another complex activity. There are multiple ways to measure (counting only the traditional four levels of measurement: nominal, ordinal, interval, and ratio, but there are many others), and the process of constructing the research instrument, based on the chosen level of measurement for each indicator, is an art. It is especially complex as the concepts should also be equivalent in different cultures, and huge efforts are being spent to ensure the compatibility between the research instruments from different languages (translation being a very sensitive activity).

The entire process ends up with some numerical measurements for each indicator, and a final task to aggregate all these numbers to a single composite measure that should be large if the concept is strong, and small if the concept is weak. In the above example, highly altruistic people should be allocated large numbers, while unconcerned people should be allocated low numbers, both on a certain numerical scale.

In set theory, calibration is the process of transforming these (raw) numbers into set membership scores, such that a completely altruistic person should receive a value of 1, while a non-altruistic person should receive a value of 0. This process is far from straightforward, even for the natural sciences.

Describing the procedure, Ragin (2008) makes a distinction between 'calibration' and 'measurement' processes and exemplifies with temperature as it is a directly measurable physical property. While exact temperatures can be obtained from absolute zero to millions of degrees, no such procedure would

even be able to automatically determine what is 'hot' and what is 'cold'. These are human interpreted concepts and need to be associated with some subjective numerical anchors (thresholds). On the Celsius scale, 0 degrees is usually associated with cold, while 100 degrees is usually associated with very hot, and these numbers are not picked at random. They correspond to the points where the water changes states: to ice at 0 degrees and to steam at 100 degrees, when the water boils.

The choice of thresholds is very important, for it determines the point where something is completely out of a set (for instance at 0 degrees, the ice is completely out of the set of hot matter) and the point where something is completely inside the set (at 100 degrees, steam is completely inside the set of hot matter). A third threshold is also employed called the 'crossover': the point of maximum ambiguity where it is impossible to determine whether something is more in than out of a set, corresponding to the set membership score of 0.5.

The set of thresholds (exclusion, crossover, and inclusion) is not universal, even for the same concept. A 'tall' person means one thing in countries like Norway and Netherlands, where the average male height is more than 1.8 m, and another thing in countries like Indonesia and Bolivia, where the average is about 1.6 m. It is the concept that matters – not its exact measurement – therefore, different thresholds need to be used in different cultural contexts, depending on the local perception.

Traditionally, there are two types of calibrations for each type of sets, crisp and fuzzy. Calibrating to crisp sets is essentially a matter of recoding the raw data and establishing a certain number of thresholds for each value of the calibrated set. When binary crisp sets are intended to be obtained, a single threshold is needed to divide the raw values in two categories: those below the threshold will be allocated a value of 0 (out of the set) and for those above the threshold, a value of 1 (in the

set). When multivalued crisp sets are intended, there will be two thresholds to divide into three categories, and so on. The general formula for the number of thresholds is the number of values minus 1.

Even for this (crude) type of recoding, the values of the thresholds should not be mechanically determined. A statistician will likely divide the values using the median, which would, in many cases, be a mistake. It is not the number of cases that should determine the value of the threshold, but rather the meaning of the concept and the expert's intimate knowledge about which cases belong to which category.

For instance, there will be a certain value of the threshold to divide countries' GDP in the set of 'developed countries' and a different value of the threshold for the set of 'very developed countries'. The exact value should be determined only after an inspection of the distribution of GDP values, especially if they are not clearly clustered. In such a situation, the researcher's experience should act as a guide in establishing the best threshold value that would correctly separate different countries in different categories, even if the difference is small. The whole of this process should be thoroughly described in a dedicated methodological section, with strong theoretical justifications for the chosen value of the threshold.

Calibrating to fuzzy sets is more challenging and, at the same time, more interesting because there are multiple ways to obtain fuzzy membership scores from the same raw numerical data. The most widely used is called the 'direct method', first described by Ragin (2000). It uses the logistical function to allocate membership scores, using the exclusion, cross-over, and inclusion thresholds.

Table 57.4 below displays the two relevant columns extracted from Ragin's book, the first showing the national income in US dollars and the second showing the degree of membership (the calibrated counterparts of the national income) into the set of developed countries.

**Table 57.4 Per capita income (INC), calibrated to fuzzy sets membership scores (fsMS)**

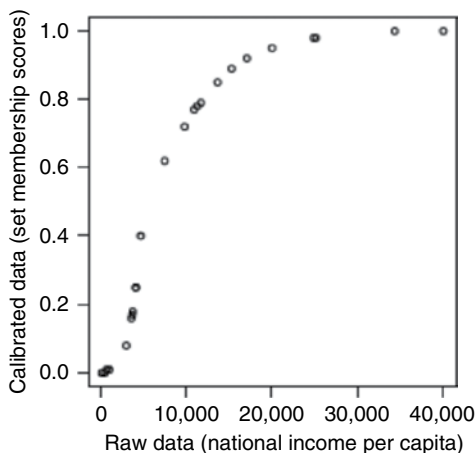
	<i>INC</i>	<i>fsMS</i>
Switzerland	40,110	1.00
United States	34,400	1.00
Netherlands	25,200	0.98
Finland	24,920	0.98
Australia	20,060	0.95
Israel	17,090	0.92
Spain	15,320	0.89
New Zealand	13,680	0.85
Cyprus	11,720	0.79
Greece	11,290	0.78
Portugal	10,940	0.77
Korea, Rep.	9,800	0.72
Argentina	7,470	0.62
Hungary	4,670	0.40
Venezuela	4,100	0.25
Estonia	4,070	0.25
Panama	3,740	0.18
Mauritius	3,690	0.17
Brazil	3,590	0.16
Turkey	2,980	0.08
Bolivia	1,000	0.01
Cote d'Ivoire	650	0.01
Senegal	450	0.00
Burundi	110	0.00

At the top of the list, Switzerland and the United States are highly developed countries, which explains their full membership score of 1, while Senegal and Burundi, with national incomes of 450 USD and 110 USD respectively, are too poor to have any membership whatsoever in the set of developed countries.

What threshold values best describe this set, and how are the membership values calculated? A quick quantitative solution would be to calculate the ratio of every other country from the income of Switzerland, the richest country in that data.

Aside from the fact such a method is mechanical and data driven, it would immediately become obvious that, for instance, the Netherlands (which currently has an almost full inclusion of 0.98 in the set of developed





**Figure 57.1 Calibration in the set of developed countries**

countries) would have a ratio equal to 0.628, which does not seem to accurately reflect our knowledge. Likewise, a median value of 8,635 USD would leave Argentina more out of the set than more in, and the average of 11,294 USD is even more misleading, leaving Greece more out than in.

Ragin started by first deciding the crossover threshold at a value of 5,000 USD, which is the point of maximum ambiguity about a country being in more in than more out of the set of developed countries. He then applied some mathematical calculations based on the logistic function and the associated log odds, arriving at a full inclusion score of 20,000 USD (corresponding to a membership score of at least 0.95 and a log odds of membership of at least +3) and a full exclusion score of 2,500 USD (corresponding to a membership score of at most 0.05 and a log odds of membership lower than -3).

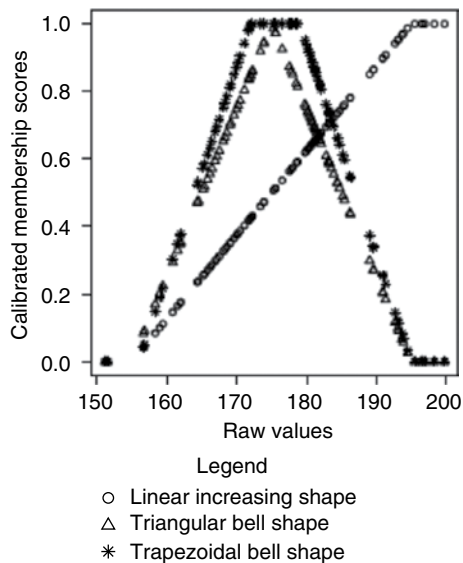
Employing the logistic function, the generated set membership scores follow the familiar increasing S shape displayed in Figure 57.1, but this function is only one among many other possible ones to perform calibration. Linear mathematical transformations are also possible, such as the one from the Equation (4), as extracted from Duşa (2019: 84):

$$dm_x = \begin{cases} 0 & \text{if } x \leq e, \\ \frac{1}{2} \left( \frac{e-x}{e-c} \right)^b & \text{if } e < x \leq c, \\ 1 - \frac{1}{2} \left( \frac{i-x}{i-c} \right)^a & \text{if } c < x \leq i, \\ 1 & \text{if } x > i. \end{cases} \quad (4)$$

where:

- *e* is the threshold for full exclusion
- *c* is the crossover
- *i* is the threshold for full inclusion
- *x* is the raw value to be calibrated
- *b* determines the shape below the crossover (linear when *b* = 1 and curved when *b* > 1)
- *a* determines the shape above the crossover (linear when *a* = 1 and curved when *a* > 1)

The calibration functions in Figure 57.2 refer to the calibration of 100 randomly selected heights ranging from 150 cm to 200 cm. These values are calibrated in the set of ‘tall people’ (the linear increasing function that could act as a replacement for the logistical S shape) as well as in the set of ‘average height



**Figure 57.2 Other possible calibration functions**

people’ (with a triangular shape, and also with a trapezoidal shape). This is an example that shows how the calibrated values depend on the conceptual meaning of the calibrated set. All shapes refer to people’s heights and all use the exact same raw values, but the meaning is different for ‘average height’ and for ‘tall’ people.

The set of three threshold values (155 cm for full exclusion, 175 cm for the crossover, and 195 cm for full inclusion) can be used only for the increasing linear that approximates an S shape for the set of ‘tall’ people. The other linear functions that approximate a bell shape (for the set of ‘average height’ people) are more challenging, and need a set of six values for the thresholds (three for the first part that increases towards the middle, and the other three for the second part that decreases from the middle towards the higher heights). There are two full exclusion thresholds, two crossover values, and, finally, two full inclusion thresholds (that coincide for the triangular shape), with the calibrated values being obtained via the mathematical transformations from Equation (5):

$$dm_x = \begin{cases} 0 & \text{if } x \leq e_1, \\ \frac{1}{2} \left( \frac{e_1 - x}{e_1 - c_1} \right)^b & \text{if } e_1 < x \leq c_1, \\ 1 - \frac{1}{2} \left( \frac{i_1 - x}{i_1 - c_1} \right)^a & \text{if } c_1 < x \leq i_1, \\ 1 & \text{if } i_1 < x \leq i_2, \\ 1 - \frac{1}{2} \left( \frac{i_2 - x}{i_2 - c_2} \right)^a & \text{if } i_2 < x \leq c_2, \\ \frac{1}{2} \left( \frac{e_2 - x}{e_2 - c_2} \right)^b & \text{if } c_2 < x \leq e_2, \\ 0 & \text{if } x > e_2. \end{cases} \quad (5)$$

Apart from the direct method, Ragin also presented an ‘indirect’ one in order to obtain fuzzy membership scores from interval level raw data. In this method, no qualitative anchors (thresholds) need to be specified in advance, but, rather, it involves creating an

artificial dependent variable where each case is allocated a certain fuzzy membership category from 0 to 1 (usually six, an even number to avoid the point of maximum ambiguity 0.5), then performs a (quasi)binomial logistic regression using a fractional polynomial equation with the raw values as an independent variable against the newly formed dependent variable containing the fuzzy membership categories (for more details, see Duşa, 2019: 92).

A different type of calibration is applied for categorical causal conditions (for instance, containing values from a response Likert type scale). It is not possible to determine any thresholds because the variation is extremely small and data can sometimes be severely skewed, which limits the variation even more. For the same reasons, no regression equation can be applied with the indirect method, since it assumes at least the independent variable to be metric.

A possible solution to this problem is to manually allocate fuzzy membership scores for each category (the so-called ‘direct assignment’ method), introduced by Verkuilen (2005) who also criticized it for containing bias due to researcher’s subjectivity. Verkuilen mentions a possibly better solution, by employing the Totally Fuzzy and Relative (TFR) method (Cheli and Lemmi, 1995), making use of the empirical cumulative distribution function of the observed data  $E$ , then calculating the fraction between the distance from each CDF value  $E(x)$  to the CDF of the first value from the Likert scale  $E(1)$ , and the distance from 1 (the maximum possible fuzzy score) to the same  $E(1)$ :

$$TFR = \max \left( 0, \frac{E(x) - E(1)}{1 - E(1)} \right) \quad (6)$$

Calibration is a very important topic in set theoretical methods, as many of the subsequent results depend on this operation. It should not be a mechanical process, but rather an informed activity where the researcher should present the methodological reasons that led to one method or another.

## SET MEMBERSHIP SCORES VS. PROBABILITIES

Despite this topic being discussed numerous times before (Dubois and Prade, 1989; Kosko, 1990; Zadeh, 1995; Ragin, 2008; Schneider and Wagemann, 2012), and despite attempts to combine set theory and statistics (Heylen and Nachtegael, 2013), set membership scores and probabilities can still be confused as both range from 0 to 1 and, at a first glance, seem very similar.

Before delving into the formal difference, consider the following example involving a potentially hot stove. If the stove has a 1% probability of being very hot, there is still a small (but real) chance to get severely burned when touching it. However, if we say the stove has a 1% inclusion in the set of hot objects, the stove can be safely touched without any risk of getting burned.

Ragin's example with the glass of water has the same interpretation. If there is a 1% probability the glass will contain a deadly poison, there is a small but definite chance of dying after drinking that water. But if the glass has a 1% inclusion in the set of poisonous drinks, there is absolutely no risk of dying.

Intuitive as they may seem, these two examples still don't explain the fundamental difference. At the formal level, the probability has to obey the Kolmogorov axioms:

- the probability of an event that is certain is equal to 1:  $P(C) = 1$
- the probability of an impossible event is equal to 0:  $P(\emptyset) = 0$
- if two events do not overlap ( $A \cap B = \emptyset$ ), then  $P(A + B) = P(A) + P(B)$

The probability can essentially be interpreted as a relative frequency obtained from an infinite repetition of an experiment. It is a frequentist statistic (based on what is called a frequentist approach), where the conclusions are drawn from the relative proportions in the data.

However, frequencies can only be computed for categorical variables, in this situation: for events either happening or not. To calculate probabilities (relative frequencies) there are only two possible values for the event: 1 (true, happening) or 0 (false, not happening). The first section already presented the different types of sets, and this corresponds to the definition of a binary crisp set.

Therefore, the meaning of probability is necessarily related to crisp sets, while membership scores are related to fuzzy sets. They simply refer to different things, given that crisp sets are only particular cases of fuzzy sets. Set membership scores refer to various degrees of membership to a set, they are related to the uncertainty about set membership that cannot be computed the same as a probability because the set itself is not crisp, but fuzzy.

When flipping a coin, there are only two possible outcomes (heads or tails) and an exact probability of occurrence for each can be computed by flipping the coin numerous times. These are clear-cut categories (either heads or tails), but not all concepts are so clear. Whether a person is 'young' is a matter of uncertainty, and every person can be included (more, or less) in the set of young people. Same with 'smart', 'healthy', etc., all of which cannot be determined unequivocally.

There are situations where probabilities and fuzzy sets can be combined (Singpurwalla and Booker, 2004; Demey et al., 2017), especially with Bayesian probabilities (Mahoney, 2016; Barrenechea and Mahoney, 2017; Fairfield and Charman, 2017) in conjunction with process tracing, but these two concepts do not completely overlap. In the words of Zadeh (1995) himself, they are 'complementary rather than competitive'.

## POLARITY AND ASYMMETRY

There is an even deeper layer of understanding that needs to be uncovered with respect to

probabilities and fuzzy sets. Describing probability, Kosko (1994: 32) shows that it works with bivalent sets only (an event either happens or it does not happen), and another important difference refers to how a set relates to its negation.

In probability theory,  $A \cap \sim A = \emptyset$ , and  $A \cup \sim A = 1$ . For fuzzy sets, it turns out that  $A \cap \sim A \neq \emptyset$ , and  $A \cup \sim A \neq 1$ . These inequalities (especially the first one) essentially entail that objects can be part of both a set and its negation, and the union of the two sets might not always be equal to the universe.

This has deep implications over how we relate to events, their negation, and the common misperception of bipolarity. Sets are unipolar, therefore a bipolar measurement scale (for instance, a Likert type response scale) cannot be easily accommodated with a single set.

In a bipolar space, ‘good’ is the opposite of ‘bad’; but a ‘not bad’ thing is not precisely the same as a ‘good’ thing: it is just not bad. Same with ‘ugly’ vs. ‘beautiful’: if a thing is not ugly, that does not mean it is necessarily beautiful, or, the other way around, something that is not beautiful is not necessarily ugly. Bauer et al. (2014) encountered similar difficulties in evaluating a bipolar scale with left-right political attitudes, analyzing the vagueness of the social science concepts in applied survey research.

Things, or people, can have membership scores of more than 0.5 in both a set and its negation. A person can be both happy and unhappy at the same time, therefore translating a bipolar scale into a single set is difficult, if not impossible. There should be two sets, first for the happy persons and the second for the unhappy ones, and a person can be allocated membership scores in both, such that the sum of the two scores can exceed 1 (something impossible with probabilities).

The set negation leads to another point of misunderstanding between quantitative statistics (especially the correlation-based techniques, for instance the regression analysis) and set theoretic methods. Numerous

articles have been written comparing empirical results (Katz et al., 2005; Grendstad, 2007; Fujita, 2009; Grofman and Schneider, 2009; Woodside, 2014), pointing to the deficiencies of regression techniques (Pennings, 2003; Marx and Soares, 2015), criticizing fuzzy sets (Seawright, 2005; Paine, 2015; Munck, 2016), and revealing the advantages of fuzzy sets (Cooper and Glaesser, 2010), or, more recently, focusing on the integration and complementarity between the two methods (Skaaning, 2007; Mahoney, 2010; Fiss et al., 2013; Radaelli and Wagemann, 2019).

The sheer amount of written publications suggest at least a couple of things. First, that set theoretic methods are increasingly used in a field traditionally dominated by the quantitative analysis, and second, there is a lot of potential for these methods to be confused (despite the obvious differences) as they both refer to explanatory causal models for a given phenomenon.

Correlation-based techniques assume an ideal linear relation between the independent and dependent variables. When high values of the dependent variable (that can be interpreted as the ‘presence’ of the outcome, in set theory) are explained by high values of the independent variable(s), then low values of the dependent (‘absence’ of the outcome) are necessarily explained by low values of the independent variable(s).

By contrast, set theoretical methods do not assume this kind of linearity. While the presence of the outcome can be explained by a certain configuration of causal conditions, the absence of the outcome can have a very different explanation, involving different causal combinations. If welfare can be explained by the combination of education and health, it is perfectly possible for the absence of welfare to be explained by different causes.

While the correlation-based analyses are symmetric with respect to the dependent variable, the set theoretic methods are characterized by an asymmetric relation between a set of causes and a certain outcome. This is a fundamental ontological difference that

separates the two analysis systems, which should explain both why they are sometimes confused, as well as why their results are seemingly different.

**NECESSITY AND SUFFICIENCY**

Natural language abounds with expressions containing the words ‘necessary’ and ‘sufficient’. In trying to identify the most relevant conditions that are associated with an outcome, theorists often ask: what are the necessary conditions for the outcome? (without which the outcome cannot happen), or what conditions are sufficient to trigger an event? (that, when present, the event is guaranteed to happen).

The contrast between the correlational perspective and the set theoretic methods can be further revealed by analyzing Figure 57.3. The crosstable on the left side is a typical, minimal representation of the quantitative statistical perspective, focused on the perfect correlation from the main diagonal. Everything off the main diagonal is problematic and decreases the coefficient of correlation.

The crosstable on the right side, however, tells a different story. In the language of statistics, the 45 cases in the upper left quadrant potentially ruin the correlation, but they make perfect sense from a set theoretical point of view: since there are no cases in the lower right quadrant, this crosstable tells the story of X being a perfect subset of Y. The ‘problematic’ upper left quadrant simply says there are cases where Y is present and X is

absent – in other words, X does not cover (does not ‘explain’) all of Y.

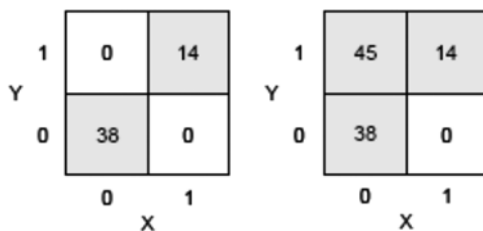
The zero cases in the lower right quadrant – combined with the 14 cases in the upper right quadrant – say there is no instance of X where Y is absent, which means that X is completely included in Y (it is a subset of Y). Whenever X happens, Y happens as well, that is to say X is ‘sufficient’ for Y (‘if X, then Y’).

This is a different type of language, a set theoretical one, that is foreign to the traditional quantitative analysis. Regression analysis and the sufficiency analysis have the very same purpose, to seek the relevant causal conditions for a given phenomenon. However, when inspecting for sufficiency, the focus is not the main diagonal (correlation style) but rather on the right side of the crosstable where X happens (where X is equal to 1).

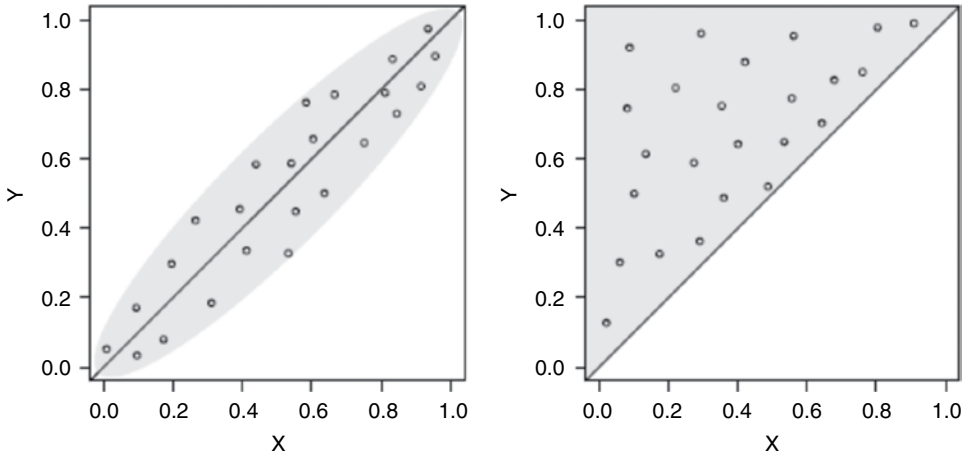
This is naturally a very simplified example using just two values for both X and Y. Quantitative researchers would be right to argue that, when the dependent variable is binary, a logistic regression model is more appropriate than a linear regression model. However, set theoretical data need not necessarily be crisp, they can also be fuzzy with a larger variation between 0 and 1 – a crosstable is not enough to represent the data.

At a closer inspection on Figure 57.4, the situation is identical for fuzzy sets. The left plot displays the characteristic ellipse shape of the cloud, with a very positive correlation between the independent and the dependent variables. It does not really matter whether the points are located above or below the diagonal, as long as they are close.

The cloud of points from the right plot would be considered problematic. Not only are the points located far from the main diagonal (ideally, the regression line), but they also display inconstant variance (a phenomenon called heteroskedasticity). However, this is not problematic for set theory: as long as the points are located above the main diagonal (values of X are always smaller than corresponding values of Y), it is a perfect representation of a fuzzy subset relation. In set theoretical language,



**Figure 57.3 Correlation (left) and subset sufficiency (right)**



**Figure 57.4 Fuzzy correlation (left) and fuzzy subset sufficiency (right)**

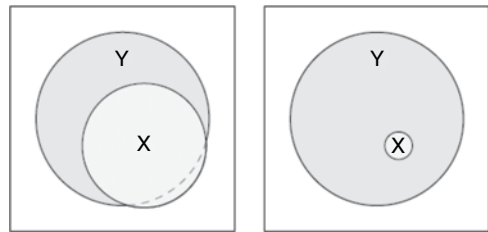
such a subset relation, it is also described as perfectly ‘consistent’.

Not all subset relations are that perfect. In fact, there can be situations where X can happen and Y is absent, without affecting the sufficiency relation (too much). Just as there are no countries with perfect democracies (they are ‘more or less’ democratic), situations with perfect sufficiency are also extremely rare. When perfect sufficiency happens, it is mainly the result of our calibration choice: it can happen in crisp sets, but this is almost never observed with fuzzy sets.

The concept of fuzziness teaches us that conditions can be ‘more or less’ sufficient, just as two sets can be more or less included one into the other. The causal set should be ‘consistent enough’ with (or ‘included enough’ in) the outcome set, to be accepted as sufficient.

The big question is how much of outcome set Y is explained by causal set X, a very common question in traditional statistics that is usually answered with the  $R^2$  coefficient in the regression analysis. In set theory, this is a matter of coverage. There can be situations with imperfect consistency but large coverage, and perfect consistency but low coverage.

Out of the two situations in Figure 57.5, the relation from the left plot is the most relevant. Despite the imperfect consistency (inclusion),



**Figure 57.5 Incomplete inclusion/large coverage (left) and complete inclusion/low coverage (right)**

the causal condition X covers a lot of the cases in the outcome Y, qualifying as a highly relevant (albeit imperfect) sufficient condition for Y.

In the plot from the right side, X is perfectly consistent with Y but it covers only a very small area, which means there are very many cases in Y that are not explained by X, suggesting we should search for more causal conditions that explain the entire diversity of the outcome’s presence. In such situations, X is called sufficient but not necessary, an expression which is also described by the concept of ‘equifinality’: the very same outcome can be produced via multiple causal paths, just as there are many roads that lead to the same city.

Inclusion and coverage can be precisely measured, with the same formula being valid

for both crisp and fuzzy sets. Equation (7) calculates the consistency for sufficiency (*inclS*), while Equation (8) calculates the coverage for sufficiency (*covS*), where the sufficiency relation is denoted by the forward arrow sign  $X \Rightarrow Y$ :

$$inclS_{X \Rightarrow Y} = \frac{\sum \min(X, Y)}{\sum X} \quad (7)$$

$$covS_{X \Rightarrow Y} = \frac{\sum \min(X, Y)}{\sum Y} \quad (8)$$

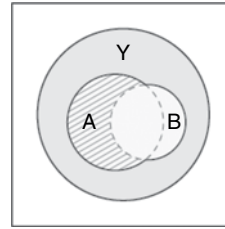
In the regression analysis, independent variables may be collinear, meaning they explain the same part of the dependent variable’s variation. This is usually detected with the contribution of each independent variable to the model’s  $R^2$  coefficient: only those variables that contribute a significant increase of the  $R^2$  are preferred.

Similarly, in set theory, the causal conditions have a so called ‘raw’ coverage and also a ‘unique’ coverage. Their unique coverage (*covU*) is the area from the outcome Y which is solely covered by a certain causal condition, as shown in Equation (9) and Figure 57.6.

$$covU_{A \Rightarrow Y} = \frac{\sum \min(Y, A)}{\sum Y} - \frac{\sum \min(Y, A, \max(B, C, \dots))}{\sum Y} \quad (9)$$

In Figure 57.6, the unique coverage of condition A can be computed as the area of Y covered by A, minus the intersection of A and B (its area jointly covered by condition B). More generally, minus the intersection between A and the union of all other causal conditions that cover the same area of Y covered by A.

Necessity and sufficiency are mirrored concepts. While sufficiency is about the subset relation of the causal condition within the outcome set, necessity is the other way around: the superset relation of the causal condition over the outcome set. A causal



**Figure 57.6 Unique coverage of A (hashed area)**

condition is necessary *iff* it is a superset of the outcome: when Y happens, X is always present. When the outcome Y does not occur in the absence of X, it means that X is necessary.

The upper left quadrant in a  $2 \times 2$  crosstable should be empty (where  $Y = 1$  and  $X = 0$ ), and, correspondingly, the area above the main diagonal in a fuzzy XY plot should also be empty in order to determine necessity.

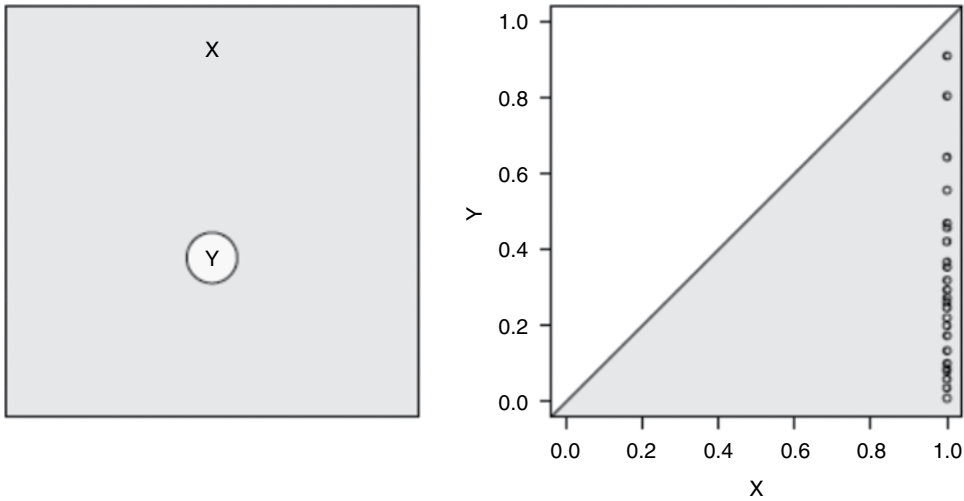
Mirrored scores for the consistency of necessity (*inclN*, how much of Y is included in X), as well as for the coverage of necessity (*covN*, how much of X is covered by Y) can be calculated, as shown in Equations (10) and (11):

$$inclN_{X \Leftarrow Y} = \frac{\sum \min(X, Y)}{\sum Y} \quad (10)$$

$$covN_{X \Leftarrow Y} = \frac{\sum \min(X, Y)}{\sum X} \quad (11)$$

When analyzing necessity, the most important thing is to determine how relevant a necessary condition is. Oxygen is a necessary condition for a fire, but it is an irrelevant necessary condition as oxygen can be found everywhere, and in most situations where oxygen is present, a fire is not observed. A more important necessary condition would be heat, and another necessary condition may be a spark. Both of these are truly necessary (hence relevant) to start a fire.

The relevance of a necessary condition is revealed by the coverage score. If the outcome Y covers only a very small area of the



**Figure 57.7** X as a trivial necessary condition for Y

causal condition, it is a sign that X might be irrelevant.

Goertz (2006a) is a leading scholar in the analysis of necessity, further differentiating between irrelevant and trivial necessary conditions. Irrelevance and triviality are somewhat similar, and they are frequently used as synonyms, but there is a subtle difference between them. Triviality is a maximum of irrelevance, to the point that not only the subset outcome Y covers a very small area of the causal condition X, but the superset condition X becomes so large that it fills the entire universe. When trivial, the causal condition is omnipresent, with no empirical evidence of its absence.

In the previous example, oxygen is an irrelevant, but not exactly a trivial, necessary condition for a fire, as there are many (in fact, most) places in the Universe where oxygen is absent. In the Euler/Venn diagram from Figure 56.7, Y is completely consistent with X but it covers a very small area. Moreover, it can be noticed that X occupies the entire universe represented by the rectangle: it is an omnipresent necessary condition.

The same line of reasoning can be applied on the XY plot from the right side, where the focus on necessity is the area below the main diagonal, and X is trivial since all of its points are

located on the extreme right where X is always equal to 1, and most of the points are located in the lower half of the plot where Y is more or less absent (below the crossover 0.5 point).

A condition becomes less and less trivial (hence more and more relevant) when the points move away from the extreme right, where X is always equal to 1, towards the main diagonal. Goertz proposed a measure of triviality by simply measuring the distance between the fuzzy values and 1. Later, Schneider and Wagemann (2012) advanced Goertz’s work and proposed a measure called Relevance of Necessity (*RoN*), that is the current standard to complement the coverage score for necessity:

$$RoN = \frac{\sum (1 - X)}{\sum (1 - \min(X, Y))} \quad (12)$$

The XY plots and Venn/Euler diagrams have a couple of more interesting properties to discuss. If points are located all over the plot, there is no clear relationship between the cause and the outcome. We expect the points to be positioned either above the main diagonal (for sufficiency) or below (for necessity). If that happens, it means that if a cause is perfectly sufficient, it is usually not necessary



and, conversely, if it is perfectly necessary, it is usually not sufficient.

Ideally, we would like to find a causal condition that is both necessary and sufficient for an outcome, thus having the greatest explanatory power. At a first sight, that would seem impossible since a causal set  $X$  cannot be a subset and a superset of the outcome  $Y$  at the same time. It may in fact happen when the two sets coincide: the subset  $X$  covers 100% of the set  $Y$ .

In terms of  $XY$  plots, the points are located neither above nor below the main diagonal. When the two sets coincide, the points are located exactly along the main diagonal, which would also correspond to a (close to) perfect correlation in the statistical tradition, similar to the left plot from Figure 57.4.

However, such a perfect correlation is difficult to obtain in practice, and it would usually mean we are not dealing with two different concepts (for the cause and for the effect) but with one and the same concept under two different measurements. No causal set is perfectly correlated with the outcome, and, perhaps more importantly, a single causal set is neither necessary nor sufficient by itself. It is very rare to obtain an explanatory model with a single causal condition, a typical outcome being produced by various combinations of causes.

Causal factors combine in conjunction, which in set theory is set intersections. Where a single cause might not be (sufficiently) included into an outcome set, an intersection with other condition(s) might be small enough to fit.

The same thing happens for necessity, but in reverse. If a single causal condition is not big enough to qualify as a necessary superset of the outcome, disjunctions (set unions) of two or more causal conditions might eventually form a big enough superset to cover the outcome. However, if conjunctions are easy to interpret (the simultaneous presence of two causal sets), disjunctions need to have theoretically valid interpretations, much like the quantitative researchers having to find a meaningful interpretation for the latent constructs resulted from the principal component analysis.

More recent and interesting developments in the analysis of necessity include the Necessary Condition Analysis (NCA) by Dul (2016), while on sufficiency, Schneider and Rohlfing (2016) bring important insights in the cutting edge, so called Set Theoretic Multi-Method Research (STMMR) which is an entire topic on its own and deserves a separate and more extended presentation.

## SET THEORY AND THE QUALITATIVE COMPARATIVE ANALYSIS

Having presented the background of set theory, the stage is set to introduce a (third) way to tackle research problems traditionally approached through the qualitative and quantitative methods.

The trouble with quantitative research is that it needs many cases (a large  $N$ ) to make the Central Limit Theorem work, and a typical political science research compares only a handful of countries or events and does not have that many cases. There are only 28 countries in the EU, and a comparative study on the eastern European countries will have even less cases. When studying very rare events such as revolutions, Skocpol (1979) had only three cases to work upon: France, Russia, and China.

It is difficult to argue that there is an underlying, potentially infinite population of 'possible' such events to draw large samples from, in order to justify the use of the quantitative analysis, even with Monte Carlo simulations for small samples. On the other side, the qualitative analysis is very much case oriented and produces perfect explanations for all individual cases. This is often useful for theory formation, but it is usually regarded as too specific to have generalizable value.

With both sides having strong arguments to defend one method or another in different situations, Ragin (1987) employed set theory and Boolean algebra to import a methodology created for electrical engineering (Quine, 1955; McCluskey, 1956) into the social and political sciences. He showed how, through a systematic

**Table 57.5 Boolean minimization example**

A	B	Y
1	1	1
1	0	1
1	-	1

comparative analysis of all possible pairs of cases, the relevant causal factors can be identified and the irrelevant ones eliminated. More importantly, he showed how to identify the patterns, or the combinations of causal conditions, that are sufficient to produce an outcome.

The essence of the entire procedure can be reduced to a process called Boolean minimization, which was itself imported into the electrical engineering from the canons of logical induction formulated by J.S. Mill (1843).

The two expressions in Table 57.5 are equivalent to  $AB + A\sim B$ , which can be simplified to A alone since the condition B is redundant, present in the first, and absent in the second:  $A(B + \sim B) = A$ . In such an example, B is said to be ‘minimized’ (or eliminated), hence the name of the Boolean minimization procedure.

Each case that is added to the analysis displays a certain combination (of presence or absence) of causal conditions, and the algorithm exhaustively compares all possible pairs cases to first identify if they differ by only one literal, then iteratively and progressively minimize until nothing else can be further simplified. The final product of this procedure is the set of so-called ‘prime implicants’, which are simpler (more parsimonious) but equivalent expressions to the initial, empirically observed cases.

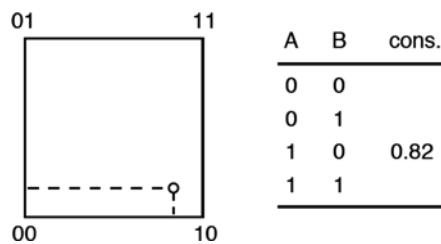
Since pairs of cases are compared, the process is more qualitative than quantitative, therefore the ‘Q’ in QCA stands for the ‘Qualitative’ Comparative Analysis. It has absolutely nothing to do with traditional statistics, yet it employs a systematical and solid mathematical algorithm such as the Boolean minimization to identify the minimal configurations of (relevant) causal conditions which are sufficient to produce an outcome.

Crisp sets are very attractive as they allow one to map the empirically observed configurations over a finite number of combinations of presence/absence for the causal conditions (equal to  $\prod l_c$ , where  $l$  is the number of levels for each causal condition  $c = 1 \dots n$ ). This finite space is called a truth table, and it contains all positive and negative observed configurations, as well as those for which there is no empirical information (called ‘remainders’).

However, it is precisely the ‘Boolean’ nature of the algorithm that attracted a lot of criticism (Goldstone, 1997; Goldthorpe, 1997), since it suggests a very deterministic view of reality (Liebersson, 1991) and, as pointed many times before, most social phenomena are not simply present or absent, but somewhere in between.

The debate led to an upgrade of QCA from Boolean to fuzzy sets (Ragin, 2000, 2008). Instead of crisp values, each case has a membership score for each of the causal condition sets. The challenge, that was also solved by Ragin (2004), was to translate fuzzy membership scores to truth table crisp scores, because the minimization process is Boolean.

In the fuzzy version, the truth table configurations act as the corners of a multidimensional vector space where the set membership scores play the role of fuzzy coordinates for the position of each case. Figure 57.8 presents the simplest possible vector space with two dimensions, and a case having two fuzzy membership scores of 0.85 on the horizontal and 0.18 on the vertical. For only two



**Figure 57.8 Bidimensional vector space (left) and the corresponding truth table (right)**

causal conditions, the truth table contains  $2 \cdot 2 = 4$  rows, represented by the corners of the square, and the case is located close to the lower right corner 10 (that is, presence of the first condition and absence of the second).

It is rather clear to which truth table configuration does the case belong to in this example, but it would be more difficult to assess if the case were located close to the middle of the vector space. Ragin's procedure uses the fuzzy coordinates of each case to calculate consistency scores for each of the corners and determines which corners are the cases more consistent with (or closest to).

The consistency score of this case is the set intersection between the first membership score (0.85) and the negation of the second ( $1 - 0.18 = 0.82$ ), which is the fuzzy minimum between 0.85 and 0.82, equal to 0.82. Provided there is no fuzzy membership score of exactly 0.5 (the point of maximum ambiguity), there is only one corner to which cases have a higher than 0.5 consistency.

The corners of the vector space can be interpreted as genuine ideal types in the Weberian tradition, which an imperfect fuzzy configuration is most similar to. Upon determining where each case is ideally positioned in the truth table configurations, the algorithm proceeds with the same Boolean minimization procedure as in the crisp version in order to identify minimally sufficient configurations that are related to the presence (or absence) of an outcome.

It is beyond the purpose of this chapter to offer a complete presentation of the QCA procedure with all its details. There are entire books written for this purpose (Ragin, 2000, 2008; Rihoux and Ragin, 2009; Schneider and Wagemann, 2012; Duşa, 2019), and the interested reader is invited to consult the relevant literature. The main purpose was to reveal how the language of sets and the Boolean algebra can be employed for social and political research.

To conclude, set theoretic methods are rather young compared with the long-established quantitative tradition, but they

already compensate through a sound and precise mathematical procedure that uses set relations (subsets and supersets) to identify multiple conjunctural causation, where the outcome can be produced via several (sufficient) combinations of causal conditions.

Different to the strict statistical assumptions in the quantitative analysis, the causal conditions in QCA are not assumed to be independent of each other. What matters is how they conjunctively combine to form sufficient subsets of the outcome, and their relevance in terms of both coverage of the outcome and how well they explain the empirically observed configurations.

## REFERENCES

- Aiton, Eric J. 1985. *Leibniz. A Biography*. Bristol and Boston: Adam Hilger Ltd.
- Arfi, Badredine. 2010. *Linguistic Fuzzy Logic Methods in Social Sciences*. Vol. 253 of *Studies in Fuzziness and Soft Computing* Berlin and Heidelberg: Springer Verlag.
- Babst, Dean V. 1964. 'Elective Governments. A Force for Peace'. *The Wisconsin Sociologist* 3(1): 9–14.
- Barrenechea, Rodrigo and James Mahoney. 2017. 'A Set-Theoretic Approach to Bayesian Process Tracing'. *Sociological Methods & Research* 48(3): 451–484.
- Bauer, Paul C., Pablo Barberá, Kathrin Ackermann and Aaron Venetz. 2014. 'Vague Concepts in Survey Questions: A General Problem Illustrated with the Left-Right Scale'. *SSRN Electronic Journal*. Available at: <http://ssrn.com/abstract=2467595>.
- Cheli, Bruno and Achille Lemmi. 1995. 'A 'Totally' Fuzzy and Relative Approach to the Multidimensional Analysis of Poverty'. *Economic Notes* 1: 115–134.
- Cooper, Barry and Judith Glaesser. 2010. 'Contrasting Variable-Analytic and Case-Based Approaches to the Analysis of Survey Datasets: Exploring How Achievement Varies by Ability across Configurations of Social Class and Sex'. *Methodological Innovations* 5(1): 4–23.
- Dauben, Joseph Warren. 1979. *Georg Cantor. His Mathematics and Philosophy of the Infinite*. Princeton: Princeton University Press.

- Demey, Lorenz, Barteld Kooi and Joshua Sack. 2017. Logic and Probability. In *The Stanford Encyclopedia of Philosophy (Summer 2019 Edition)*, ed. Edward N. Zalta. Metaphysics Research Lab, Stanford University. Available at <https://plato.stanford.edu/archives/sum2019/entries/logic-probability/> (Accessed on 20 January 2020).
- Dubois, Didier and Henri Prade. 1989. 'Fuzzy Sets, Probability and Measurement'. *European Journal of Operational Research* 40: 135–154.
- Dul, Jan. 2016. 'Necessary Condition Analysis (NCA). Logic and Methodology of 'Necessary but Not Sufficient' Causality'. *Organizational Research Methods* 19(1): 10–52.
- Duşa, Adrian. 2019. Calibration. In *QCA with R. A Comprehensive Resource*, ed. Adrian Duşa. Cham: Springer International Publishing, pp. 61–98.
- Esping-Andersen, Gøsta. 1990. *The Three Worlds of Welfare Capitalism*. Princeton, New Jersey: Princeton University Press.
- Fairfield, Tasha and Andrew E. Charman. 2017. 'Explicit Bayesian Analysis for Process Tracing: Guidelines, Opportunities, and Caveats'. *Political Analysis* 25(3): 363–380.
- Fiss, Peer C., Dmitry Sharapov and Lasse Cronqvist. 2013. 'Opposites Attract? Opportunities and Challenges for Integrating Large-N QCA and Econometric Analysis'. *Political Research Quarterly* 66(1): 191–198.
- Fujita, Taisuke. 2009. 'Developed and Democratic Countries' Policy-making on Dispute Settlement in the GATT/WTO: Exploring Conjunctural and Multiple Causations by Comparing QCA and Regression Analysis'. *Sociological Theory and Methods* 24: 181–202.
- Goertz, Gary. 2006a. 'Assessing the Trivialness, Relevance, and Relative Importance of Necessary or Sufficient Conditions in Social Science'. *Studies in Comparative International Development* 41(2): 88–109.
- Goertz, Gary. 2006b. *Social Science Concepts. A User's Guide*. Princeton and Oxford: Princeton University Press.
- Goldstone, Jack A. 1997. 'Methodological issues in comparative macrosociology'. *Comparative Social Research* 16: 121–132.
- Goldthorpe, John H. 1997. 'Current issues in comparative macrosociology: a debate on methodological issues'. *Comparative Social Research* 16: 1–26.
- Grendstad, Gunnar. 2007. 'Casual Complexity and Party Preference'. *European Journal of Political Research* 46: 121–149.
- Grofman, Bernard and Carsten Q. Schneider. 2009. 'An Introduction to Crisp Set QCA, with a Comparison to Binary Logistic Regression'. *Political Research Quarterly* 62(4): 662–672.
- Halmos, Paul R. 1974. *Naive Set Theory*. New York: Springer.
- Heylen, Ben and Mike Nachtegaele. 2013. 'The Integration of Fuzzy Sets and Statistics: Toward Strict Falsification in the Social Sciences'. *Quality & Quantity* 47(6): 3185–3200.
- Hsieh, Hsin-I. 1980. 'Set Theory as a Meta-Language for Natural Languages'. *International Journal of Human Communication* 13(3): 529–542.
- Katz, Aaron, Matthias vom Hau and James Mahoney. 2005. 'Explaining the Great Reversal in Spanish America: Fuzzy-Set Analysis Versus Regression Analysis'. *Sociological Methods & Research* 33(4): 539–573.
- Kosko, Bart. 1990. 'Fuzziness vs. Probability'. *International Journal of General Systems* 17(1): 211–240.
- Kosko, Bart. 1994. 'The Probability Monopoly'. *IEEE Transactions on Fuzzy Systems* 2(1): 32–33.
- Lieberson, Stanley. 1991. 'Small N's and Big Conclusions: An Examination of the Reasoning in Comparative Studies Based on a Small Number of Cases'. *Social Forces* 70(2): 307–320.
- Łukasiewicz, Jan. 1970. On 3-Valued Logic. In *Jan Łukasiewicz: Selected Works*, ed. Leo Borkowski. Warsaw: Polish Scientific Publishers.
- Mahoney, James. 1980. 'Qualitative Methodology and Comparative Politics'. *Comparative Political Studies* 40(2): 122–144.
- Mahoney, James. 2010. 'After KKV: The New Methodology of Qualitative Research'. *World Politics* 62(1): 120–147.
- Mahoney, James. 2016. 'Mechanisms, Bayesianism, and process tracing'. *New Political Economy* 21(5): 493–499.
- Marx, Axel and Jadir Soares. 2015. 'Applying New Methodological Tools in Human Rights Research. The Case of Qualitative Comparative Analysis'. *The International Journal of Human Rights* 20(3): 365–385.
- McCluskey, E. J. 1956. 'Minimization of Boolean Functions'. *Bell System Technical Journal* 35(6): 1417–1444.

- Mendel, Jerry M. and Mohammad M. Korjani. 2018. 'A New Method for Calibrating the Fuzzy Sets Used in fsQCA'. *Information Sciences*. doi: 468. 10.1016/j.ins.2018.07.050.
- Mill, John Stuart. 1843. *A System of Logic, Ratiocinative and Inductive*. Vol. 1 London: John W. Parker, Harrison & Co.
- Munck, Gerardo L. 2016. 'Assessing Set-Theoretic Comparative Methods: A Tool for Qualitative Comparativists?' *Comparative Political Studies* 49: 1–6.
- Novák, Vilém. 1991. Fuzzy Set Theory and Modelling of Natural Language Semantics. In *Interactive Fuzzy Optimization*, eds. Mario Fedrizzi, Janusz Kacprzyk and Marc Roubens Vol. 368 of *Lecture Notes in Economics and Mathematical Systems*. Berlin Heidelberg: Springer Verlag.
- Paine, Jack. 2015. 'Set-Theoretic Comparative Methods: Less Distinctive Than Claimed'. *Comparative Political Studies* 49(6): 1–39.
- Pennings, Paul. 2003. 'Beyond Dichotomous Explanations: Explaining Constitutional Control of the Executive with Fuzzy-Sets'. *European Journal of Political Research* 42(4): 541–567.
- Pinter, Charles C. 2014. *A Book of Set Theory*. Mineola and New York: Dover Publications.
- Quine, W. V. 1955. 'A Way to Simplify Truth Functions'. *The American Mathematical Monthly* 62(9): 624–631.
- Radaelli, Claudio M. and Claudius Wagemann. 2019. 'What did I Leave Out? Omitted Variables in Regression and Qualitative Comparative Analysis'. *European Political Science* 18: 275–290.
- Ragin, Charles C. 1987. *The Comparative Method*. Berkeley and London: University of California Press.
- Ragin, Charles C. 2000. *Fuzzy-Set Social Science*. 1 ed. Chicago: University of Chicago Press.
- Ragin, Charles C. 2004. 'From Fuzzy Sets to Crisp Truth Tables'. COMPASS Working Paper. Available at: <http://www.compass.org/wpseries/Ragin2004.pdf>
- Ragin, Charles C. 2008. *Redesigning Social Inquiry. Fuzzy Sets And Beyond*. Chicago and London: University of Chicago Press.
- Ragin, Charles C. and Peer Fiss. 2017. *Intersectional Inequality: Race, Class, Test Scores, and Poverty*. Chicago: University of Chicago Press.
- Ramzipoor, Roxanna. 2014. 'Set Theory and Fuzzy Sets: Their Relationship to Natural Language. An Interview with George Lakoff'. *Qualitative & Multi-Method Research* 12(1): 9–14.
- Rihoux, Benoît and Charles C. Ragin. 2009. *Configurational Comparative Methods - Qualitative Comparative Analysis (QCA) and Related Techniques*. Vol. 51 of *Applied Social Research Methods*. Newbury Park: Sage.
- Schneider, Carsten and Claudius Wagemann. 2012. *Set-Theoretic Methods for the Social Sciences. A Guide to Qualitative Comparative Analysis*. Cambridge: Cambridge University Press.
- Schneider, Carsten Q. and Ingo Rohlfing. 2016. 'Case Studies Nested in Fuzzy-set QCA on Sufficiency: Formalizing Case Selection and Causal Inference'. *Sociological Methods and Research* 45(3): 536–568.
- Seawright, Jason. 2005. 'Qualitative Comparative Analysis vis-à-vis Regression'. *Studies in Comparative International Development* 40(1): 3–26.
- Singpurwalla, Nozer D. and Jane M. Booker. 2004. 'Membership Functions and Probability Measures of Fuzzy Sets'. *Journal of the American Statistical Association* 99(467): 867–889.
- Skaaning, Svend-Erik. 2007. 'Explaining Post-Communist Respect for Civil Liberty: A Multi-Methods Test'. *Journal of Business Research* 60(5): 493–500.
- Skocpol, Theda. 1979. *States and Social Revolutions: A Comparative Analysis of France, Russia and China*. Cambridge University Press.
- Verkuilen, Jay. 2005. 'Assigning Membership in a Fuzzy Set Analysis'. *Sociological Methods and Research* 33(4): 462–496.
- Woodside, Arch G. 2014. 'Embrace Perform Model: Complexity Theory, Contrarian Case Analysis, and Multiple Realities'. *Journal of Business Research* 67(12): 2495–2503.
- Zadeh, Lotfi A. 1965. 'Fuzzy Sets'. *Information and Control* 8: 338–353.
- Zadeh, Lotfi A. 1983. A Fuzzy-Set-Theoretic Approach to the Compositionality of Meaning: Propositions, Dispositions and Canonical Forms. In *Fuzzy Sets, Fuzzy Logic, and Fuzzy Systems: Selected Papers by Lotfi A. Zadeh*, eds. George J. Klir and Bo Yuan. Vol. 6 of *Advances in Fuzzy Systems: Application and Theory*. Hackensack: World Scientific Pub Co Inc, pp. 594–613.
- Zadeh, Lotfi A. 1995. 'Probability Theory and Fuzzy Logic are Complementary Rather Than Competitive'. *Technometrics* 37: 271–276.

# Mixed-Methods Designs

Imke Harbers and Matthew C. Ingram

## INTRODUCTION

Mixed-methods (MM) research designs are those that combine more than one methodological approach within the same study. MM designs are also commonly called multimethod (Seawright, 2016), hybrid (Palinkas et al., 2015; Schoonenboom and Johnson, 2017), or simply mixed designs (Johnson et al., 2007: 118). MM research has exploded in popularity and prestige since the mid-1990s. This explosion was driven in part by methodological debates in the social and behavioral sciences. Initially, these debates focused on the relative merit of different methodological traditions (e.g., quantitative versus qualitative), but a productive debate eventually emerged about the complementary and synergistic strengths of different traditions, and therefore on the value of leveraging different approaches in MM designs. Thus, the appeal of MM to researchers, funding bodies, and publishers is rooted in the belief that combining methods adds value by

enhancing the validity of conclusions in ways that would be impossible using any one method in isolation. The basic intuition is that studying complex social processes requires researchers to be prepared to use diverse elements from the full breadth of the methodological toolbox of the social sciences. From a practical perspective, MM designs also tend to provide realistic, pragmatic, and feasible approaches even for some of the most compelling and probing research questions, making MM designs attractive to a wide range of researchers, including graduate students and junior scholars.

In this chapter, we proceed as follows. First, we clarify what we understand as mixed-methods research (MMR). We focus on designs that combine quantitative and qualitative methods, though we recognize that it is possible to combine a wide variety of methods. Second, we document the rise in (and still rising) popularity of MMR. Third, we disaggregate different types of MMR designs. We do so by differentiating types of

research designs according to (1) the manner in which methods are combined, or the degree of integration, (2) the sequence in which they are combined, and (3) the analytic motivations for such combinations. We also discuss some of the challenges inherent in MMR and identify innovations and future directions for overcoming them.

## WHAT IS MIXED-METHODS RESEARCH?

There is a lively debate about what does and does not constitute MMR. The popularity of the term, and its appeal to a broad range of audiences, has contributed to a situation where many seek to claim the label. We take as our starting point the broad definition offered by the founding editors of the *Journal of Mixed Methods Research*, who define MMR as ‘research in which the investigator collects and analyzes data, integrates the findings, and draws inferences using both qualitative and quantitative approaches or methods in a single study or a program of inquiry’ (Tashakkori and Creswell, 2007: 4).<sup>1</sup> Such research stands in contrast to ‘mono-method’ or ‘purist’ designs that rely on a single methodological approach (Johnson and Onwuegbuzie, 2004). While MMR potentially spans the full epistemological spectrum from positivism to interpretivism, our emphasis throughout this chapter is on the more positivist expressions of MMR. That is, we focus on designs where the analytic goal is inference, either causal or descriptive, and where the purpose of the design is to ensure that methods are employed in a way that is rigorous, systematic, and replicable (see also Gerring, 2017: xxi). Following Brady and Collier (2004: 291), we use the term ‘inference’ here to mean ‘the process of using data to draw broader conclusions about concepts and hypotheses that are the focus of research’. Especially relevant to MMR is Brady and Collier’s notion

of ‘nested inference’, where researchers seek to explain social phenomena by drawing ‘on both data-set observations and causal-process observations, sometimes at different levels of analysis’ (Brady and Collier, 2004: 298). This is the kind of research that has been most visible in mainstream political science journals. Moreover, by limiting ourselves to studies where the goal is inference, we are able to circumvent debates about the compatibility of qualitative and quantitative research on ontological and epistemological grounds (see Howe, 1988, on this debate). Our point of departure is that quantitative and qualitative methods can be combined fruitfully, and that doing so offers researchers additional opportunities for answering their respective research questions. While our definition of MMR is broad, it is worth pointing out what it does not cover. Research designs that combine two different quantitative methods, or two different qualitative methods, are not considered instances of MMR under our definition, even though these are also sometimes labeled as such in the literature.<sup>2</sup>

At its best, MM designs allow researchers to ‘draw from the strengths and minimize the weaknesses’ of different methods in a single study (Johnson and Onwuegbuzie, 2004: 14–15). The relationship between quantitative and qualitative methods is then one of ‘mutual illumination’ (Sammons, 2010: 699–700). This speaks to two methodological commitments that are sometimes seen as the defining core of the broad tent that is MMR. The first is a high degree of methodological eclecticism, where researchers ‘are constantly looking for other methods to explore a research problem or answer a research question through a synergistic process’ (Teddlie and Tashakkori, 2010: 16). The underlying motivation is the belief that through the successful integration of methods, the whole becomes greater than the sum of its parts.<sup>3</sup> The second principle is the subscription to an ‘iterative, cyclical approach to research’ (Teddlie and Tashakkori, 2010:

17; also, Yom, 2015). Conceptualizing the research process as an iterative cycle allows MM researchers to alternate between inductive and deductive phases, and between exploratory and confirmatory phases, during the course of a research project. Rather than thinking about inquiry as a process with discrete starting and ending points, the iterative understanding is motivated by the belief that each phase should inform the next, and that there is not a single, unique point of departure for investigations, but rather there are multiple possible points of departure.

That said, it is worth noting that even the most enthusiastic proponents of MMR will readily accept that MMR is not a silver bullet for every research problem. From an analytic perspective, the integration of methods offers opportunities, but it also creates challenges that do not arise in monomethod work (Rohlfing and Starke 2013). For instance, Rohlfing (2008) examines MMR designs where cases are selected for in-depth study on the basis of a quantitative model and notes that if the statistical model in the quantitative stage is misspecified, then the contribution of the qualitative phase is unclear because the basis for case selection was flawed. Rather than each phase of the research cycle guiding and improving the next one, problems that arise in one phase may then carry forward and undercut subsequent phases. In such cases, subsequent phases add little value to the understanding of the phenomenon under study, and researchers may draw incorrect inferences. We return to some of the challenges inherent in MMR below.

While MMR researchers share a commitment to methodological eclecticism, it is important not to equate this with an ‘anything goes’ attitude. There is a lively debate among MMR scholars about methodology that acknowledges not only the analytic benefits of MMR but also potential pitfalls. As with monomethod designs, the value of MMR depends on the quality of execution. All else being equal, a strong monomethod study is better than a weak multimethod one.

That said, if both designs are strong and well-executed, then MM scholars would argue that the findings of the multimethod one are likely to be more informative and valid, since they are able to incorporate a broader range of evidence. They may also be able to speak to a broader audience if they succeed in engaging quantitative as well as qualitative scholars. Overall, our purpose in this chapter is not to argue that MMR is always better, but to provide an overview of the menu of options that exists for researchers interested in pursuing MMR and to highlight some of the choices researchers will have to make. Throughout, we aim to provide citations to the different strands of methodological scholarship we discuss to facilitate further reading.

## THE RISE OF MIXED-METHODS RESEARCH

MMR has exploded in popularity and prestige since the mid-1990s, driven in part by broader methodological debates in the social and behavioral sciences. Initially, these debates focused on the relative merit of the quantitative and qualitative traditions. For instance, a major driver in the social sciences was *Designing Social Inquiry* (King et al., 1994) – advocating that quantitative and qualitative designs share a single logic of inference – and the responses to it, including a prominent set of review articles (and a rejoinder) in the *American Political Science Review* (Caporaso, 1995; Collier, 1995; King et al., 1995; Laitin, 1995; Rogowski, 1995; Tarrow, 1995), a response that culminated with *Rethinking Social Inquiry* (Brady and Collier, 2004/2010) – advocating for different logics of inference for quantitative and qualitative research, though with rigorous standards in either approach for reaching those inferences. Beyond this quantitative/qualitative divide in the social sciences, a productive debate eventually focused on the complementary nature of different traditions,



and therefore on the value of leveraging the strengths of different approaches in MMR. Social and behavioral scientists advanced full, book-length treatments on mixing methods. Major efforts in this regard include textbooks by Tashakkori and Teddlie (2003, 2010) and Creswell and Plano Clark (2007, 2011). Both of these volumes are intended for audiences from across the social and behavioral sciences. In political science, books by Seawright (2016), Gerring (2017), Goertz (2017), and Weller and Barnes (2014) offer guidance for the design and implementation of MMR, and many other methodological texts incorporate at least a brief discussion of MMR (e.g., Cyr, 2019: 25–32; Neumayer and Plümper, 2017: 73–5). Handbooks focusing on subfields such as comparative policy studies (Biesenbender and Héritier, 2014) and violence studies (Thaler, 2019) now frequently include chapters specifically discussing MM designs and their role in the study of the respective field. Such publications clearly responded to the

demand among political scientists for methodological advice on mixing methods. Indeed, the use of MM designs has continued to increase over time, suggesting they are more than a passing fad and are here to stay.

We can document this rise with citations to individual publications, interest in related topics in online searches, the emergence of new specialized journals, organized communities of scholars, and in awards by major funding sources. Regarding citations, as of January 2019, the handbook by Abbas Tashakkori and Charles Teddlie (2003, 2010) had more than 11,000 citations, and together with two additional books by these authors on the topic (1998, and Teddlie and Tashakkori 2009) combined more than 22,000 citations.<sup>4</sup> A single book by Creswell and Plano Clark (2007/2011) tallied more than 25,000 citations – that is, this book alone (both the 1st and 2nd editions) is being cited more than 2,000 times a year over a span of more than 10 years. A single article that is prominent

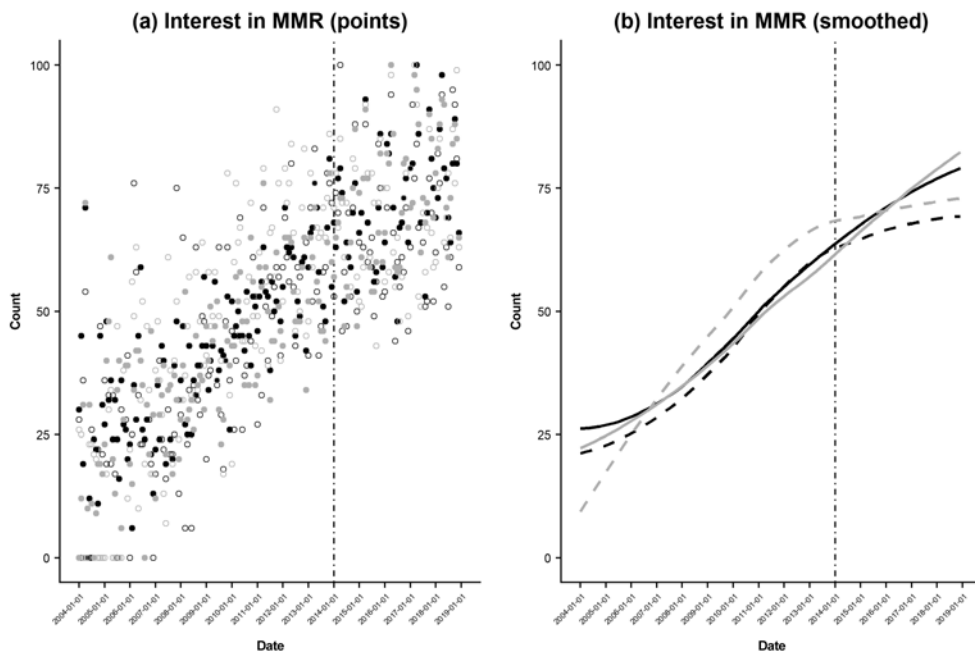


Figure 58.1

among political scientists (Lieberman, 2005), had more than 1,100 citations.

Interest in related topics in online searches also demonstrates the rise in MMR. Relying on Google Trends data (2019), we searched for two related terms: ‘mixed methods’ and ‘multimethodology’. Figure 58.1 shows the results of this search. In the monthly data from January 2004 to December 2018, each point represents the number of online searches for each term.

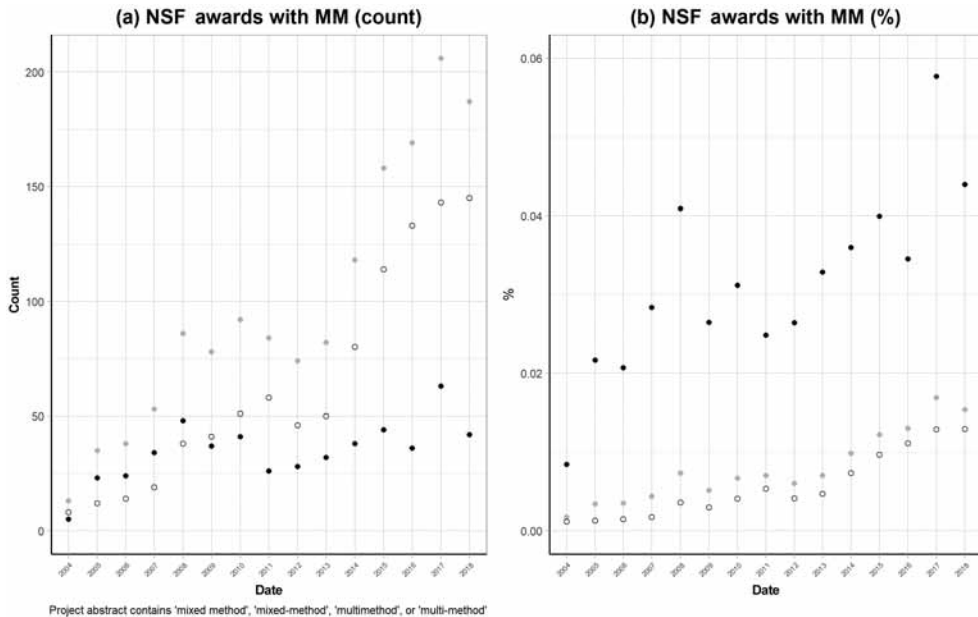
Panel (a) shows the raw data points, with filled circles representing worldwide data and empty circles representing the United States. Panel (b) shows smoothed trend lines: solid lines represent worldwide data and dashed lines represent the United States. In both panels, black represents the term ‘mixed methods’ and grey represents the term ‘multimethodology’; dates are on the x-axis and the number of searches is on the y-axis. Google Trends normalizes the variable on the y-axis so that 100 represents the maximum value in the time span examined for each search. The dashed vertical line in both panels represents January 2014.

Two major patterns emerge from Figure 58.1. First, interest in MMR has been increasing steadily over the last 15 years. This is evident in the upward, left-to-right sweep of the cloud of points in panel (a) and the trend lines in panel (b). Further, focusing on the vertical dashed line that marks January 2014 and looking at the upper-right corner of panel (a), all of the maximum values for each of the four measures appear in 2014 or later, that is, within the last few years. Second, interest in MMR appears to have risen faster in the United States and slowed in the last few years, starting around 2013–14, as shown by the flattening of the two dashed trend lines near the vertical dashed line. In contrast, interest in MMR worldwide keeps climbing; the solid trend lines retain a fairly steep slope, especially the gray one representing ‘multimethodology’.

Beyond individual publications or general interest, several journals now focus on,

or foreground, MMR (Creswell 2010: 61), including the *Journal of Mixed Methods Research* (JMMR). JMMR, founded in 2007, has had an impact factor of around 2.00 from 2010–16, and its impact factor rose to 3.27 in 2017 and 3.54 in 2018, with a 5-year impact factor of 3.28 in 2018 (Clarivate Analytics, 2018). This record made JMMR the top-ranked journal in both 2017 and 2018 among around 100 journals (98 in 2017 and 104 in 2018) in the category ‘Social Science – Interdisciplinary (SSCI)’. In the field of political science, the rising interest in MMR lead to the creation of a new section on ‘Qualitative and Multi-Method Research’ within the American Political Science Association. The section newsletter, which has been published since 2003, provides a platform for section members to showcase and discuss developments in MMR, and it enjoys a wide readership. In sum, both the number of outlets and the prominence of these outlets in their respective areas support concluding that both the popularity and prestige of MMR are increasing.

Lastly, data from the US National Science Foundation (NSF) strongly indicate that funding interest in MMR has been increasing over time. Specifically, Figure 58.2 reports data based on a search of 180,639 proposals that were funded by the NSF between 2004 and 2018. This time span is the same as the one for data from Google Trends above, facilitating a comparison of trends across Figure 58.1 and Figure 58.2. In the NSF data, we searched the abstract narrative of each award for key terms that showed the project included a mixed methods component. Specifically, we searched for any one of four terms (‘mixed methods’, ‘mixed-methods’, ‘multimethod’, or ‘multi-method’). Based on these criteria, we identified 1,479 projects that included a mixed methods component. We also distinguished projects in the social sciences from other programs. We did so by searching for the term ‘social’ in the ‘organization’ field of the data, which is the field that identifies the main NSF



**Figure 58.2**

unit in charge of reviewing and funding the proposal. For instance, this process identified all awards issued by the Directorate for Social, Behavioral, and Economic Sciences (SBE). Thus, we were able to classify MMR research across social and other sciences. Out of the 180,639 projects, 16,284 (9%) were in the social sciences. In Figure 58.2, panel (a) graphs the total count of projects by year within each disciplinary category, and panel (b) graphs the percentage of MMR projects within each disciplinary category. In both panels, gray dots identify projects across all disciplines, filled black dots identify projects in social sciences, and hollow black circles identify projects in disciplines other than social sciences.

Panel (a) shows that overall, the number of awards for MM projects in disciplines other than social sciences has been steadily trending upwards, and, since 2014–15, it has been two or three times as high as the number for the social sciences. However, prior to 2009, the number of awards for MMR projects in the social sciences was higher than

the number for MMR projects in other disciplines (except for 2004). This is somewhat striking because awards in the social sciences constitute only about 9% of all awards (see the earlier discussion of NSF data), so the number of awards in SBE relative to all other awards is proportionately small. Thus, the pattern in panel (a) is strong evidence that MMR designs were well-regarded in the social sciences before they began to grow in popularity in other disciplines.

This conclusion is supported by panel (b), which shows that the proportion of all awards that include an MMR component – a metric that is comparable across disciplines – is much higher for the social sciences than for other sciences. Specifically, the black filled dots show that about 2% to 4% of all NSF awards in the social sciences included an MM component from 2005 to 2016, and this increased to above 4% in the last two years of data (2017 and 2018). In contrast, the percentage in other disciplines was below 1% through 2015; the proportion of MMR projects in other disciplines has been rising,

but it has never risen above 1.5%. To sum up, MMR projects are increasingly successful in obtaining highly competitive NSF funding, and NSF awards in the social sciences are two to three times more likely to include an MM component than awards in other disciplines.

## TYPES OF MIXED-METHODS DESIGNS

MMR, as outlined above, is a broad church. In light of the range of designs labeled as MMR, it is helpful to highlight the nature of methodological choices different types of designs involve. There has been a lively debate about possible ways to classify MMR designs, and a number of scholars have offered typologies (Biesenbender and Héritier, 2014; Morse, 2010; Teddlie and Tashakkori, 2009). Gerring (2017: chapter 7) distinguished different types of MM designs based on whether the quantitative phase precedes the qualitative phase or vice versa. He also offers a range of research examples for each type (also Seawright and Gerring 2008). Even though the debate about various ways to classify MM designs and the respective advantages and drawbacks of different designs has been fruitful, the utility of such efforts for applied researchers and for teaching remains limited. Given the diversity of MM designs, it is impossible to provide a comprehensive and exhaustive menu of options from which researchers may simply select the appropriate one for their research problem based on a pre-defined set of criteria. Instead, ‘researchers using mixed methods are encouraged to continuously reexamine the results from one strand of the study compared to the results from another and to make changes both in the design and data collection procedures accordingly’ (Teddlie and Tashakkori, 2010: 23).

While proponents of MMR stress this ‘inclusive, pluralistic, and complementary’ approach to method selection as its core

selling point (Johnson and Onwuegbuzie, 2004: 17), it is also somewhat unsatisfying to researchers seeking to implement MM in their own work. If there is no authoritative overview of designs from which the right one may be selected, how are researchers supposed to know where to start? Rather than proposing an additional typology of MM designs, or discussing previous efforts to classify MM designs, we aim to highlight the nature of key methodological choices and the analytic tradeoffs involved in these choices. In this section, we identify three of the key dimensions along which MM designs can vary. These are: (1) the manner in which the methods are combined, that is, the degree of integration; (2) the sequence in which they are combined; and (3) the analytic motivations for such combinations.

First, by degree of integration, or the *manner* in which quantitative and qualitative methods are combined, we refer to the extent to which one method is embedded within the other, that is, the extent to which each method is analyzing the same data or information. At one extreme, there might be no integration at all. Following Domínguez and Hollstein (2014), we refer to these designs as ‘parallel’. Parallel designs execute each individual method separately and could even examine separate data or samples of data. The basic idea of parallel designs is to check whether different methods yield the same substantive conclusions, that is, whether they corroborate each other. If the results from a quantitative analysis are in line with those of a qualitative study of the same relationship or vice versa, then researchers can be more confident in their results.<sup>5</sup> At the other extreme, the combined methods might be highly integrated. Integrated designs – also called nested (Lieberman, 2005) or embedded (Hollstein, 2014) designs – ensure that each method is examining observations from the same sample of data, or that one method examines a subset of observations from the sample examined by the other method (Seawright, 2016).

Second, the *sequence* of methods refers to the temporal ordering of the methods. Some authors advocate starting with a qualitative study and then complementing that study with a quantitative one (Rohlfing, 2008), and others advocate reversing that sequence of methods (Lieberman, 2005). Note that the issue of sequencing is particularly relevant to integrated designs, where the quantitative and qualitative phase are supposed to build on each other. Sequence is less important for parallel designs. Parallel designs, where quantitative and qualitative phases are conducted independently of one another, are therefore also called ‘concurrent’ or ‘simultaneous’ designs (Morse, 1991; Teddlie and Tashakkori, 2010: 11). Within political science, the sequence from quantitative to qualitative has been particularly prominent, and there is a broad literature on how diagnostics from regression models can be leveraged for case selection. Indeed, a recent special issue on case selection emphasized that quantitative tools ‘should be enlisted, wherever possible, as the procedure is transparent and replicable, it limits opportunities for cherry-picking, and it enhances claims for representativeness’ (Elman et al., 2016: 383). Arguably, however, regression-based case selection is only possible where the state-of-the-art in the field allows researchers to build a robust model. Where less is known about the phenomenon, or where data availability restricts opportunities for quantitative analysis, it may be better to start with a case study. Ultimately, given the emphasis MMR places on an iterative conceptualization of the research cycle, the discussion about sequence should be seen in the context of a broader commitment to going back and forth between quantitative and qualitative phases. The exact starting point is therefore somewhat arbitrary and depends on previous knowledge – conceptual, theoretical, and empirical – and data availability.

The third dimension along which MM designs may vary are the *analytic motivations* of the project as a whole, and of mixing

methods in particular. As outlined above, we restrict the discussion in this chapter to projects where the overarching goal is inference. This includes descriptive inference, which ‘employs data to reach conclusions about what happened’, and causal inference, which ‘employs data to reach conclusions about why it happened’ (Brady and Collier 2004: 291). Under the broad label of inference, scholars pursue a broad range of analytic goals. The most obvious distinction might be the one between theory-testing and theory-building (George and Bennett, 2005; Lieberman, 2005). Equally important, however, and prominent among MM researchers, are concerns about concept formation and measurement (Adcock and Collier, 2001; Ahram, 2013). Moreover, MM designs are increasingly leveraged for policy evaluation (Bamberger et al., 2010).

The pre-eminent example of an MM design in political science is the nested design as introduced by Lieberman (2005). Nested designs combine quantitative and qualitative methods for the purpose of inference. The manner in which qualitative and quantitative methods are combined is integrated, and the design sequences the methods from quantitative to qualitative. In nested analysis, the combination of methods is highly systematic, explicitly conducting diagnostics on the quantitative phase of analysis in order to integrate the cases analyzed qualitatively within the same sample of observations that was examined quantitatively. The analytic motivation for this kind of design can be either theory-testing or theory-building. In nested analysis, the MM component is explicit, and the methods section of the eventual article or book would outline how the case selection was informed by the initial quantitative phase and post-estimation diagnostics.

Yet, the MM component of a broader research project may also be more implicit. Jensenius (2014) offers an insightful discussion about the fieldwork of quantitative data collection. While the goal of spending time

in the field is to collect data that will be analyzed with quantitative techniques, the quantitative phase of the research is then informed and refined by case knowledge. Qualitative techniques like (informal) interviews and participant observation may be indispensable to appreciate how large datasets were collected, and to evaluate data quality. Especially for data originally collected by public or private institutions, rather than professional survey organizations, understanding the origins of the data can be crucial to know how to spot irregularities, and to appreciate which inferences are and are not possible on the basis of the data. We believe that such ‘implicit’ MMR, where the eventual article or book only reports the results of the quantitative phase without discussing how the qualitative techniques contributed to the development of this analysis, are prominent across the social sciences. Note that the reverse sequence (i.e., from quantitative to qualitative) is probably also prominent. Researchers building their statistical model are encouraged not only to examine the overall model-fit, but also to examine outliers. Puzzling patterns of outliers may lead researchers to take a closer look at these cases, and to uncover what might be missing from the model. While such ‘lightweight case studies’ might not rise to the level of extensive, in-depth analysis that many associate with case study research, we would still argue that this constitutes an instance of MMR, even though this may remain implicit in the eventual presentation of results.

We illustrate the dimensions of MMR based on two examples of empirical research in political science: one book and one article. Both come from scholarship on violence in Latin America.

In her book on displacement in civil wars, Steele (2017) leverages subnational variation in patterns of displacement in Colombia. To understand when and where armed groups engage in ‘political cleansing’, that is, the strategic targeting of certain civilian groups for displacement by armed actors, she first

conducted in-depth fieldwork in a region affected by the violence. An initial round of archival research allowed her to link electoral results with locations of violence and thus to connect partisan preferences and patterns of displacement. The integration of different types of data from local archives constituted the basis for a test of the argument that armed groups target civilians when they are trying to conquer territory and have reason to believe that the local population supports a rival group. To test the argument beyond the region where it was developed, Steele then collected quantitative data on election results and displacement in other parts of the country. In terms of our dimensions, this constitutes an instance of an integrated design, where different phases of the research build on each other. The sequence is from qualitative to quantitative, though this is also an excellent example of the iterative nature of MMR. Quantitative data on key variables of interest informed case selection during the initial phase of the project, and insights gleaned there then informed further quantitative analysis. The analytic goal of the project was to explain patterns of displacement.

An article by Trejo and Ley (2018) on the origins of large-scale criminal violence in Mexico offers an illustration of how even an ambitious MMR design can be communicated within the space constraints of a journal article. The authors argue that political alternation in gubernatorial power eroded the informal networks of protection that organized crime groups relied on. To defend their turf, these groups invested in private militias, which they subsequently deployed to expand their territory and to fight rivals. The authors drew on spatial data of cartel murders to map the development of spatial patterns of violence. This reveals that violence is more pronounced in states that experienced alternation in gubernatorial power. To better understand networks of protection, the authors then conducted in-depth interviews with former governors of the states affected by the violence.

They submitted the implications of their argument about the timing and location of violence to quantitative tests and probed the mechanisms with case studies. Their project again provides an illustration of an integrated design, where quantitative and qualitative phases directly speak to each other. The primary sequence in the article is from quantitative to qualitative, but the hypothesized mechanisms are elaborated on the basis of interview data about informal networks. This, again, indicates how researchers may draw on quantitative and qualitative data to build more robust and convincing causal arguments.

### **OVERCOMING CHALLENGES INHERENT IN MMR: INNOVATIONS AND FUTURE DIRECTIONS**

As highlighted above, despite the promises and advantages of MMR, mixing methods also creates challenges for researchers beyond those inherent in monomethod work. In this section, we discuss some of these challenges, as well as ways in which researchers are addressing them methodologically and practically. The challenges we focus on here arise primarily from the increasing sophistication of methods, which place additional burdens on researchers aiming to make use of the full breadth of methodological innovation. The variety of chapters in this *Handbook* illustrates the wealth of methodological literature available to empirical scholars in political science and international relations, and a substantial share of this work is relevant to scholars seeking to conduct high-quality MMR.

New, more advanced, and specialized techniques are continuously emerging in the quantitative and qualitative traditions. In quantitative research, development has moved away from OLS, which has long been the workhorse of statistical analysis, to more complex types of modeling that can

take into consideration different types of dependent data structures (e.g., panel data, networks, spatial data, or complex systems) or that are geared towards causal identification with observational data (Angrist and Pischke, 2009; Dunning, 2012), as well as moving away from frequentist assumptions and towards Bayesian approaches. In qualitative research, new work has been particularly rich in the area of within-case analysis, including Bayesian process tracing (Bennett 2008; Bennett and Checkel, 2015; Fairfield and Charman, 2017; Rohlfing, 2012). There is also a lively debate about transparency in qualitative research, which raises the bar for communicating methodological choices and procedures inherent in qualitative research. The challenge for MM research is to keep up by offering guidance on how to build research designs that integrate the strengths of new (and old) techniques from each tradition.

First, from a practical perspective, given the complexity and sophistication of methodological approaches, individual researchers may find it challenging to be proficient in state-of-the-art approaches from more than one tradition. MMR compounds the challenge of keeping up with relevant methodological developments by asking scholars to stay on top of methods from different traditions. This requires effort and time, which necessarily involves compromises and opportunity costs and, depending on the availability of additional training, may also require funding, travel, and other investments and commitments, as well as integrating into new research communities. There are clear barriers for individual scholars who want to build expertise in quantitative and qualitative methods.

This has prompted researchers to work in teams, and to divide tasks according to methodological expertise. Collaboration can help resolve the challenge of becoming expert in two or more methods. Yet, collaboration also poses its own challenges. A good collaboration requires complementary interests and skills, but it also requires complementary

personal attributes, including mutual respect, balance of effort, communication, and a larger sense of partnership. Costs of collaboration can include ‘heightened transaction costs, shirking by a coauthor, monitoring differing levels of individual commitment and effort, duplication of effort, difficulty in coordinating tasks, and not receiving credit for individual performance’ (McDermott and Hatemi, 2010: 52–3). Moreover, if the collaboration is interdisciplinary, then there might be additional costs, including publishing in outlets valued in one discipline but not another, and the added effort of communicating ideas across different disciplinary vocabularies, norms, and audiences. Undoubtedly, these costs are also distributed unevenly among scholars based on several factors, including career stage, institutional profile, subfield, and global region. In addition, reporting findings and justifying methodological choices from multiple traditions is difficult within the space constraints of a journal article. Researchers may therefore struggle to explain methodological choices sufficiently while also leaving enough space to discuss the substantive importance of their findings. These practical challenges of conducting research and publishing results are generally more pronounced in MMR than in monomethod work.

Beyond such practical challenges, in this section we illustrate some of the methodological issues that arise for MMR from the increasing sophistication of methods. We do this by focusing on dependent data structures. This is a field of quantitative analysis that has expanded considerably, and the substantive importance of dependent data structures for understanding phenomena of interest to the social sciences is widely recognized. The standard regression assumption that units of analysis constitute independent observations that do not interact with each other is often hard to sustain and may even be antithetical to the ‘social’ part of social science research (Darmofal 2015). Yet, especially for integrated designs, ensuring that

dependencies are conceptualized and modeled consistently throughout the quantitative and qualitative phase of the research is often more challenging than it may appear at first glance. Even though qualitative research has long emphasized path dependence and how causal processes unfold over time, as well as issues like cross-unit interactions in processes of diffusion or contagion, integrating insights from both traditions systematically is often challenging.

Panel data, where dependencies may be temporal as well as cross-sectional, are one prominent example of an area where the field has advanced substantially. Case selection strategies based on the diagnostics of a quantitative model (Elman et al., 2016) may then not adequately identify promising cases in which to conduct in-depth analysis. For instance, if one were to follow Lieberman’s nested analysis, it is not clear that identifying a single well-predicted observation for model-testing analysis, or one poorly predicted observation for model-building analysis, makes sense when units are contributing multiple observations over time. Rather, researchers may want to identify the properties of observations over time from different units and select units for in-depth analysis based on the overall pattern or totality of properties (e.g., well-predicted, poorly predicted, influential) exhibited by all observations contributed over time by any one cross-sectional unit. There is limited methodological guidance for MM designs employing panel data during the quantitative analysis. Similar challenges confront MM designs with multi-level or hierarchical data. For illustrations of how researchers attempt to work through these challenges, see Giraudy (2015) and Ingram (2016a).

Furthermore, spatial data and the greater prominence of Geographic Information Systems (GIS) have prompted scholars to explicitly examine spatial dependence (Darmofal 2015). Especially in IR and conflict studies, there is widespread evidence that the processes underpinning phenomena



of interest often have a distinctly spatial component (Gleditsch and Weidmann, 2012). If a spatial process is at work, then selecting and studying one unit without examining its interactions with neighbors is insufficient. Spatial dependence therefore challenges MM researchers to examine their conceptualization of what constitutes a ‘case’. It may well be that, rather than a single unit, the in-depth analysis needs to consider the unit along with its neighbors to better understand the causal process. As we have argued elsewhere (Harbers and Ingram, 2017, Ingram and Harbers, Forthcoming), MMR has tremendous potential to enhance our understanding of spatial processes in political science. Nevertheless, the systematic integration of insights from the quantitative and qualitative phase is challenging, regardless of whether the quantitative phase precedes the qualitative phase or vice versa. For interdisciplinary perspectives and illustrations of how quantitative and qualitative approaches to spatial processes may be combined, see Matthews et al. (2005), Kwan and Ding (2008), Fielding and Cisneros-Puebla (2009), and Yeager and Steiger (2013).

Separately, there are novel approaches in MM network analysis (Domínguez and Hollstein 2014; Ingram 2016b). Given the analytic emphasis on the relational ties among units in network analysis, MM designs in network research include those that investigate how units themselves perceive or attach meaning to those ties. For instance, does the structure of relational ties exert an influence on units regardless of how units perceive those ties? Or, alternately, if a quantitative, structural analysis concludes that relational ties exert some influence, then how can units be selected to investigate the causal channels or mechanisms undergirding that influence? For example, in research examining the spread of legal ideas among a network of judges in Mexico, Ingram (2016c) identifies a network influence in a quantitative phase and then follows up by selecting individual judges based on their centrality in

the network. He conducts personal interviews with these judges to assess what influence or meaning they attribute to their network ties in understanding how these judges came to hold specific legal ideas.

There are also innovative multi-method approaches with Bayesian approaches. Humphreys and Jacobs (2015) provide one prominent and pioneering example. Specifically, they identify a key problem in MM research, namely, that in combining two or more methodological techniques, it is difficult to determine how much weight to attribute to inferences from one approach relative to the weight attributed to inferences from another approach. Their solution is a unified, Bayesian approach that ‘allows qualitative evidence to update the assumptions underlying quantitative analysis, and vice versa’ (Humphreys and Jacobs 2015: 654). They illustrate their approach with two examples from major research areas – origins of electoral systems and the causes of armed conflict. Further, they also provide guidance on how to use their approach to determine the optimal balance between quantitative and qualitative techniques in an MM design by estimating relative benefits expected from: (1) adding more cases (i.e., emphasizing the cross-sectional, correlational logic underlying the quantitative component); or (2) focusing more deeply on one or a small number of cases (i.e., emphasizing the within-case, process-based logic of the qualitative component).

Lastly, there are promising new developments in MM research that employ the growing toolkit of techniques associated with causal identification. For instance, Nielsen (2016) advocates using matching methods – which are increasingly used in quantitative analysis – as a novel tool for case selection in order to design ‘most similar’ systems for subsequent, in-depth qualitative analysis. These types of research designs are not new, but the use of matching software (and its many variations) and other tools to generate these designs is novel.

## CONCLUSION

As noted, MMR means different things to different people. We have clarified what MMR means – both in general and to us specifically, for the purposes of this chapter. We also documented the rise of MMR – including data on citations and interest from general internet audiences and funders. Further, we identified analytic goals that motivate MMR. Perhaps most importantly, we categorized major types of MM approaches according to the key decisions that researchers need to consider: (1) the manner in which methods will be combined (parallel or integrated); (2) the ways in which methods will be sequenced; and (3) the analytic motivations for such combinations. Lastly, we identified challenges researchers are likely to encounter in pursuing MMR, as well as some new, innovative approaches and promising future directions for additional work in this area.

We close by re-iterating two methodological commitments at the core of MMR: methodological eclecticism, and an iterative, cyclical understanding of the research process. As we emphasized earlier, these commitments do not mean that ‘anything goes’. Rather, good MMR research needs to be structured according to how diverse methods will be integrated, in what sequence, and for what analytic purpose. Good MMR is hard to do but, done well, the MMR whole promises insights larger than the sum of its parts.

## Notes

- 1 This is similar to the definition proposed by Johnson et al. (2007: 123): ‘Mixed methods research is the type of research in which a researcher or team of researchers combines elements of qualitative and quantitative research approaches (e.g., use of qualitative and quantitative viewpoints, data collection, analysis, inference techniques) for the broad purposes of breadth and depth of understanding and corroboration’.
- 2 This kind of data triangulation, where all methods employed come from the qualitative methodological tradition (e.g., interviews, archival analysis,

observation), is discussed by Tarrow (1995) and King et al. (1995: 479–80), and the latter express the possibility of a mixed-methods design that is purely qualitative and would meet the standards of inference they espouse. In addition to qualitative and quantitative methods, scholars may also mix methods by including components based on formal (e.g., Chapters 3, 7, 8, this *Handbook*), experimental, or design-based (e.g., Bowers and Leavitt, Chapter 41 and Sinclair, Chapter 52, this *Handbook*), as well as computational or machine-learning, approaches (see Olivella and Shoub, Chapter 56, this *Handbook*). For instance, the Empirical Implications of Theoretical Models (EITM; Chapters 3, 7, 8, this *Handbook*) tradition emphasizes a sequence from formal modeling to quantitative analysis. While we readily acknowledge the analytic value of such combinations, they are beyond the scope of our discussion in this chapter.

- 3 Creswell phrases a similar idea as the ‘sum of quantitative and qualitative is greater than either approach alone’ (Creswell, 2009: 104; citing Hall and Howard, 2008).
- 4 All citation counts are based on a Google Scholar search of these authors and titles on January 11, 2019.
- 5 For a critical discussion of this notion see Seawright (2016: 4–10). He argues that it is unlikely that the quantitative and qualitative study in fact address the same question. Rather, both methods ask and answer slightly different questions, so that reconciling the results of the two studies may be difficult, if not impossible.

## REFERENCES

- Adcock, R., and D. Collier. 2001. Measurement Validity: A Shared Standard for Qualitative and Quantitative Research. *American Political Science Review* 95: 529–546.
- Ahram, A. I. 2013. Concepts and Measurement in Multimethod Research. *Political Research Quarterly* 66(2): 280–291.
- Angrist, Joshua D., and J. Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist’s Companion*. Princeton: Princeton University Press.
- Bamberger, M., V. Rao, and M. Woolcock. 2010. Using Mixed Methods in Monitoring and Evaluation. In: Tashakkori, Abbas and Charles Teddlie, eds. *Sage Handbook of*

- Mixed Methods in Social & Behavioral Research*, 2nd ed. Thousand Oaks: Sage, pp. 613–641.
- Bennett, A. 2008. Process Tracing: A Bayesian Perspective. In: Janet M. Box-Steffensmeier, Henry E. Brady, and David Collier, eds. *The Oxford Handbook of Political Methodology*. Oxford: Oxford University Press, pp. 702–721.
- Bennett, A., and J. T. Checkel. 2015. *Process Tracing: From Metaphor to Analytic Tool*. New York: Cambridge University Press.
- Biesenbender, S., and A. Héritier. 2014. Mixed-Methods Designs in Comparative Public Policy Research: The Dismantling of Pension Policies. In: I. Engeli and C. Rothmayr eds. *Comparative Policy Studies: Conceptual and Methodological Challenges*. Basingstoke: Palgrave, pp. 237–264.
- Brady, H. E., and D., Collier eds. 2004. *Rethinking Social Inquiry: Diverse Tools, Shared Standards*. Lanham, MD: Rowman & Littlefield Publishers.
- Brady, H. E., and D., Collier eds. 2010. *Rethinking Social Inquiry: Diverse Tools, Shared Standards*. Lanham, MD: Rowman & Littlefield Publishers.
- Caporaso, J.. 1995. Research Design, Falsification, and the Qualitative-Quantitative Divide Reviewed Work(s): Designing Social Inquiry: Scientific Inference in Qualitative Research by Gary King, Robert O. Keohane and Sidney Verba. *American Political Science Review* 89(2): 457–460.
- Clarivate Analytics. 2018. InCites Journal Citation Reports. Accessed via Web of Science, University at Albany, January 11, 2019.
- Collier, D. 1995. Translating Quantitative Methods for Qualitative Researchers: The Case of Selection Bias. *American Political Science Review* 89(2): 461–466.
- Collier, D., H. E. Brady, and J. Seawright. 2004. Sources of Leverage in Causal Inference: Toward an Alternative View of Methodology. In: Henry E. Brady, David Collier, eds. *Rethinking Social Inquiry: Diverse Tools, Shared Standards..*, Lanham: Rowman and Littlefield, pp. 229–266.
- Creswell, J. W. 2009. Mapping the Field of Mixed Methods Research. *Journal of Mixed Methods Research* 3(2): 95–108.
- Creswell, J. W. 2010. Mapping the Developing Landscape of Mixed Methods Research. In: Tashakkori and Teddlie, eds. *SAGE Handbook of Mixed Methods Research*. Thousand Oaks: Sage, pp. 45–68.
- Creswell, J. W., and P. Clark, V. L. 2007. *Designing and Conducting Mixed Methods Research*. Thousand Oaks: Sage.
- Creswell, J. W., and P. Clark, V. L. 2011. *Designing and Conducting Mixed Methods Research*, 2nd ed. Thousand Oaks: Sage.
- Cyr, J. 2019. *Focus Groups for the Social Science Researcher*. New York: Cambridge University Press.
- Darmofal, D.. 2015. *Spatial Analysis for the Social Sciences*. Cambridge: Cambridge University Press.
- Domínguez, S., and B. Hollstein, eds. 2014. *Mixed Methods Social Networks Research. Design and Applications*, New York: Cambridge University Press.
- Dunning, T.. 2012. *Natural Experiments in the Social Sciences: A Design-Based Approach*. Cambridge: Cambridge University Press.
- Elman, C., J. Gerring, and J. Mahoney. 2016. Case Study Research: Putting the Quant Into the Qual. *Sociological Research & Methods* 45(3): 375–391.
- Fairfield, T., and Charman, A. 2017. Explicit Bayesian Analysis for Process Tracing: Guidelines, Opportunities, and Caveats. *Political Analysis*, 25(3), 363–380.
- Fielding, N., and Cisneros-Puebla, C. A. 2009. CAQDAS-GIS Convergence: Toward a New Integrated Mixed Method Research Practice? *Journal of Mixed Methods Research*, 3(4): 349–370.
- George, A. L., and A. Bennett. 2005. *Case Studies and Theory Development in the Social Sciences*. Cambridge: MIT Press.
- Gerring, J.. 2017. *Case Study Research: Principles and Practices*. 2nd ed. New York: Cambridge University Press.
- Giraudy, A.. 2015. *Democrats and Autocrats. Pathways of Subnational Undemocratic Regime Continuity within Democratic Countries*. Oxford: Oxford University Press.
- Gleditsch, K. S., and N. B. Weidmann. 2012. Richardson in the Information Age: Geographic Information Systems and Spatial Data in International Studies. *Annual Review of Political Science* 15: 461–481.
- Goertz, G. 2017. *Multimethod Research, Causal Mechanisms, and Case Studies: An*

- Integrated Approach*. Princeton: Princeton University Press.
- Google Trends. 2019. URL: <https://trends.google.com/trends> (accessed Jan. 11, 2019).
- Hall, B., and K. Howard 2008. A Synergistic Approach: Conducting Mixed Methods Research with Typological and Systemic Design Considerations. *Journal of Mixed Methods Research* 2(3): 248–269.
- Harbers, I., and M. C. Ingram. 2017. Geo-Nested Analysis: Mixed-Methods Research with Spatially Dependent Data. *Political Analysis* 25(3): 289–307.
- Hollstein, B. 2014. Mixed Methods Social Networks Research: An Introduction. In: S. Dominguez, and B. Hollstein eds. *Mixed Methods Social Networks Research. Design and Applications*. New York: Cambridge University Press, pp. 3–34.
- Howe, K. R. 1988. Against the Quantitative-Qualitative Incompatibility Thesis or Dogmas Die Hard. *Educational Researcher* 17(8): 10–16.
- Humphreys, M., and A. M. Jacobs. 2015. Mixing Methods: A Bayesian approach. *American Political Science Review* 109(4): 653–673
- Ingram, M. C. 2016a. *Crafting Courts in New Democracies: The Politics of Subnational Judicial Reform in Brazil and Mexico*. New York: Cambridge University Press.
- Ingram, M. C. 2016b. Mandates, Geography, and Networks: Explaining Subnational Diffusion of Criminal Procedure Reform in Mexico. *Latin American Politics and Society* 58(1): 121–145.
- Ingram, M. C. 2016c. Networked Justice: Judges, the Diffusion of Ideas, and Legal Reform Movements in Mexico. *Journal of Latin American Studies* 48(4): 739–768.
- Ingram, M. C., and I. Harbers. Forthcoming. Spatial Tools for Case Selection: Using LISA Statistics for Mixed-Methods Research Designs. *Political Science Research and Methods*, doi: <https://doi.org/10.1017/psrm.2019.3>.
- Jensenius, F. R. (2014) The Fieldwork of Quantitative Data Collection. *PS: Political Science & Politics* 47(2): 402–404.
- Johnson, R. B., and Onwuegbuzie, A. J. 2004. Mixed Methods Research: A Research Paradigm whose Time has Come. *Educational Researcher* 33(7): 14–26.
- King, G., R. O. Keohane, and S. Verba. 1994. *Designing social inquiry: Scientific inference in qualitative research*. Princeton: Princeton University Press.
- King, G., R. O. Keohane, and S. Verba. 1995. The Importance of Research Design in Political Science. *American Political Science Review* 89(2): 475–481.
- Kwan, M.-P., and G. Ding. 2008. Geo-Narrative: Extending GIS for Narrative Analysis in Qualitative and Mixed Method Research. *The Professional Geographer* 60(4): 1–30.
- Laitin, D. D. (1995) Disciplining Political Science. *American Political Science Review* 89(2): 454–456.
- Lieberman, E. S. 2005. Nested Analysis as a Mixed-Method Strategy for Comparative Research. *American Political Science Review* 99(3): 435–452.
- Matthews, S. A., J. E. Detwiler, and L. M. Burton. 2005. Geo-ethnography: Coupling Geographic Information Analysis Techniques with Ethnographic Methods in Urban Research. *Cartographica: The International Journal for Geographic Information and Geovisualization* 40(4): 75–90.
- McDermott, R., and P. K. Hatemi. 2010) Emerging Models of Collaboration in Political Science: Changes, Benefits, and Challenges. *PS: Political Science & Politics* 43(1): 49–58.
- Morse, J. 1991. Approaches to Qualitative-Quantitative Methodological Triangulation. *Nursing Research*, 40: 120–123.
- Morse, J. M. 2010. Procedures and Practice of Mixed Method Design: Maintaining Control, Rigor, and Complexity. In: Tashakkori, Abbas and Charles Teddlie, eds. *Sage Handbook of Mixed Methods in Social & Behavioral Research*, 2nd ed. Thousand Oaks: Sage, pp. 339–352.
- Nielsen, R. A. 2016. Case Selection via Matching. *Sociological Methods & Research* 45(3): 569–597.
- Neumayer, E., and T. Plümper. 2017. *Robustness Tests for Quantitative Research*. Cambridge: Cambridge University Press.
- Palinkas, L. A., S. M. Horwitz, C. A. Green, J. P. Wisdom, N. Duan, and K. Hoagwood. 2015. Purposeful Sampling for Qualitative Data Collection and Analysis in Mixed Method Implementation Research. *Administration and Policy in Mental Health* 42(5): 533–544.

- Rogowski, R. 1995. The Role of Theory and Anomaly in Social-Scientific Inference. *American Political Science Review* 89(2): 467–470.
- Rohlfing, I. 2008. What You See and What You Get: Pitfalls and Principles of Nested Analysis in Comparative Research. *Comparative Political Studies* 41(11): 1492–1514.
- Rohlfing, I. 2012. *Case Studies and Causal Inference: An Integrative Framework*. Basingstoke: Palgrave Macmillan.
- Rohlfing, I., and P. Starke. 2013. Building on Solid Ground: Robust Case Selection in Multi-Method Research. *Swiss Political Science Review* 19(4): 492–512.
- Sammons, P. 2010. The Contribution of Mixed Methods to Recent Research on Educational Effectiveness. In: Tashakkori, Abbas and Charles Teddlie, eds. *Sage Handbook of Mixed Methods in Social & Behavioral Research*, 2nd ed. Thousand Oaks: Sage, pp. 697–724.
- Schoonenboom, J. and R. B. Johnson. 2017. How to Construct a Mixed Methods Research Design. *Köln Zeitschrift für Soziologie und Sozialpsychologie* 69(2): 107–131.
- Seawright, J.. 2016. *Multi-Method Social Science: Combining Qualitative and Quantitative Tools*. Cambridge: Cambridge University Press.
- Seawright, J., and J. Gerring. 2008. Case Selection Techniques in Case Study Research: A Menu of Qualitative and Quantitative Options. *Political Research Quarterly* 61(2): 294–308.
- Steele, A. A. 2017. *Democracy and Displacement in Colombia's Civil War*. Ithaca: Cornell University Press.
- Tarrow, S. 1995. Bridging the Quantitative-Qualitative Divide in Political Science, *American Political Science Review*, 89(2): 471–474.
- Tashakkori, A., and C. Teddlie. 1998. *Mixed Methodology: Combining Qualitative and Quantitative Approaches*. Thousand Oaks: Sage.
- Tashakkori, A., and C. Teddlie. 2003. *Sage Handbook of Mixed Methods in Social & Behavioral Research*. Thousand Oaks: Sage.
- Tashakkori, A., and C. Teddlie. 2010. *Sage Handbook of Mixed Methods in Social & Behavioral Research*, 2nd ed. Thousand Oaks: Sage.
- Tashakkori, A., and J. W. Creswell 2007. Editorial: The New Era of Mixed Methods. *Journal of Mixed Methods Research* 1(1): 3–7.
- Teddlie, C., and A. Tashakkori. 2009. *Foundations of Mixed Methods Research: Integrating Quantitative and Qualitative Approaches in the Social and Behavioral Sciences*. Thousand Oaks: Sage.
- Teddlie, C., and A. Tashakkori. 2010. Overview of Contemporary Issues in Mixed Method Research. In: A. Tashakkori and C. Teddlie, eds. *Sage Handbook of Mixed Methods in Social & Behavioral Research*, 2nd ed. Thousand Oaks: Sage, pp. 1–44.
- Thaler, K. H. 2019. Mixed Methods in Violence Studies. In: W. S. DeKeseredy, C. M. Rennison, and A. K. Hall-Sanchez, eds. *The Routledge International Handbook of Violence Studies*. New York: Routledge, pp. 19–29.
- Trejo, G., and S. Ley. 2018. Why did Drug Cartels Go to War in Mexico? Subnational Party Alternation, the Breakdown of Criminal Protection, and the Onset of Large-Scale Violence. *Comparative Political Studies* 51(7): 900–937.
- Weller, N., and Barnes, J. 2014. *Finding Pathways: Mixed-Method Research for Studying Causal Mechanisms*. Cambridge: Cambridge University Press.
- Yeager, C. D., and Steiger, T. 2013. Applied Geography in a Digital Age: The Case for Mixed Methods. *Applied Geography* 39: 1–4.
- Yom, S.. 2015. From Methodology to Practice Inductive Iteration in Comparative Research. *Comparative Political Studies* 48(5): 616–644.

# Case Study Methods: Case Selection and Case Analysis

Chiara Ruffa

Case study methods are a group of approaches in political science and international relations that aim at testing and developing theory.<sup>1</sup> The key characteristic of case study methods is their focus on one or few cases but with the ambition to understand and capture broader and more general underlying dynamics. This implies two core inter-related functions: the first is describing an observed phenomenon in all its depth and complexity; the second is attempting to generalize to a broader universe of cases. A research project using case study methods can rarely attempt to do both at the same time and to the same extent: the attempt to go deep usually has to compromise with trying to describe and study several cases in breadth. Either way, case study methods imply a more or less explicit positivist assumption in terms of the objective of research: in different ways case studies aim at generalizing beyond the case or the set of cases at hand and they acknowledge an attempt to identify and describe patterns of behavior. Notwithstanding the positivist

assumption, the spectrum of positions can vary greatly: it ranges from explicit positivist approaches to interpretivist ones, in which the ability to observe causality is somewhat challenged (della Porta, 2008: 32; Vennesson, 2008). Case study methods' opportunities and challenges partly derive from being tools that can be used by scholars with different understandings of causality and different assumptions about what a researcher wants and can ostensibly know about the world.

Case study methods are a very well established set of methods in political science and international relations. In a 2011 TRIP (Teaching, Research and International Policy) survey, most scholars in international relations declared that their main chosen methods were qualitative.<sup>2</sup> As Bennett and Elman (2010: 499) have noted on several occasions, 'qualitative research methods are currently enjoying an almost unprecedented popularity and vitality in both the international relations and comparative politics subfields'. Notwithstanding their persistent popularity

and traction, the relevance of case study methods is being eroded by a growing focus on causal identification and inference in the positivist social sciences.<sup>3</sup> More narrowly, case study methods seem to fall short even towards observational quantitative approaches due to a general tendency to strive for generalization. Notwithstanding these concerns, case study methods still hold strong.

Case study methods scholars have developed a set of sophisticated techniques to make case study research rigorous and maximize the generalizability potential of case studies (Bennett and Elman, 2010). Partly because of this attempt to engage with other approaches, most of the debate on case studies has focused on case-selection techniques for generalizability purposes and has not engaged as much with case study analysis – that is, tools to conduct a case study and interpret the material. Yet, as Seawright and Gerring (2008: 294) argue, ‘case selection and case analysis are intertwined to a much greater extent in case study research than in large-N cross-case analysis’. The necessity to engage with other approaches has led to a somewhat essentialist understanding of case study methods and has over-shadowed the importance of using case-selection techniques in eclectic and creative ways and engaging more systematically with case analysis of complex phenomena. Case study methods should be seen as methods in their own right, with their own potentials and pitfalls. I contend that more attention should be paid to discuss case selection together with case analysis. Case study methods are both art and craft, and, as such, they should combine both command of the method and creative thinking to leverage inferential power through case selection and to gauge complexity through case analysis. In this chapter, I present the most common strategies of case selection and case analysis and outline ways to combine them more creatively for maximizing inferential power while at the same time capturing complexity.

This chapter comes in six parts. I first situate case study methods in the broader

literature on research methods. Second, I discuss case study methods in relation to what a case is and what it is good for; third, I present different logics of case selection; fourth, I discuss single case study approaches and fifth, I discuss comparative designs. Sixth, I reflect on common practices for conducting case studies and finally, I draw some conclusions.

### **BETWEEN A ROCK AND A HARD PLACE: CASE STUDY METHODS BETWEEN POST-POSITIVISM AND LARGE-N APPROACHES**

Despite being so well established, case study methods find themselves between ‘a rock and a hard place’. On the one hand, post-positivist and interpretivist approaches may work well with case studies but have different underlying ontological, epistemological and methodological assumptions. By post-positivism and interpretivism, I refer to a wealth of rather diverse ontologies and epistemologies that challenge the possibility to describe and explain phenomena in terms of causal relations and the impossibility to separate the object of research from the researcher.<sup>4</sup> The very use of the term ‘case study’, as in the ‘study of a case’, may reflect an underlying objective for generalizability, thereby making the term itself not widely used among post-positivists and interpretivists. Some scholars from post-positivist and interpretivist approaches have started to increasingly reflect upon and make use of case study methods, but they are mainly concerned with issues of generality and explicitly question generalizability as a more or less explicit research objective (Salter and Mutlu, 2013).

Among positivist scholarship, case study methods also hold a slightly uncomfortable place. The focus on a small number of cases (small-N) is what makes them distinct from the other, more widespread set of approaches in political science, that is, large-N approaches making use of statistical

techniques or experimental approaches. With the increasing traction of observational quantitative and experimental methods, case study methods are often under-appreciated, and their utility and breadth of application is widely misunderstood and misused. Within the field of political science and international relations, case studies are often used and seen in a subordinate and complementary fashion to quantitative techniques.<sup>5</sup> To its extreme, within the burgeoning literature using ever more advanced quantitative techniques, I witness a worrying and growing use of illustrative examples as substitute for systematic case study evidence.<sup>6</sup> While illustrative examples are useful to give an indication of the direction and quality of causal relation, they cannot substitute full-fledged case study analysis. Embracing and adopting case study methods may indeed not be a strategic choice: at first sight, time constraints and the pressure to publish more and more quickly do not provide great incentives to graduate students to specialize in case study methods, which are, by their nature, very time-consuming and field-work intensive, are narrated at best in book-format outputs and find themselves very constrained in article format. I still remember when I was desperately trying to fit my 25,000-word case narrative into a 13,000-word limit for the journal *Security Studies* – one of the most generous outlets in terms of word count (Ruffa, 2017). At the same time, engaging with quantitative scholars provides great opportunities. For instance, recent debates on qualitative methods have discussed how to make qualitative case studies more transparent, so as to foster dialogue with quantitative scholars through active citations, among others (Barnes and Weller, 2017; Moravcsik, 2010, 2014).

Notwithstanding these considerations, case study methods also provide unique opportunities on their own for understanding, conceptualizing, developing and testing new theories, and they remain a set of widely utilized methods. As Gerring (2009: 65) puts it, the case study approach ‘is disrespected

but nonetheless regularly employed. Indeed, it remains the workhorse of most disciplines and subfields in the social sciences’. Also, case study methods still hold an important place and perform analysis and cover aspects that quantitative techniques are unable to capture. In particular, they are unique at providing nuances in theories and observing phenomena from up-close, so that one does not need to capture and measure proxies. In other words, using case study methods allows for high validity – ‘the degree to which a measure accurately represents the concept that is supposed to measure’ (Kellstedt and Whitten, 2013: 126). Importantly, they still hold a lot of promise that is not being fully exploited or always given full attention.

## WHAT IS A CASE? WHAT IS IT NOT? AND WHAT IS IT FOR?

### *What is a Case?*

In this section, I define what a case is and is not and reflect upon what it is for. A case is usually defined as an instance of a broader phenomenon under study. George and Bennett (2005: 17) define ‘a case as an instance of a class of events’ and explicitly refer to the generalizability of it, since a case study is ‘the detailed examination of an aspect of a historical episode to develop or test historical explanations that may be generalizable to other events’. Levy (2008: 2) more precisely talks about a case as ‘an attempt to understand and interpret a spatially and temporally bounded set of events’. Occasionally used as synonym of observation in the qualitative realm, those two terms actually differ. As Gerring (2009: 20-21) puts it, in contrast to a case, ‘an *observation* is the most basic element of any empirical endeavor’, yet ‘in a case study, however, the case under study always provides more than one observation’. Those observations may be scattered across time and space within the same case.



Case studies focus on phenomena of specific interest, such as revolutions, wars, decisions and military interventions. The key here – albeit often neglected – is the importance of identifying a case at the smallest unit of analysis possible given the chosen theory. A case study allows to observe the theory at play. It is also important to distinguish between the case and the broader population of cases that the theory aspires to be applicable to. For instance, in his seminal work on the Cuban missile crisis, Allison's (1971) 'case' is the Cuban missile crisis. But this specific case could be seen as part of a large population of cases, which depending on the theory's focus could be coercive diplomacy, crisis management or the operational code of political leaders. When I first started to work on military cultures in peacekeeping operations, I was undecided on whether my cases were national troops deployed in peacekeeping missions or rather peacekeeping missions or, even more broadly, international military interventions (Ruffa, 2018). I have only recently realized that what I defined as 'my case' depended on the kind of literature and contribution I was trying to make – this being, for my project, security studies and the peacekeeping literature. My cases ended up being four different national troops deployed in two different missions. Since I made an argument about how military culture influences the ways in which soldiers behave when deployed in peacekeeping missions, it was reasonable to zoom in at the unit level, since each unit deployed to an area of operation and had full responsibility of implementing the UN mandate in that area – which was part of a peacekeeping mission. As such, case studies under analysis can be anything, ranging from countries to well defined time periods. In other words, some self-reflection on what a case is in a particular research project is key in order to clarify the nature of the contribution that is being made. Relatedly, it is important to reflect upon how high or low on the ladder of abstraction one wants to place her/his contribution (Sartori, 1970).

Once one has an idea about the objective of research – be it theory testing, development

or merely descriptive – a research design requires some assumptions that the objective of research is to identify patterns that might lead to generalizable results. This assumption aside, there is still widespread variation within positivist approaches, mainly in terms of the objective of research and in terms of how explicit the research objective is. The very use of the term 'case study' entails an underlying positivist assumption, in that the objective is to generalize and find patterns that are valid beyond the case under study. The logic underlying case study methods does not fundamentally differ from that of large-N studies: both case studies and large-N studies are based on observational data, meaning a non-experimental selection of the cases or observations under study. The core difference, however, between these two types of methods relates to the number of cases or observation under investigation. Case study methods vary between single case studies ( $n=1$ ) up to 12/15 cases, although the issue is still widely debated. Finally, case studies do not necessarily equal qualitative approaches. While in political science and international relations, case study methods usually rely on qualitative strategies of data collection, they can make use of numbers and statistics for measuring some aspects of the key variables in place (Collier, 2011: 824).

### ***What is a Case... Not?***

A case study is not an illustrative example. Also (but not only) in the quantitative literature, there is a growing usage of illustrative examples to suggest the plausibility of the existence of certain causal relations and the plausibility of underlying mechanisms. Illustrative examples are tremendously useful to suggest that some causal relations might be at play in the empirical domain; to complement a significant correlation with empirical examples suggesting the plausibility of the relations and of its underlying mechanism(s) and to guide the early phases of theorizing as

reality checks. While illustrations are increasingly required and used, their scope and purpose remain different from case studies. A recent paper by Haass and Ansorg (2018) is a good example in this respect. The authors argue that in peace operations with high-quality troops, militaries are better able to protect civilians. They further argue that such better protection happens because ‘high-quality militaries are better able to deter violence from state and non-state actors and create buffer zones within conflict areas, can better reach remote locations, and have superior capabilities to monitor the implementation of peace agreements’ (Haass and Ansorg, 2018: 1). While their paper is an important contribution to the peacekeeping literature, their use of an illustrative example on the case of Mali misses the broader picture that one would hopefully likely get in a full-fledged case study.

Despite a mission strength that was significantly lower at the time than that of MINUSCA (...) in a country about twice the size of the CAR, the UN operation successfully stabilized the situation in Mali and monitored the presidential elections in August 2013. MINUSMA was in a much better position to respond to threats against civilians than MINUSCA in part due to the fact that the mission consisted of, inter alia, highly trained troops from the Netherlands, Denmark, Norway and Finland. (Haass and Ansorg, 2018: 1)

The problem with that illustrative example is that it misses some important underlying dynamics happening within the case. Dutch, Norwegian and Finnish troops are unlikely to have contributed actively to all those mechanisms described because they were not even present on the ground, they deployed in very low numbers, and were only intelligence, surveillance and reconnaissance missions tasked to report back up to the Force Commander via the intelligence command structure. While Haas and Ansorg’s theory is sound and quantitatively supported, further qualitative research could potentially try to understand why we may find this result in a mission like Mali, for instance by exploring within case variations or trying to gauge qualitatively whether there was an indirect effect.

Along similar lines, deciding to focus on a full-fledged case study rather than on illustrative examples could be beneficial from the early phases of research. For instance, in a recent article, my coauthors and I argue that an increase in terrorist threats or attacks influence the level of military involvement in politics by either the military pushing its way in politics or being pulled (Bove, Rivera and Ruffa, 2019). We test this quantitatively and conduct three illustrative case studies: Algeria before the descent into the civil war; France in 1996–98 and 2015–16. The project started with the idea of only proposing some illustrative examples but it was only when we decided to turn them into full-fledged case studies and invested time on them that we discovered a second fundamental underlying causal mechanism. We did not only find a pushing effect of the military in politics but also of a pulling effect— which we then decided to further theorize about. If at all possible for time and length constraints, full-fledged case studies are usually preferable as they allow to capture dynamics that may remain invisible at a first cursory view.

### ***What is a Case For?***

A necessary preparatory step in delving into a case has to do with deciding what kind of research object one has in relation to its level of ambition. Different case-selection techniques would then follow depending on the kind of research objective one has. In their classic book, *Case study and theory development in the social sciences*, George and Bennett (2005) distinguish between six different kinds of case studies depending on the kind of research objective one has, which they adapted from the work of Eckstein (1975) and Lijphart (1971). They distinguish between atheoretical/configurative idiographic case studies, disciplined configurative, heuristic/theory development, theory testing, plausibility probes and building-block studies (George and Bennett, 2005: 75). Importantly, in a

single-outcome study, the phenomenon of interest is not seen as an instance of some greater population of cases. A somewhat similar approach is Beach and Pedersens' third form of process tracing and they describe it as a type of process tracing used to explain why a specific outcome occurred in a specific case and this is really the only occasion in which a type of process tracing does not contribute to advance theory (Beach and Pedersen, 2013).<sup>7</sup> While usually discarded, single-outcome studies can be good descriptions of cases that might be used in subsequent studies for theory building but by themselves do not cumulate or contribute directly to theory (George and Bennett, 2005: 75). They have no ambition beyond describing the phenomenon of interest and thereby get closer to historical work. In some instances, single case studies focus on phenomena that are intrinsically important. For instance, historian Isabel Hull (2005) wrote a book on German institutional extremism during the imperial period until 1918. Even though her research is extremely case-specific, she influenced much theory development within the literature on military and strategic cultures in the following years. In other cases, the potential population of cases appears too heterogeneous to allow for any general statements. For instance, Peter Feaver (2011) published a paper in which he examines the decision-making process that led to the surge in Iraq. In the civil-military literature, many argue that civil-military relations in the United States are so particular that they cannot be generalized beyond the case under study. Single-outcome studies are increasingly used in post-positivist approaches, for instance by focusing on the essence of politics or power shifts and changes. Aside from single-outcome studies, all other types of case study aim at either theory testing or theory development at a varying level of ambition. While theory testing has a deductive approach (from theory to empirics), theory development has an inductive spin (from empirics to theory). Disciplined configurative case studies make use of established theories for explaining a

case, while plausibility probes are theory-testing exercises with a lower level of ambition. Parallely, building-block studies are theory-development exercises with a narrower focus. Even when not used explicitly, these are useful for clarifying the objective of research given the constraints one has. Once the objective of research has been clarified, we can then move on to case-selection techniques.

## CASE-SELECTION TECHNIQUES

Case selection is a powerful tool in the hands of the researcher and the first crucial step for producing convincing research designs. Case selection should be 'an integral part of a good research strategy to achieve well-defined objectives of the study' (George and Bennett, 2005: 83). In case studies, the researcher 'is asked to perform a heroic role: to stand for (represent) a population of cases that is often much larger than the case itself' (Seawright and Gerring, 2008: 294). Developing on existing work, I distinguish among three different kinds of case selection: convenience, random and strategic. For the purpose of this discussion, I am not (yet) distinguishing between single case studies and comparative designs. The first kind of case selection entails selecting a case out of convenience. For instance, one selects a case because one knows the language in a particular country or because one is particularly interested in a case for its policy implications. Such a single case study approach might give some ideas about how the theory plausibly plays out in the case at hand, but it has not been strategically chosen to maximize the generalizability potential of the theory to a population of cases. This particular category of cases is problematic in terms of selection bias, as one might not select it based on its relation to a broader universe of cases.

A second strategy is randomly selecting a case or set of cases. While this is a possibility, it is highly discouraged. In fact, 'serious

problems are likely to develop if one chooses a very small sample in a completely random fashion' (Seawright and Gerring, 2008: 295). With two Monte Carlo experiments, Seawright and Gerring (2008) show that selecting cases randomly 'will often produce a sample that is substantially unrepresentative' (Seawright and Gerring, 2008: 295). Given the considerations above, the third strategy might in fact be the best one. The so-called strategic (or purposive) case selection is based on the idea of strategically selecting a case or pair of cases based on its hypothesized characteristics in relation to a broader universe of cases. Following that logic, one selects a case or compares two or more cases, trying to maximize the chances of capturing causal connections that are occurring within the universe of cases and controlling as much as possible for confounders. Selecting cases is a daunting task, and it is important to be pragmatic about it and combine convenience (such as the profound knowledge of a language or culture) with strategic considerations. A tip I always find useful is to start thinking about cases already when one is taking key decisions about theory. We often hold a romantic idea of first deciding on theory and then moving on to research design and case selection. Yet, this is rarely the case: we often circularly refine theories based on cases and decide on cases based on theory.<sup>8</sup> Doing so allows the research to buy time and reflect upon case selection at an early stage and reflect on tradeoffs among selection criteria, feasibility, previous case knowledge and the like. I discuss options of strategic case selection, first focusing only on one case and then, second, conducting comparative designs.

## **SINGLE CASE STUDIES**

Once the research objective has been clarified, it is important to reflect upon the different kinds of case study that are possible and are available, given the available constraints. A

first important distinction is between single case studies and comparative design. As the term suggests, single case studies focus on one single case, exploring the plausibility of a theory or tracing the causal mechanism at play in a particular context. In comparison to designs with more cases, single case studies have higher levels of conceptual validity, allowing us to take into account the complexity of contextual factors. By contrast, they are unable to control for confounders and may suffer from selection bias. Comparative methods (both in their most similar and most different system versions) allow us to control for confounders and have a greater external validity.

In line with the general discussion on case selection, single case studies can be selected by convenience or in strategic ways. In the first scenario, while there might be very good reasons for doing it, such cases may suffer from severe selection bias and their generalizability potential will be diminished. When selected strategically, however, single case studies hold the promise of combining the richness of focusing on one case with the ambition of saying something about the broader population of interest. Indeed, single case studies can be selected on the basis of their relation to theory, or, in quantitative language, on whether they are situated on the regression line or not (Gerring, 2007; Gerring, 2017)

For this reason, I label them 'empowered single case studies', and they are usually distinguished among four different kinds (Gerring, 2007; Gerring, 2017; Levy, 2008). Such distinction is mainly based on whether the values of dependent or independent variables are known or not. The first possibility is to select a so-called extreme case that has extreme or unusual values on the independent or the dependent variables. It is used for hypothesis generating or probing a potential causal connection but not to perform a full-fledged test of the theory. Such a case does not score very high in terms of representativeness because it does not give us any indication of how well the theory holds, since we do not yet know anything about the relationship

between independent and dependent variables. A second option is a deviant case, which is an outlier in comparison to all the cases that follow the theory, because – as the term suggests – it deviates from a cross-case relationship. A deviant case is useful for theory-development purposes or for generating new hypotheses, because it deviates from a cross-case relationship in a theory and might help detect neglected variables. For instance, in our work on NGO–military relations in complex humanitarian emergencies, Vennesson and I noticed that identity alone would not be able to explain variations in NGO–military relations (Ruffa and Vennesson, 2014). A hidden variable was at play, but we did not know what. Therefore, we selected a deviant case from existing theories, where we conducted qualitative empirical work and identified a neglected variable of interest – namely, domestic institutional configurations. A deviant case allowed us to identify a new variable, which was then ripe for further testing. A third option, which is in fact a variation of a deviant case, is an influential case, which displays influential configurations of the independent variable and makes me accept or refute the theory. It is suited for hypothesis testing as it helps me to make some final decisions about my theory. A fourth option is ‘a crucial case, which is most likely or least likely to exhibit a given outcome. It is based on a qualitative assessment of real crucial-ness’ (Gerring, 2007: 247). Crucial cases play an important confirming/confirmatory or disconfirming/disconfirmatory role in hypothesis testing. While they do not tell us much about representativeness, they are useful to get a sense of whether the theory holds even in a hard case (least likely) or does not hold in an easy case (most likely) and therefore whether it deserves further testing either quantitatively or qualitatively: ‘The inferential logic of least likely case design is based on the “Sinatra inference”—if I can make it there I can make it anywhere’ (Levy, 2002: 442), and by contrast, ‘the logic of most likely case design is based on the inverse

Sinatra inference—if I cannot make it there, I cannot make it anywhere’ (Levy, 2002: 442). A fifth option is the so-called pathway case, which is thought to embody a distinct causal path from the independent to the dependent variable. The pathway case is useful to probe a causal mechanism, and therefore a case that is easy to study at length can be particularly useful.<sup>9</sup> The pathway case should embody some typical relation as expected by the theory and is particularly fruitful in mixed-method analysis. Its main characteristic is that it lends itself to exploring causal mechanisms. So it shares from traits with the typical case that entails the selection of a case that mirrors the typical example of a cross-case relationship: the so-called typical or paradigmatic case. Such a case lends itself well for testing theories and how they hold, and they are representative of the broader population almost by definition. Pathway and typical cases partly overlap but a pathway case is not necessarily a typical case. Sometimes one could select a pathway case which is not typical because it is easy or practical to study.

Single case studies have both strengths and weaknesses. Because they allow for much emphasis on the case, they score high on conceptual validity, in the sense that nuanced and complex concepts can be constructed and measured. Relatedly, single case study can be useful for developing new hypotheses because they allow us to explore the causal mechanism and embark on process tracing (Bennett and Elman, 2007). The work by Katzenstein (1996), for instance, with a least-likely case, has produced some of the most influential recent scholarship in international relations. Recent work on a single case, the Democratic Republic of Congo, has changed the way we think about making peace at the local level (Autesserre, 2014).

On the other hand, single case studies also have some disadvantages. The first and main problem is the case-selection bias, that is, selecting a case that is already known to display certain characteristics. For instance, the

qualitative literature on peacekeeping has suffered from this particular problem, by focusing overwhelmingly on cases of failure rather than on cases of success (Howard, 2008; Pouligny, 2005). They are also not particularly suited for identifying scope conditions or necessity. Set aside typical cases, a single case study is unable to tell us whether the phenomenon we are observing is representative for the population of cases. In terms of inferential leverage, 'empowered single case studies' are more powerful than pure single case studies. Single case studies definitely allow for delving into a case study and its richness, which is great and very important. With some minor adjustments, single case studies can have an improved chance of generalizability. As usual with case study research, it is much about trade-offs, and it is difficult to provide a main takeaway. For sure, selecting a case based on convenience is weakened by the selection bias but also by the lack of objectivity. At the same time, a profound knowledge about the case is sometimes necessary to immerse oneself into a project. In practice, we often do both: we strive to follow a systematic case selection as well as deal with practical considerations. For instance, when selecting peace and stability operations to study, I was looking for a traditional UN peacekeeping mission but I also had to select cases of ongoing missions since I could only collect ethnographic evidence about soldiers deployed (Ruffa, 2018). Compounding case selection logics and practical considerations, I opted for the UN mission in Lebanon. Finally, single case studies have in any case a fundamental problem with confounders, and if one is particularly concerned by those, s/he should consider a comparative case study approach, which will be the focus of the next section.

## COMPARATIVE CASE STUDIES

Comparing cases entails two cases being selected and compared to one another. There

are two main known types of comparative designs: the most similar and the most different system design (Lijphart, 1971; Przeworski and Teune, 1970). In a most similar system case comparison, which builds on the logic of Mill's 'method of difference', two cases are selected on the basis of them presenting similar characteristics on most confounding variables, setting aside the independent or the dependent variables of interest (Mill, 1882: 484). In its purest form, such a case entails that through strategic case selection, the researcher manages to control for confounders. Those cases are selected because those variables do not entail much explanatory power. The best-known alternative to this is the most different system case comparison, which entails two cases being selected on the basis of them varying in every possible respect except the independent or dependent variable of interest. A classic example of this is the work by Theda Skocpol (2015). These two designs are mutually exclusive when thinking of one specific case comparison. However, when it comes to more complex designs, they can actually be fruitfully combined. For instance, in my own work, I combined a most similar system design with a most different system design. I conducted, separately, two distinct most similar system designs by comparing the French and Italian units in the UN mission in Lebanon and the French and Italian units in the NATO mission in Afghanistan. But I then compared Lebanon and Afghanistan, which were the most different types of peace and stability operations I could find that they had a similar outcome and independent variable of interest (Ruffa, 2017, 2018). While in that research the most different system design is notably much less developed, it is an attempt to creatively combine those two approaches. The most similar system design within each operation studied allowed me to gauge the plausibility of my theory by controlling for confounders. The most different system allowed to check whether my theory held in two very different context – thereby maximizing external validity. While usually

cases for comparative designs are selected strategically, recent work suggests the use of matching techniques to systematically select a most similar system design (Nielsen, 2016).

When thinking about comparative methods, a useful distinction is among variable-oriented vs. case-oriented designs. I opted for a variable-oriented comparative study, as opposed to a case oriented one. 'Variable-oriented studies mainly aim at establishing generalized relationships between variables, while case-oriented research seeks to understand complex units' (della Porta, 2008: 198). This differs from case-oriented that 'aims at rich descriptions of a few instances of a certain phenomenon' (della Porta, 2008: 198). While both approaches are legitimate, disagreements persist in terms of whether the underlying logics (generalizations vs. account for complexity) differ or are the same. Most research mirrors the tension between case vs. variable-oriented comparison, including my own. To illustrate, the foci of my analysis are rather complex units, which might have lent themselves better to a case-oriented approach. However, I opted for a variable-oriented approach for three reasons. First, notwithstanding the micro-focus, my ambition was to maximize the inferential strategy beyond the cases under study. Second, I decided to talk with the language of variables as a way to navigate through each case and being able to compare them, since I was using concepts and measurements that had never been used before. Third, I aimed to show that a comparative design with a strategic case selection actually ended up close to standard large-N techniques and fundamentally had a comparable inferential leverage. When actually conducting my comparative case study in the mission, I realized early on that the complexity of the issue at hand, the richness of the cases and the comparative logic necessitated a rigorous way of conducting research. Other alternatives to control for confounders are within case comparisons and longitudinal and spatial comparisons. Within case comparisons have the advantage that

most social, cultural and structural factors are probably somewhat similar. A similar logic pertains comparisons of the same unit in different moments of time (longitudinal) or in different space (spatial).

Conducting the comparative methods is a great opportunity to strive for greater generalizability and controlling for confounders while still maintaining some of the typical richness of small-N approaches. Comparative methods are useful both for developing new hypotheses and testing old ones. They also suffer, however, from two weaknesses. The first is that selection bias persists and, if anything, is worsened by the necessity of selecting good cases with certain characteristics to compare. The second is that they also run the risk of missing the independence among cases and of comparing apples and oranges.

Similar to large-N approaches, the comparative method takes competing explanations very seriously. While statistical methods assess rival explanations through statistical control, experimental methods eliminate rival explanations through experimental control and random sample selection of both the control and the treatment group. While experimental methods have the highest inferential power, statistical methods based on observational data have a similar understanding of the 'control' idea to the comparative method. While observational statistical methods assess rival explanations through statistical control, comparative methods do so strategically. Once a case or set of cases has been selected, the case study has to be conducted, which is the focus of the next section.

## **OPTIONS ON HOW TO DO A CASE STUDY: A FEW OPTIONS**

While the literature on case selection has burgeoned, relatively little is known about what to do when one has actually selected his/her case(s). The widespread assumption is that case selection and case analysis are two

distinct phases, but that is almost never true. Case study methods entail much more of a circular back-and-forth between case selection and case analysis than is usually recognized. This section briefly reviews three non-exhaustive strategies for conducting case studies: process tracing, structured focused comparison and congruence theory.<sup>10</sup> While process tracing and congruence do not require more than one case and in fact are particularly suited for in-depth one-case analysis, the method of structured focused comparison – as the term suggests – requires at least two cases. These three types of case study analysis entail specific strategies of data collection, ranging from individual qualitative interviews, observation, archival research, focus groups and document analysis, which fall outside the scope of this chapter (Kapiszewski et al., 2015; Ruffa, 2019). In the remainder of this section, I briefly present each approach and then provide a broad-stroke discussion on the advantages and disadvantages of each approach to case study analysis.

*Process tracing* is one of the most common techniques and entails ‘the systematic examination of diagnostic evidence selected and analyzed in light of research questions and hypotheses posed by the investigator’ (Collier, 2011: 823). Recent studies have introduced a similar method, called causal process observations (CPOs). A CPO systematically ‘highlights the contrast between a) the empirical foundation of qualitative research, and b) the data matrices analyzed by quantitative researchers, which may be called data-set observations (DSOs)’ (Collier, 2011: 823). Collier and others consider CPOs as equivalent to process tracing.<sup>11</sup> Process tracing is used both for descriptive and for explanatory purposes, and it is particularly well suited for within-case analysis (Beach and Pedersen, 2013; Checkel, 2006; George and Bennett, 2005). Since its first systematic and comprehensive treatment in George and Bennett (George and Bennett, 2005), different kinds of process tracing have emerged, mainly clustering around positivist

vs. more interpretivist research approaches (Beach and Pedersen, 2013). As Vennesson (2008: 232) points out ‘now the most common conceptions of process tracing are more standardized than the original formulation, and they emphasize the identification of a causal mechanism that connects independent and dependent variables. The emphasis is on causality, deduction and causal mechanisms’. In this particular, standardized version of process tracing, four empirical tests have been identified, originally formulated by Bennett and Elman (2010), who built on Van Evera (1997). Those four tests help understand how strongly a theory holds. Passing the test is necessary and/or sufficient for accepting the inference. When delving into the empirics and exploring the validity of the hypotheses, those four tests allow us to understand the strength of the hypothesis based on the empirical support found. The idea is that the hypothesis has to pass an increasingly difficult test; the harder the test it passes, the stronger the hypothesis is holding. If a hypothesis passes the straw-in-the-wind test, the hypothesis is relevant but not confirmed. If it fails, the hypothesis is not eliminated but weakened. If a hypothesis passes a hoop test, the relevance of the hypothesis is affirmed but not confirmed. However, if the hypothesis fails the hoop test, it suggests that the hypothesis has to be eliminated. When the hypothesis passes the smoking-gun test, it is confirmed, but if it fails, it is not eliminated but somewhat weakened. The fourth test is doubly decisive: if the hypothesis passes the test, it is confirmed and eliminates the others; if it fails, the hypothesis is eliminated (Collier, 2011).

This standardized idea of process tracing has gained traction in recent decades, which has implied that ‘although the idea of process tracing is often invoked by scholars as they examine qualitative data, too often this tool is neither well understood nor rigorously applied’ (Collier, 2011: 823). Sometimes scholars claim they are using process tracing when they are not. Alternatively, the



process-tracing method has become so technical that it can be hardly applied.

As Vennesson (2008: 232) points out, ‘something has been lost in the most recent formulations of process tracing’. A more pluralistic and less mechanistic understanding of process tracing could help capture more complex political phenomena – that can be linear, circular and interacting – than can hardly pass the four tests presented above. Finally, a particularly important yet somewhat less emphasized aspect is the sequencing aspect on process tracing, which gives close attention to phenomena as they unfold throughout time (Mahoney, 2010). In whatever form, however, tracing a process between two variables still differs from storytelling. As Flyvbjerg (2006: 237–41) suggests, process tracing differs from a pure narrative in three ways: process tracing is focused, structured and aims at providing a narrative explanation of a causal path that leads to a specific outcome. I contend that process tracing is a great framework but that it should be used pragmatically, otherwise there is a risk of missing the bigger picture. Relating to this, it is important that process tracing is not equated with testing causal mechanisms. As Gerring (2010: 1499) wrote, one should avoid ‘a dogmatic interpretation of the mechanistic mission’ – an explanation that is overly concerned with mechanisms.

A second common approach is less formal and entails systematically developing a set of observable implications to understand which kind of empirical referent one would need in order to find support for the theory. One particular version of it applies only to the comparative method, which is about conducting a *structured focused comparison*. Such comparison is structured because the same questions are asked across the cases, and it is focused because only the questions relevant to understanding the plausibility of the theory are asked, therefore focusing only on certain aspects (Flyvbjerg, 2006; George and Bennett, 2005). Such a method entails the ‘use of a well-defined set of theoretical

questions or propositions to structure an empirical inquiry on a particular analytically defined aspect of a set of events across different cases’ (George and Bennett, 2005). Structured focused comparisons are widely used in both research and teaching contexts, since they allow for stringent and rigorous comparison and at the same time provide a good toolkit on how to actually carry out a comparison (Friesendorf, 2018).

A third alternative to this is the method of congruence, which is often considered important to complement process tracing. Outcomes are congruent with the expectations of a particular theory (Blatter and Haverland, 2012). The method of congruence is insufficient when it comes to causal mechanisms and can be complemented with the consistency of observable implications of theory at different stages of the causal mechanism. While less technical than the previous two, congruence in whatever form is a pragmatic way to approach complex fieldwork terrain.

## CONCLUSION

The greater traction of quantitative methods has posed challenges to the role played by case study methods in political sciences and international relations. Even though case studies can be fruitfully used to complement quantitative techniques, they remain a method on their own, which maintains its specificities, advantages and disadvantages. The challenge for the future of case study methods is to strive for greater empirical rigor and greater transparency (Elman et al., 2010). With the current ongoing cross-fertilization unfolding across subfields (Gerring, 2017), further reflections on what case studies are and what they entail are important, as are the calls for more pluralism and pragmatism. This chapter has highlighted the importance of three dimensions to bear in mind in all phases of research, be it the setting up, the case selection or the conducting the case study phase. First, at the cost of

sounding obvious, it is important to clarify and be explicit about the kind of research objective one has and whether one is opting for the best design given the circumstances one is in. Second, within the case study methods, it is important to be well aware of the most important trade-offs and which design to opt for given the constraints. It is also important to be explicit about the inferential leverage that matters, given the time, budget and expertise constraints. But it ultimately boils down to asking what is the most fruitful way to answer one's research question given one's constraints? Third, case study research implies the fruitful combination of the craft and art of research in the social sciences. While it is always important to engage with a broader population of cases and master the techniques available for case selection and case analysis, it is also important to nurture creativity and 'let the case speak to you'. But, it has also to do with using case study methods to answer research questions by not overemphasizing mechanistic technicism of case-selection techniques but rather appreciating the unique opportunities provided by case study methods to answer complex research questions. The research objective should be to strive for greater transparency and maximize the inferential leverage so as to be able to dialogue with other approaches but, ultimately, to answer relevant research questions.

## Notes

- 1 I gratefully acknowledge Annekatri Deglow's excellent comments on this chapter.
- 2 <https://trip.wm.edu>.
- 3 Very crudely, by positivism I refer to a wealth of diverse approaches that have as ultimate objective of research to identify causal relations (by developing or testing theories; by conceptualizing or describing phenomena) and that believe that reality is distinct from the observer/researcher and knowable (even if only to some extent). For a more thorough discussion, see: Della Porta and Keating (2008: 19–39).
- 4 Della Porta and Keating distinguish between positivist, post-positivist, interpretivist and humanistic approaches in terms of ontology and

epistemology (Della Porta and Keating, 2008: 19–39). I also note that positivism and post-positivism lie on a continuum.

- 5 To the point that several scholars label only quantitative research as being 'empirical', notwithstanding the very strong and deep engagement of case study research with the 'empirics'.
- 6 Often quantitative scholars are asked to add illustrative examples when going through the review process. Ideally, the system could allow for more systematic mixed-methods approaches, where quantitatively focused and qualitatively focused could collaborate.
- 7 I thank Annekatri Deglow for pointing this out to me.
- 8 This process is often called 'abduction'.
- 9 For an alternative perspective to causal mechanisms, see process patterns (Friedrichs, 2016)
- 10 Causal process observations (CPOs) are often used but they will be treated here as synonymous to process tracing.
- 11 For an excellent example of CPOs, see Deglow (2018).

## REFERENCES

- Allison, G. T. (1971). *Essence of decision: explaining the Cuban missile crisis*. Boston, MA: Little, Brown and Company.
- Autesserre, S. (2014). *Peaceland: conflict resolution and the everyday politics of international intervention*. New York: Cambridge University Press.
- Barnes, J. & Weller, N. (2017). Case Studies and Analytic Transparency in Causal-Oriented Mixed-Methods Research. *PS: Political Science & Politics*, 50(4), 1019–1022.
- Beach, D. & Pedersen, R. B. (2013). *Process-tracing methods: foundations and guidelines*. Ann Arbor, MI: University of Michigan Press.
- Bennett, A. & Elman, C. (2007). Case Study Methods in the International Relations Subfield. *Comparative Political Studies*, 40(2), 170–195.
- Bennett, A. & Elman, C. (2010). Case Study Methods. I C. Reus-Smit & D. Snidal (Eds), *The Oxford handbook of international relations* (499–517). Oxford: Oxford University Press.
- Blatter, J. & Haverland, M. (2012). *Designing case studies: explanatory approaches in small-n research*. London: Palgrave MacMillan.

- Bove, Vincenzo, Mauricio Rivera, and Chiara Ruffa. 2019. 'Terrorist Violence and Nonviolent Military Involvement in Politics'. *European Journal of International Relations*, <https://doi.org/10.1177/1354066119866499>
- Checkel, J. T. (2006). Tracing Causal Mechanisms. *International Studies Review*, 8(2), 362–370.
- Collier, D. (2011). Understanding Process Tracing. *PS: Political Science and Politics*, 44(4), 823–830.
- Deglow, A. (2018). *Forces of destruction and construction: local conflict dynamics, institutional trust and postwar crime*. Uppsala: Uppsala University Press.
- della Porta, D. (2008). Comparative Analysis: Case-oriented versus Variable-oriented Research. I D. della Porta & M. Keating (Eds), *Approaches and methodologies in the social sciences: a pluralist perspective* (198–222). Cambridge: Cambridge University Press.
- Eckstein, H. (1975). Case Studies and Theory in Political Science. I F. I. Greenstein & N. W. Polsby (Eds), *Handbook of political science. Vol. 7 of political science: Scope and theory* (s. 79–138). Reading, MA: Addison-Wesley.
- Elman, C., Kapiszewski, D. & Vinuela, L. (2010). Qualitative Data Archiving: Rewards and Challenges. *PS: Political Science & Politics*, 43(1), 23–27.
- Feaver, P. D. (2011). The Right to Be Right: Civil-Military Relations and the Iraq Surge Decision. *International Security*, 35(4), 87–125.
- Flyvbjerg, B. (2006). Five Misunderstandings About Case-Study Research. *Qualitative Inquiry*, 12(2), 219–245.
- Friedrichs, J. (2016). Causal Mechanisms and Process Patterns in International Relations: Thinking Within and Without the box. *St Antony's International Review*, 12(1), 76–89.
- Friesendorf, C. (2018). *How Western soldiers fight: organizational routines in multinational missions*. Cambridge: Cambridge University Press.
- George, A. L. & Bennett, A. (2005). *Case studies and theory development in the social sciences*. Cambridge, MA; London: MIT Press.
- Gerring, J. (2007). *Case study research: principles and practices*. Cambridge: Cambridge University Press.
- Gerring, J. (2009). Case Study: What it is and What it Does. I R. E. Goodin (Ed.), *The Oxford handbook of political science*. Oxford: Oxford University Press, DOI: 10.1093/oxfordhb/9780199566020.003.0004
- Gerring, J. (2010). Causal Mechanisms: Yes, But.... *Comparative Political Studies*, 43(11), 1499–1526.
- Gerring, J. (2017). Qualitative Methods. *Annual Review of Political Science*, 20(1), 15–36.
- Haass, F. & Ansorg, N. (2018). Better Peacekeepers, Better Protection? Troop Quality of United Nations Peace Operations and Violence against Civilians. *Journal of Peace Research*, 55(6), 742–758.
- Howard, L. M. (2008). *UN peacekeeping in civil wars*. Cambridge: Cambridge University Press.
- Hull, I. V. (2005). *Absolute destruction: military culture and the practices of war in Imperial Germany*. Ithaca, NY and London: Cornell University Press.
- Kapiszewski, D., MacLean, L. M. & Read, B. L. (2015). *Field research in political science: practices and principles*, Cambridge: Cambridge University Press, chapters 5–10. [available in the library]. Cambridge, MA: Cambridge University Press.
- Katzenstein, P. J. (1996). *Cultural norms and national security*. Ithaca, NY: Cornell University Press.
- Kellstedt, P. M. & Whitten, G. D. (2013). *The fundamentals of political science research*. Cambridge: Cambridge University Press.
- Levy, J. S. (2002). Qualitative methods in international relations. In Brecher, M., and F. P. Harvey (Eds), *Millennial reflections on international studies*, (432–454). Ann Arbor: University of Michigan Press.
- Levy, J. S. (2008). Case Studies: Types, Designs, and Logics of Inference. *Conflict Management and Peace Science*, 25(1), 1–18.
- Lijphart, A. (1971). Comparative Politics and the Comparative Method. *American Political Science Review*, 65(03), 682–693.
- Mahoney, J. (2010). After KKV: The New Methodology of Qualitative Research. *World Politics*, 62(1), 120–147.
- Mill, J. S. (1882). *A system of logic, ratiocinative and inductive, being a connected view of the principles of evidence, and the methods of scientific investigation* (8<sup>th</sup> edition) New York: Harper & Brothers, publishers.
- Moravcsik, A. (2010). Active Citation: A Precondition for Replicable Qualitative Research. *PS: Political Science and Politics*, 43(1), 29–35.

- Moravcsik, A. (2014). Trust, but Verify: The Transparency Revolution and Qualitative International Relations. *Security Studies*, 23(4), 663–688.
- Nielsen, R. A. (2016). Case Selection via Matching. *Sociological Methods & Research*, 45(3), 569–597.
- Poulligny, B. (2005). Civil Society and Post-Conflict Peacebuilding: Ambiguities of International Programmes Aimed at Building 'New' Societies. *Security Dialogue*, 36(4), 495–510.
- Przeworski, A. & Teune, H. (1970). *The logic of comparative social inquiry*. New York: Wiley.
- Ruffa, C. (2017). Military Cultures and Force Employment in Peace Operations. *Security Studies*, 26(3), 391–422.
- Ruffa, C. (2018). *Military cultures in peace and stability operations*. Philadelphia, PA: University of Pennsylvania Press.
- Ruffa, C. (2019). 'The comparative method' in *SAGE Research Methods Cases*, <https://doi.org/10.1080/13533312.2019.1596742>.
- Ruffa, C. & Vennesson, P. (2014). Fighting and Helping? A Historical-Institutionalist Explanation of NGO-Military Relations. *Security Studies*, 23(3), 582–621.
- Salter, M. B. & Mutlu, C. E. (2013). *Research methods in critical security studies: an introduction*. London: Routledge.
- Sartori, G. (1970). Concept Misformation in Comparative Politics. *The American Political Science Review*, 64(4), 763–772.
- Seawright, J. & Gerring, J. (2008). Case Selection Techniques in Case Study Research: A Menu of Qualitative and Quantitative Options. *Political Research Quarterly*, 61(2), 294–308.
- Skocpol, T. (2015). *States and social revolutions*. Cambridge: Cambridge University Press.
- Stephen Van Evera, S. V. (1997). *Guide to methods for students of political science*. Ithaca, NY: Cornell University Press.
- Vennesson, P. (2008). *Case study and process tracing: theories and practices* (D. della Porta & M. Keating, Eds). Cambridge: Cambridge University Press.

# Comparative Analyses of Foreign Policy

Klaus Brummer

## INTRODUCTION

What makes Foreign Policy Analysis (FPA) a distinct sub-field of International Relations (IR) is the deliberate focus on the 'black box state'. Subscribing to a liberal understanding of international politics, scholarship in this tradition ties a country's external behavior back to its decision makers and/or domestic political structures (Snyder et al., [1954] 2002). Thus, on a *conceptual* level, FPA specifies and 'fine-tunes' Kenneth Waltz's (1959) rather general and abstract 'first image' and 'second image' explanations of international politics by offering a more nuanced and detailed understanding as to what exactly it is that drives behavior on the level of individual actors and the state level respectively. The ensuing greater number of levels of analysis as proposed by FPA represents one promising starting point for comparative research in this particular sub-field of IR since comparisons can be made across a broad set of substantive dimensions.

Another difference between FPA and, particularly, systemic IR theories is *empirical* in nature. Contrary to grand IR theories' tendency to focus on great powers, as most prominently showcased by neo-/realism, FPA is much less centered on explaining the behavior of that particular, and numerically rather small, group of states. Rather, FPA scholarship devotes much more attention to the decisions and actions of 'middle powers' and 'small states'. While entailing the risk of restricting the scope of empirical findings to specific types of states (which, of course, is also the case for theories that focus on great powers only), this broader perspective leads to a marked increase in the number of possible objects of investigation. In so doing, it represents another promising starting point for comparative analysis since the universe of cases is significantly expanded.

Hence, FPA and comparative research go together particularly well. It comes as no surprise, then, that a sub-field called 'Comparative Foreign Policy' exists within

FPA. Yet, efforts in this direction, which were strongly influenced by the behavioralist turn in IR and Political Science more broadly, did not produce the aspired results in terms of identifying ‘a grand unified theory of all foreign policy behavior for all nations for all time’ (Hudson, 2005: 9). After having fallen into disrepute for several decades, a renewed interest in comparative analyses of foreign policy has materialized more recently.

Rather than (unrealistically) aiming for the aforementioned ‘grand unified theory’, the current trend toward ‘back to comparison’ (Kaarbo, 2003) is driven more by middle-range theorizing, knowledge accumulation, and the aspiration to produce policy-relevant research. As Kaarbo (2003: 157) notes, ‘If a researcher wants to investigate many of the traditional factors that explain foreign policy... it is necessary to compare foreign policies across time, space, and issues to understand the general explanatory power of these various influences on governments’ behavior’. That scholars have heeded this call for comparative analyses of foreign policy (though hardly in the tradition of the grand unified theory-building efforts associated with ‘Comparative Foreign Policy’) is illustrated by the fact that around 75% of the articles published in the field’s main journal, *Foreign Policy Analysis*, in the year 2018 featured a comparative angle (esp. across time and/or space).

Building on this renewed interest in comparative analyses of foreign policy, the remainder of this chapter proceeds in four steps: the next section shows how FPA scholarship has specified possible drivers for international politics located on either the level of individual decision makers or the domestic-political level. It is within and across those levels of analysis that a comparative approach to the study of foreign policy can be fruitfully employed. This is followed by a brief overview of landmark comparative studies in FPA along with the presentation of more recent calls to recover and reinvigorate comparative analyses of foreign policy. The final section weaves the different strands of the argument

together and discusses two methods that have become increasingly popular when engaging in comparative analyses of foreign policy. On the level of individual decision makers, it shows how automated content analysis has made huge inroads over the last two decades with respect to the at-a-distance assessment of political leaders, thereby rendering possible meaningful comparisons between leaders within and across countries. On the state level of analysis, the chapter shows how Qualitative Comparative Analysis (QCA) has been fruitfully employed for establishing necessary and/or sufficient domestic-political drivers of foreign policy. The concluding section briefly summarizes the discussion and offers suggestions for future advances in comparative research in FPA.

## **FPA’S MULTI-FACETTED LEVELS OF ANALYSIS**

FPA scholarship does not dispute the importance of systemic factors in explanations of international politics. As Juliet Kaarbo (2015: 208) argues, ‘the international is not by definition exogenous to an FPA approach’. At the same time, FPA does not privilege such factors in its explanations, since many of them – for instance, the distribution of capabilities across the units of the international system and, even more so, the ordering principle of the international system (Waltz, 1979) – are very abstract and also quite stable over time. Therefore, they are ill-equipped to explain the specifics of foreign policy decision making processes or rapid changes in a country’s foreign policy behavior, to give but two examples. Moreover, while a country’s relative material capabilities suggest, or foreclose for that matter, certain types of behavior, states do not always engage in that (allegedly optimal) behavior, as, for instance, the literature on ‘underbalancing’ (Schweller, 2004) has shown. Overall, then, systemic IR theories cannot and, from the viewpoint of on its

key proponents, also should not be used to 'explain the particular policies of states', such as 'why state X made a certain move last Tuesday' (Waltz, 1979: 121).

This is where FPA comes in. This sub-field of IR is not only interested in explaining particular foreign policy processes, decisions, and behaviors of states but has also developed the analytical tools to accomplish those goals. For that purpose, FPA has opened up the 'black box state'. In doing so, it has arrived at more nuanced understandings of the drivers of policy that are associated with leaders and domestic institutions respectively. It is those factors that are privileged by FPA, rather than systemic/international ones, which is why FPA could, at minimum, represent a 'complement' to IR theories and maybe even an outright 'competitor' (Kaarbo, 2015: 208–9).

### ***Specifying the First Image***

A key dictum of much of FPA scholarship is 'who leads matters' (Walker, 2003: 245). Of course, this is hardly a novel argument per se as it, for instance, merely reiterates the general insight offered by Waltz (1959: 16), according to whom 'the nature and behavior of man' is one possible driver of international politics (here: cause for war). However, Waltz did not fully explicate what this 'first image' actually entails. That is, he refrained from systematically elaborating on both the specific characteristics of leaders that might actually guide their decisions and the situations in which individual leaders may be more (or less, for that matter) influential in the making of their respective country's foreign policies and thus in international politics more broadly. Those two gaps have been addressed by FPA scholarship.

A number of situations or contexts have been identified in which decision makers should have the most impact on, and thus make the most difference for, their country's foreign policy. Following Holsti (1976: 30), individual decision makers may matter the

most under several specific circumstances. Those include: 'non-routine situations' or 'unanticipated events' in which established procedures cannot be applied and cognitive mechanisms are likely to kick in; 'highly ambiguous' decision contexts which allow for a multitude of interpretations; or situations in which too much information is available ('information overload') so that certain coping mechanisms are required. More recently, though, Robert Jervis (2013) has argued, with a particular emphasis on US post-Cold War foreign policy, that individual decisions makers might actually matter less than commonly expected. Hence, there is no agreement about the contexts in which – let alone the specific extent to which – decision makers matter: 'The question of the extent to which leaders matter in international politics is as familiar as it is impossible to fully answer' (Jervis, 2013: 154). Still, decades of FPA scholarship do suggest that leaders can make a difference at certain points in time.

Against this background, FPA scholarship has identified several distinct characteristics of decision makers that are supposed to impact the making and outcome of their foreign policies and translated them into a number of analytical approaches. Those approaches are united by a set of underlying assumptions. First, challenging strict rationality assumptions as contained in many mainstream IR theories, decision makers are supposed to behave in a 'boundedly rational' way (Simon, 1957). That is, while they seek to engage in objectively rational behavior, limitations in humans' information processing and problem-solving capacities put limits on accomplishing this goal. To compensate for those limitations, decision makers 'construct a simplified model of the real situation in order to deal with it. [They] behave [] rationally with respect to this model, and such behavior is not even approximately optimal with respect to the real world' (Simon, 1957: 199). Second, the individual characteristics and idiosyncrasies of leaders are held to be key to understand how they construct their

simplified models of the world and thus ultimately their foreign policy choices. Finally, decision makers are supposed to differ with respect to those characteristics in the first place. Based on that assumption, it is possible to explain variation even in the foreign policies of countries with similar material capabilities or domestic institutions. So, it is not only that individuals matter *per se* but, more often than not, it is the specific individuals with their distinct characteristics that are crucial for the explanation of foreign policy.

This begs the question as to what exactly it is about individuals that matters. Drawing on insights from a number of other disciplines, such as social psychology and behavioral economics, FPA scholarship offers a broad range of possible answers, and thus levels of analysis, to this question. Possible explanatory factors include: risk propensities; the use of analogies or metaphors; learning; the influence of ‘hot’ and ‘cold’ emotions; motivations; identities; and attitudes. The following paragraphs offer in some greater detail two additional examples for FPA scholarship in this tradition, pertaining to leaders’ political beliefs on the one hand and information processing capacities on the other.

Operational Code Analysis (OCA) focuses on the political beliefs of decision makers (Leites, 1951; George, 1969; Walker et al., 1998). Operational codes are defined as ‘a set of general beliefs about fundamental issues of history and central questions of politics as these bear, in turn, on the problem of action’ (George, 1969: 191). The approach features 10 beliefs that are grouped into two clusters. Philosophical beliefs refer to decision-makers’ assumptions concerning the fundamental nature of politics, above all whether ‘the political universe [is] essentially one of harmony or conflict’ (George, 1969: 201). Hence, philosophical beliefs are crucial for the definition of a situation. In turn, instrumental beliefs relate to decision-makers’ understanding about ends-means relationships regarding action in the political realm. Most importantly, they provide answers to the question ‘what is

the best approach for selecting goals or objectives for political action?’ (George 1969: 205). Instrumental beliefs are thus decisive for the selection of the appropriate responses or instruments for the policy challenge at hand. Overall, political beliefs are conceptualized ‘as causal mechanisms with steering effects’ (Walker and Schafer, 2006: 7). Numerous applications of OCA have shown that individual decision makers, including non-Western leaders (e.g., Malici and Malici, 2005; He and Feng, 2015), do indeed vary in their political beliefs. Moreover, studies have successfully linked variation in beliefs with distinct foreign policy behaviors and outcomes, for instance by bringing in event data (e.g., Walker et al., 1999).

Integrative Complexity (e.g., Suedfeld et al., 1977; Suedfeld et al., 2011) focuses on leaders’ information processing capacities. A leader’s level of complexity is conceptualized as a ‘state’ in ‘which a person is functioning at a specific time and in a specific situation’ (Suedfeld et al., 2005: 247). The two key dimensions are differentiation and integration. Whereas the former refers to leaders’ ability to grasp multidimensionality by identifying alternative viewpoints and/or acknowledging different dimensions associated with a topic, the latter concerns leaders’ ability to interconnect those perceptions in the form of trade-offs, etc. (Suedfeld et al., 2011: 1009). A leader’s level of integrative complexity at any given point in time can be influenced by a number of factors, both intrapsychic (e.g., value conflicts) and context-specific (e.g., decision environment, political considerations, or the nature of the task) ones. Since integrative complexity is situation-specific, there is not only variation on this variable among leaders, but also a single leader can possibly show huge variation over time. This opens up a multitude of starting points for comparative analyses, for instance with respect to the integrative complexity of one or more leaders depending on the decision context, foreign policy domain, or the political stakes associated with the decision, to name



but a few. Like OCA and the Leadership Trait Analysis (see below for details), Integrative Complexity is also linked up with at-a-distance assessment techniques to render research feasible. Those techniques draw on leaders' speech acts from which quantitative data and ultimately 'scores' of leaders on the variable(s) of interest are derived. In case of Integrative Complexity, a broad array of leaders has been examined, including US president Barack Obama (Suedfeld et al., 2011) or leaders who were responsible for surprise attacks (Suedfeld and Bluck, 1988).

### ***Specifying the Second Image***

FPA scholarship offers an equally broad array of explanatory factors with respect to the 'second image' and thus possible domestic drivers of foreign policy processes and outcomes. They range from exploring interaction patterns in formal or informal small-group settings (e.g., cabinets or committees of the executive branch) to coordination- and decision-making processes within and between governmental organizations and the influence of other types of institutions (e.g., parliaments, the media, or lobby groups) on a country's foreign policy to general attributes of political systems (e.g., parliamentary versus presidential), regimes (e.g., democracy versus autocracy), or societies (e.g., strategic cultures or national role conceptions) and their impact on foreign policy behavior and outcomes. Resulting from this focus on domestic political actors, institutions, and processes, there are multiple cross-connections to Comparative Politics, where many of the same factors are addressed (see Brummer et al., 2019). The following paragraphs briefly outline two perspectives in this tradition in some greater detail, which focus on bureaucratic politics and decision-making groups and coalition governments respectively.

The Bureaucratic Politics Model focuses on inter-ministerial decision-making processes and the outcomes that they produce (Allison, 1971; Allison and Halperin, 1972; Allison and

Zelikow, 1999). The model's core assumptions can be summarized as follows: first, bureaucratic actors' policy preferences are strongly influenced, albeit not determined, by their position in government, hence the famous aphorism 'Where you stand depends on where you sit' (Allison and Zelikow, 1999: 307); second, bureaucratic actors engage in inter-ministerial bargaining processes since none of them wields sufficient power resources as to simply impose their preferences on other actors; and finally, the outcomes of bargaining processes among bureaucratic actors is best described as '*resultants*' in the sense that what happens is not chosen as a solution to a problem but rather results from compromise, conflict, and confusion of officials with diverse interests and unequal influence' (Allison and Zelikow, 1999: 294–55; emphasis in the original). The model's assumptions have been predominantly tested in qualitative case studies dealing with decision-making processes in 'industrialized nations' (Allison and Halperin, 1972: 43). In a few instances, the model has also been applied to non-Western settings, such as the Soviet Union (Dawisha, 1981).

Another type of domestic institution is addressed in the literature on the implications of coalition government for foreign policy. Scholarship on this topic can be grouped into two strands. One explores the implications of coalition government on the process of foreign policy making, for instance, with respect to the effects of junior coalition parties, or different types of coalitions more generally, on decision making procedures (e.g., Kaarbo, 1996; Oppermann and Brummer, 2014). Studies in this direction have predominantly used qualitative case studies, often times in a comparative fashion. The other strand in the literature focuses on the outcomes of foreign policy making by coalitions, for instance whether they end up with more 'extreme' foreign policies, that is, more cooperative or aggressive policies than single-party governments (e.g., Beasley and Kaarbo, 2014). Studies in that direction have predominantly employed quantitative methods.

## LANDMARK COMPARATIVE FPA SCHOLARSHIP

The previous section has shown that FPA scholarship contains a broad array of possible explanatory factors pertaining respectively to leaders and domestic institutions. Comparative approaches to the study of foreign policy have used those variables as their point of departure. The most ambitious among them have tried to integrate several of those factors in a comprehensive analytical framework. However, since such efforts – which date back to the 1960s and the 1970s – did not produce the aspired results, comparative approaches to foreign policy lost importance. This has changed in recent years, however, when calls to recover the comparative analysis of foreign policy has reinvigorated scholarship in this direction.

The foundational text for ‘Comparative Foreign Policy’ (CFP) as a distinct strand of research within FPA, and one of the ‘paradigmatic works’ of the field of FPA more generally (Hudson, 2012: 14), is a book chapter by James Rosenau entitled *Pre-theories and Theories of Foreign Policy* (Rosenau, [1966] 2006). The goal of that ‘pre-theory’ was to process information pertaining to foreign policy in a comparable fashion as a precondition for the subsequent building of actual theories of foreign policy, hence the term ‘pre-theory’. For that purpose, Rosenau came up with five sets of variables that serve as sources for the external behavior of states. Those were: characteristics of individual decision makers; the (bureaucratic or governmental) roles that decision makers occupy; governmental variables; societal variables; and systemic variables. Depending on different types of societies (large or small countries; developed or underdeveloped countries; open or closed polities), those variables were ranked according to their ‘relative potency’ (Rosenau, [1966] 2006: 175). For instance, in the case of a large, developed country with an open polity, the most important variable for explaining that country’s foreign policy was supposed to

be the roles of the decision makers, followed by the societal, governmental, systemic, and individual variables. Building on this foundation, additional concepts were added that were supposed to help at arriving at even more specific and nuanced understandings of the drivers of foreign policy. These included the distinction between penetrated and non-penetrated political systems as well between four different types of issues (territorial, status, human resources, and non-human resources). Those differentiations were supposed to help in ‘facilitat[ing] the formulation of if-then propositions’ (Rosenau, [1966] 2006: 177). However, since ‘[t]he empirical results were less than the protagonist had hoped’ the ultimate goal of arriving at ‘a grand unified theory of foreign policy behavior applicable to all nations and time periods’ did not materialize (Hudson, 2012: 19).

The CREON project (Comparative Research on the Event of Nations) pursued a quantitative cross-national comparative approach that sought to explore connections between different types of nations and specific foreign policy behavior (e.g., East et al. 1978; Callahan et al., 1982). A novel event data set for 38 states for the decade between 1959 and 1968 was compiled as an empirical base for the analysis (for details on the data, see Hermann et al., 1973). With the goal of developing ‘multivariate theories of foreign policy...interrelationships of independent variables with each other and with the dependent variables’ were explored in order to come up with ‘a more complete statement of the processes that produce foreign policy behavior’ (Callahan, 1982: 32). Mirroring Rosenau’s pre-theory, CREON featured three independent variables pertaining to national attributes (size, accountability, development). The combination of those attributes led to eight different ‘nation types’ (e.g., large, open, and developed nations or small, closed, and underdeveloped nations). CREON also featured nine dependent variables with respect to different types of foreign policy behavior. Those included the involvement of bureaucracy, participation by heads of states, different types

of government action ('verbal behavior' vs. 'physical deeds'), and different types of events (e.g., diplomatic, military, economic), among other things (East and Hermann, 1974).

However, neither Rosenau's pre-theories nor CREON produced the desired outcomes. As a result, the comparative analysis of foreign policy 'was largely discredited even by many of its own original founders...The label "CFP" came close to being pejorative in nature' (Kaarbo, 2003: 157). Against this background, Kaarbo (2003: 157) lamented in the early 2000s that 'One of the most disappointing features of contemporary FPA is the relative dearth of comparative study'. She concluded her criticism with a call for a 'return to comparison' along three dimensions: between issue areas, between countries, and over time, using either quantitative methodologies (as Rosenau and others did) or systematic qualitative methodologies (Kaarbo, 2003). Similarly, in the opening article of the then newly established journal *Foreign Policy Analysis*, Valerie Hudson noted that 'the term "comparative foreign policy" has largely disappeared from the sub-field' of FPA (Hudson, 2005: 14). At the same time, Hudson welcomed recent comparative works in which, for instance, several of FPA's leader-oriented frameworks such as OCA and Leadership Trait Analysis (LTA) have been employed in a comparative fashion against the same leader/s since this 'allow[ed] inspection of these frameworks' relative strengths and weaknesses' (Hudson, 2005: 17).

### **COMPARING LEADERS AND DOMESTIC INSTITUTIONS: CURRENT EXAMPLES OF COMPARATIVE FPA RESEARCH**

This section demonstrates that the aforementioned calls for a reinvention of comparative analysis in FPA have been heeded in recent years. Arguably, the main reason for this development has been that explanations

that focus on systemic imperatives and material factors were incapable of accounting for differences in the foreign policy behavior of states with similar capabilities, such as the non-/participation of key US allies in the Iraq war of 2003 (Kaarbo, 2003: 157). Such shortcomings not only suggested that it is necessary to explore the drivers of foreign policy that are located 'inside' a state – be it domestic institutions or individual leaders – but also to do so from a comparative perspective in order to explore how those factors play out differently in varying political environments in order to account for variation in behavior across countries.

FPA is particularly well-suited for comparative analysis given the high number of levels of analysis that it features as a result of the specification of first- and second-image drivers of foreign policy. The challenge is thus not to identify potentially meaningful substantive factors or variables for comparison but, rather, how such a comparison can be conducted in methodological terms. In the remainder of this section, two examples for state-of-the-art comparative research in FPA are presented. In accordance with FPA's focus inside the 'black box state', one example zooms in on leaders and the other on domestic institutions. On the level of individual decision makers, automated content analysis has been introduced to identify and ultimately compare specific characteristics of leaders in a systematic fashion within and across countries based on a huge amount of source material (in form of verbal utterances or speech acts). On the state level of analysis, Qualitative Comparative Analysis (QCA) has asserted itself as a viable tool for ascertaining domestic political drivers, and the interactions among them, of foreign policy from a comparative perspective.

To be sure, quantitative content analysis and QCA are not the only methods that can be used for comparative analyses of foreign policy. Meaningful comparative insights can also be gleaned by well-conducted and systematic qualitative case studies that employ

congruence tests and/or process tracing in conjunction with structured focused comparisons on a number of strategically selected cases to control for confounding factors as much as possible (Kaarbo and Beasley, 1999; George and Bennett, 2005), to give but one additional example. Still, since case-study methods are discussed in other chapters of this *Handbook*,<sup>1</sup> and since automated content analysis-schemes and QCA certainly belong to the best developed and most widely used instruments for that purpose, the following paragraphs aim to provide reasonably detailed insights into those state-of-the-art tools for the comparative analysis of foreign policy.

### ***Profiling Leaders Using Automated Content Analysis***

One of the main dictums of FPA is that leaders matter in foreign policy. This is not solely to suggest that analysts should take into account people in positions of political power when trying to explain foreign policy. Rather, it is the idiosyncratic features and characteristics of those leaders that drives foreign policy making and outcomes and which therefore need to be examined. As mentioned above, those characteristics could relate to decision makers' political beliefs, integrative complexity, risk propensities, or leadership traits and styles. Certain manifestations of those characteristics can be linked up, for instance, to certain types of policy behavior (e.g., a greater or lesser likelihood to use military force), and differences in those characteristics among leaders can help in accounting for different foreign policy behaviors among countries (even among those with similar material capabilities and domestic institutions).

Hence, techniques are required to ascertain the individual characteristics of leaders. Arguably the main challenge associated with an assessment, or 'profiling', of political leaders is access, or rather the lack thereof, to those actors. Indeed, researchers 'rarely have direct access to a leader in a way that

would allow for traditional psychological analysis' (Schafer, 2000: 512). In response, 'at-a-distance assessment techniques' have been developed to ascertain certain characteristics of leaders. Quantitative content analysis has become the primary research method in this regard since it allows for a systematic development of leaders' profiles based on their own statements without requiring direct access to them.

The substantive approaches that have been linked up with at-a-distance assessment techniques look at a variety of different variables, such as: political beliefs in case of OCA, leaders' ability for differentiation and integration in case of Integrative Complexity, or, as discussed in greater detail below, traits in case of Leadership Trait Analysis (LTA). What unites those approaches is that they draw on leaders' speech acts in order to identify specific manifestations of their respective variables of interest. By extension, they also face similar challenges with respect to their source material or choices pertaining to: impression management; audience effects; role effects; or issues of authorship (for details, see Schafer, 2000). Having said that, the internal and external validity of the approaches have been tried and tested in numerous studies so that researchers can use them with reasonable confidence.

Another issue that unites several of those at-a-distance approaches is that over the last 15 years or so, the quantitative content analysis of leaders' speech acts has become automated through the development of coding schemes that can be integrated in specific software packages. Based on LTA (Hermann, 2005), the following discussion shows how leadership traits can be identified based on the automated content analysis of verbal statements and how the ensuing results could be meaningfully used for comparative analysis.

### ***Leadership Trait Analysis***

The main proponent of LTA defines leadership style as 'the ways in which leaders relate

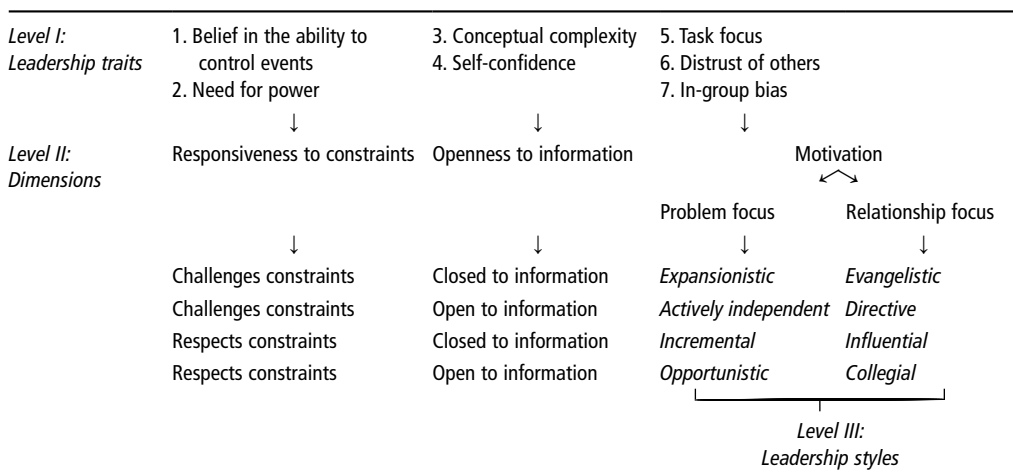
to those around them – whether constituents, advisers, or other leaders – and how they structure interactions and the norms, rules, and principles they use to guide such interactions’ (Hermann, 2005: 181). Hence, a specific leadership style, as well as the distinct leadership traits from which a specific leadership style emerges, exert a significant impact on the way that leaders seek to cope with the challenges and dilemmas of foreign policy decision making. In the final analysis, leadership traits and styles impact the conduct as well as the substance of foreign policy. LTA has been employed in numerous empirical studies. The latter have covered not only US presidents (Keller, 2005; Yi, 2010) or European prime ministers (Van Esch and Swinkels, 2015; Brummer, 2016), but also heads of international organizations (Kille and Scully, 2003; Brummer, 2014) as well as leaders from the Global South (Hermann, 1987; Taysi and Preston, 2001).

LTA can be broken down into three distinct levels (see Table 60.1). The first level focuses on seven specific leadership traits. Those traits, which represent the core of the approach, refer to: belief in ability to control events; need for power; conceptual complexity (i.e., ability for differentiation);<sup>2</sup> self-confidence; task focus; distrust of others; and

in-group bias. On the second level, certain traits are interacted to provide information on more general dimensions. Whether leaders respect or challenge constraints results from their belief in the ability to control events and need for power. Conceptual complexity and self-confidence provide information on leaders’ openness to information, and leaders’ motivation for office-seeking can be inferred from whether they focus on tasks or rather on group maintenance, their general distrust or suspiciousness of others, and whether they exhibit an in-group bias. On the third level, those three dimensions interact to produce a total of eight distinct leadership styles. For instance, a leader who challenges constraints, is closed to information, and focuses on problems is called ‘expansionistic’, whereas a leader who respects constraints, is closed to information, and focuses on relationships is called ‘influential’.

LTA assumes that leaders’ traits (and thus also the ensuing styles) are rather stable over time as well as similar across issue areas.<sup>3</sup> That is, a leader who challenges constraints is supposed to behave this way throughout his or her political career and to do so irrespective of the substantive domain that has to be addressed. Those assumptions have implications for the compilation of the ‘proper’

**Table 60.1 Leadership trait analysis**



Source: own depiction based on Hermann (2005).

source material that can be used for the identification of traits and styles.

Of course, the key question is how the key variables contained in LTA can be identified empirically. This is where the 'at-a-distance technique' for assessing leaders comes into play. LTA assumes that leaders' personalities or, more precisely, leaders' traits and styles can be inferred from their verbal statements (Hermann, 2005: 178–9). Generally speaking, then, leadership traits can be inferred from the systematic content analysis of leaders' speech acts.

More specifically, LTA stipulates a total of six requirements regarding the source material that should be used for empirical analysis (Hermann, 2005, 2008). For every leader, at least 50 verbal statements (ideally: at least 100) must be assembled (requirement one). Each statement should contain at least 100 words (ideally: at least 150) (requirement two). However, those requirements are a legacy of the time period when the coding process was still performed manually; with the advent of automated, computer-based coding (see below), the text corpus can be significantly larger nowadays as the coding even of several hundred thousand words does not take longer than a couple of minutes. Moreover, not just any verbal statement should be used for analysis. Rather, the text corpus must contain only spontaneous statements (requirement three). These are much more likely to reveal the 'true self' of leaders than scripted statements: '[I]nterviews are the material of preference. In the interview, political leaders are less in control of what they say and, even though still in a public setting, more likely to evidence what they, themselves, are like than is often possible when giving a speech' (Hermann, 2005: 179). Relatedly, the verbal statements should have been uttered in different contexts, that is, in front of different audiences, such as parliament or the press (requirement four). Two further requirements follow from the above-mentioned substantive assumptions contained in LTA with respect to the stability of traits across time and issue

domain. Following from this, the verbal statements should cover the leader's entire tenure in office rather than, say, just from his or her first or second term (requirement five), and they should also focus on different substantive issue areas (requirement six).

Source material (i.e., spontaneous verbal statements) assembled in light of those requirements is then content-analyzed. Until the early 2000s, the coding was done manually. Since then, automated coding has become possible with the development of a coding scheme for LTA that has been integrated in the software program 'Profiler Plus' developed by Social Science Automation. The automated coding scheme enables the analysis of much larger amounts of source material in virtually 'no time'. Indeed, the collection of the verbal statements is now much more time consuming than the actual coding process. What is more, concerns about the accuracy of the coding and intercoder reliability are no longer an issue (see below for possible other downsides, however).

The automated coding focuses exclusively on the identification of the above-mentioned seven leadership traits (level I). The ensuing three dimensions (level II) and eight leadership styles (level III) must be inferred by the analyst based on the results for the seven traits. Every trait is associated with a particular set of words or expressions that are supposed to indicate the existence of either high or low manifestations of the respective trait. For all seven traits, the range for the score is between 0 and 1. The lower the score, the weaker the manifestation of the respective trait, and vice versa.

For the purpose of illustration: the source material (i.e., the spontaneous verbal statements that the analyst has assembled for a specific leader in accordance with the six requirements outlined above) contains 100 words that are associated with the trait conceptual complexity; 51% of those words point to a high manifestation of that trait, such as 'maybe', 'possibly', or 'for example'. Conversely, 49% of those words suggest

the opposite, such as ‘always’, ‘certainly’, or ‘unequivocally’. Since the score for conceptual complexity is determined by calculating the percentage share of high-complexity words among all high and low complexity words contained in the source material, our fictional leader would have a conceptual complexity score of 0.51. Indeed, this is the conceptual complexity score for the former British Prime Minister David Cameron, who was responsible for the Brexit referendum in 2016, among other things.

The explanatory power of this score is somewhat limited, though. On the spectrum ranging from 0 to 1 Cameron’s score suggests that his ability to distinguish the complexities of the political environment is rather average. However, the score itself tells us nothing in terms of how well Cameron performs (i.e., above average, below average, or just average) *relative* to his peers of other key foreign policy makers, and hence the main reference group.

At this point, one of the main strengths of LTA – which makes it particularly useful for comparative analysis – comes to the fore. It results from a so-called ‘norming group’. This norming group comprises a total of 284 ‘world leaders’ (Table 60.2). It contains mean scores for all seven leadership traits, not only for the full list of leaders (comprising presidents, prime ministers, and foreign ministers), but is also broken down into, for the most part, six geographically defined sub-groupings, namely: Western Europe, Eastern Europe, Middle East and Northern Africa, Pacific Rim, Latin America, and Anglo-America.<sup>4</sup> The added value of this norming group is straightforward: in practical terms, it renders possible meaningful comparisons between leaders without having to create a profile for more than one leader, and, in substantive terms, its helps to put into context the scores of one’s own analysis.

To return to our previous example: Cameron’s score for conceptual complexity was 0.51. The norming group shows that the average score for this particular trait among close to 300 world leaders is 0.59, with a

**Table 60.2 The ‘norming group’ of leadership trait analysis**

Leadership traits	‘World leaders’ (n=284)
Belief in the ability to control events	0.35 (0.05)
Need for power	0.26 (0.05)
Conceptual complexity	0.59 (0.06)
Self-confidence	0.36 (0.10)
Task focus	0.63 (0.07)
Distrust of others	0.13 (0.06)
In-group bias	0.15 (0.05)

*Source:* own depiction based on Cuhadar et al. (2017): 47 (average scores for ‘world leaders’; standard deviations in parentheses).

standard deviation of 0.06; for the Anglo-American sub-group, it is 0.60 with a standard deviation of 0.05. Based on those numbers, we can tell that Cameron’s trait score is more than one standard deviation below both the world average and the regional sub-group. This is not only a finding in itself but becomes all the more pertinent when taking into account possible implications of different manifestations of conceptual complexity on leaders’ foreign policy behavior. For instance, leaders with a low conceptual complexity are more likely to engage in diversionary behavior (Foster and Keller, 2014). In a sense, one could argue that the domestic political challenges that Cameron had been facing both within his own (in parts strongly anti-European) Conservative Party and from the staunchly anti-European United Kingdom Independence Party (UKIP) provided strong diversionary incentives for him so that he promised a ‘Brexit’ referendum in the hope of rallying at least his own party behind him.

In short, leadership profiling based on automated content analysis represents an ideal starting point for engaging in comparative analysis of foreign policy decision makers, both within and across countries.<sup>5</sup> However, in conclusion, some caveats are in order. First, LTA remains essentially silent about how leadership traits are formed in the first place. Admittedly, however, answering this question

would require a different theory altogether (Horowitz et al., 2015). Second, the norming group contains certain problems. Most importantly, it is not clear which leaders are actually included in that group. As a result, one cannot tell whether the norming group features certain biases for instance with respect to the party-political/ideological background of the leaders or the types of political systems (presidential or parliamentarian) or regimes (democratic or autocratic) within which the leaders operate. It is also impossible to tailor a norming group based on one's specific needs (e.g., a group that includes only female leaders or leaders of democratic countries). Finally, LTA-based automated content analysis is, for the most part, confined to English source material. The only exceptions include Spanish coding schemes for the traits 'conceptual complexity' and 'distrust' and a newly developed German coding scheme that covers all seven traits (Rabini et al., 2018).<sup>6</sup> Hence, analyzing foreign policy decision makers whose first language is not English can become a challenge due to a dearth in source material. To be clear, however, this 'language challenge' relates to the availability of appropriate material, and not to the applicability of LTA to the non-English speaking world, to which it has been applied in numerous empirical studies already.

### ***Comparing Domestic-political Drivers of Foreign Policy Using Qualitative Comparative Analysis<sup>7</sup>***

The method of Qualitative Comparative Analysis (QCA) is a rather recent addition to the methodological toolbox of political science/IR (see Ragin 1987, 2000, 2008; Rihoux and Ragin 2009). QCA brings together insights from variable-oriented and case-oriented research. In so doing, it offers a 'diversity-oriented research [that] emphasizes "types" and "kinds," a formulation that views cases configurationally, as combinations of aspects' (Ragin, 2000: 14) that produce

certain results in a specific-context. As a result, QCA enables analysts to uncover causal relationships between an 'outcome' (similar to a dependent variable) and certain 'conditions' (similar to independent variables). The method is particularly well-suited for (but by no means limited to) 'medium-n' empirical analyses where the number of cases yields too few observations for statistical analyses but too many for in-depth case studies.

More specifically, QCA represents a set-theoretic approach that draws on Boolean algebra. The method can be used to establish the necessary and sufficient conditions for a certain outcome to occur. A *necessary* condition must be present whenever the outcome is observable. However, its presence is not a guarantee for the actual occurrence of the outcome. In other words, there might be instances where the necessary condition is present, but the outcome is not. Hence, the outcome is a sub-set of the necessary condition. In turn, if a *sufficient* condition is present, then the outcome is also always present. However, the outcome might also occur as a result of other conditions. Hence, the outcome is a super-set of the sufficient condition. Moreover, there are conditions that are neither necessary nor sufficient for producing a certain outcome but nonetheless can contribute to producing that outcome (so-called 'INUS' and 'SUIN' conditions). Overall, then, QCA highlights not only that one individual condition can lead to a certain outcome, but also, and more importantly, that certain 'configurations of conditions can be jointly necessary and/or sufficient, while their constituent parts might be neither necessary nor sufficient for an outcome' (Mello, 2012: 429) to occur.

Arguably, the key distinguishing features of QCA as a method are that it focuses on a combinations of factors (i.e., conditions) in producing certain results (i.e., outcomes) – or the non-occurrence (i.e., negation) of an outcome, for that matter – and that it allows for one and the same outcome to be produced by different combinations of conditions. The



literature refers to this as ‘causal complexity’ (Schneider and Wagemann 2012: 76–90). Causal complexity comprises three elements: equifinality, conjunctural causation, and asymmetry. Equifinality means that multiple causal pathways (i.e., individual conditions or, more often than not, specific combinations of conditions) can produce the same outcome. Conjunctural causation suggests that the effects of single conditions unfold only in combination with other conditions. Finally, asymmetry means ‘that insights on the causal role of a condition are of only limited use for the causal role of its absence, and the explanation of the occurrence of an outcome does not necessarily help us much in explaining its non-occurrence’ (Schneider and Wagemann, 2012: 81). To give an example: just because a certain combination of conditions (e.g.,  $x$  is present,  $y$  is absent, and  $z$  is present) leads to a certain outcome, this does not necessarily also mean that the mirror image (i.e.,  $x$  is absent,  $y$  is present, and  $z$  is absent) automatically leads to the non-occurrence of the outcome. Therefore, since knowing why a certain outcome has occurred does not necessarily tell us much about the non-occurrence of the outcome, separate empirical analyses are required: one for the occurrence of a certain outcome and one for its non-occurrence.

QCA features different variants. Arguably, the most prominent ones are crisp set QCA (csQCA) and fuzzy set QCA (fsQCA). The key difference between them is that they are based on different types of sets:

csQCA operates exclusively on conventional sets where case can either be members or non-members in the set. Their set membership is either 0 or 1. In fsQCA, by contrast, cases are allowed to have gradations of their set membership. A case does not have to be a full member or a full non-member of a set but can also be a partial member. The membership scores can all anywhere between the two extremes of full membership value of 1 and full non-membership value of 0. (Schneider and Wagemann, 2012: 13–14)

By implication, csQCA is a special variant of fsQCA, in that it only knows either full

membership or non-membership in a set, rather than also gradated levels of membership as is the case in fsQCA (Ragin, 2000: 156).

Overall, then, whereas csQCA ‘only’ looks at qualitative distinctions, or differences in kind (in the sense of ‘more or less’), fsQCA also focuses on quantitative distinctions, or differences in degree (how much ‘more or less’) (Schneider and Wagemann, 2012: 24–7). Which variant of QCA should be used in empirical analysis is up for the researcher to decide in light of the explanatory factors (conditions and outcome) that are included in the analysis. If the factors yield to a dichotomous classification, then csQCA would be an obvious choice (e.g., single-party versus coalition government). However, if some, most, or all of the factors can be depicted in a more nuanced fashion that also gets at quantitative distinctions, then fsQCA should be used (e.g., when it is not only of importance whether there is a coalition government or not, but whether also the number of coalition parties is relevant for the argument, for instance, with respect to the number of veto players).

### ***Examples for Qualitative Comparative Analyses in FPA***

Contrary to LTA as discussed above – which is first and foremost a substantive approach that has been associated with a specific method (in form of automated content analysis) – QCA does not entail any specific substantive assumptions based on which specific drivers of foreign policy can be ascertained. As a result, QCA as a tool for comparative analysis needs to be linked up with substantive approaches in order to arrive at meaningful findings. With this in mind, virtually any of FPA’s (domestic) levels of analysis can be used in conjunction with QCA. More precisely: a multitude of combinations of FPA’s domestic levels of analysis can be employed in principle, with the specific ‘mix’ being determined not by QCA, but by the analyst’s

theory-driven assumptions about possible drivers (conditions) for a specific phenomenon (outcome). This openness in terms of explanatory factors, or conditions, that can be investigated by using QCA becomes obvious in the following two examples, one relating to csQCA and the other to fsQCA.

csQCA has been used to explain the allocation of the foreign ministry, and thus one of the key portfolios within the ‘foreign policy executive’ (Hill, 2016), in governing coalitions. More specifically, Oppermann and Brummer (forthcoming) explore why the senior party of a coalition government (i.e., the party with the largest seat share in parliament of all parties involved in the governing coalition) refrains from taking over the foreign ministry and instead agrees to allocate this portfolio to a junior member of the coalition government (i.e., a party with a relatively smaller seat share in parliament). The latter represents the ‘outcome’ under examination. Drawing on the portfolio allocation literature from Comparative Politics (Laver and Schofield, 1990) and the FPA literature on parties (Rathbun, 2004) as well as coalition governments and foreign policy (Kaarbo and Beasley, 2008), the authors identify five ‘conditions’ as possible drivers for that outcome. Those are: the relative size of coalition parties; the proximity of the foreign policy positions; the party family of the junior coalition party; the salience that coalition parties ascribe to foreign policy; and allocations of the foreign ministry to junior coalition partners in the past. Representing INUS conditions, they find that for a sample of 18 European countries, the combination of the conditions ‘relative size’ of the junior party (i.e., when it is almost as large as the senior party) and ‘past allocations’ to a junior party holds the most explanatory power.

The fsQCA example selected for this chapter addresses the use of economic statecraft in international politics. Squarely embedded in the literature on the ‘normative power Europe’ (Manners, 2002), Boogaerts (2018) examines the sanctions behavior of

the European Union (EU) in reaction to the Arab Spring. Although using fsQCA, the outcome variable is dichotomized in the sense of whether the EU has adopted (partially) autonomous ‘restrictive measures’ (i.e., sanctions) against a total of 13 Arab states in the years 2011 and 2012. The four conditions, however, are coded on a fuzzy scale. They refer to: the violent suppression of protest by Arab states; the EU’s material and security interests vis-à-vis those states; prior instances of restrictive measures in the economic realm; and transitional void (akin to the breakdown of an authoritarian regime). Perhaps most importantly, the article finds that ‘human rights violations do not constitute a sufficient condition for EU sanctions in the MENA region’ (Boogaerts, 2018: 423) but only play a role in conjunction with other conditions, which casts doubts on the (self-) depiction of the EU as norm-driven actor.

## CONCLUSION

The most characteristic feature of FPA as a distinct sub-field of IR is that it opens up the ‘black box state’ and examines the decision makers that act as well as the political processes that unfold inside that box. In so doing, FPA has ‘fine-tuned’ Waltz’s first- and second-image perspectives by offering a broad array of distinct characteristics respectively of leaders and domestic institutions that plausibly serve as drivers of states’ external behavior. It is within and across the ensuing greater number of levels of analysis that a comparative approach to the study of foreign policy can be fruitfully employed.

It goes without saying that comparative analyses, especially when comparing countries (rather than, say, the evolution of a specific policy within one country), is demanding. It requires researchers to be familiar with the histories of several countries, the functioning of those countries’ political institutions, and the political culture of the respective countries,

among other things. In addition to expert empirical knowledge, familiarity with the languages of the countries under examination is also extremely valuable (though, I would contend, not a *sine qua non* condition since it would significantly restrict the number of countries, hence cases, for examination, especially in non-collaborative projects). Having said that, there are several upsides of comparative approaches to the study of foreign policy. Most importantly, it allows researchers to get a more nuanced and, at the same time, also a more comprehensive understanding of how key drivers of foreign policy – which can be associated with either individual leaders or domestic institutions – play out in different political environments which, in turn, is a prerequisite for ascertaining the more general, cross-country effects and influence of those factors on state behavior.

To accomplish the goal of comparative analyses of foreign policy, a number of well-specified analytical tools are at the disposal of analysts. This chapter has presented two of them in some greater detail. First, automated content analysis can be used to identify idiosyncratic characteristics of leaders which, in turn, can be linked to specific types of foreign policy behaviors by those leaders and ultimately to specific foreign policy outcomes. Second, QCA can be used to examine the interaction of several domestic-political drivers of foreign policy. Automated content analysis enables analysts to develop personality profiles of individual leaders ‘at-a-distance’ based on the processing of huge amounts of leaders’ verbal statements. It thus solves the problem that researchers do not have (regular) access to decision makers. In turn, one of the key strengths of QCA is that it takes both conjunctural causation and equifinality seriously. So rather than narrowing down explanations of foreign policy behaviors or outcomes to one specific variable, QCA allows for the interaction of several explanatory factors, and it also suggests that one and the same outcome can be produced by different combinations of factors.

However, both methods feature certain drawbacks from which the following avenues for future research can be inferred. To date, automated content analysis, which primarily relies on the mining of leaders’ speech acts, is for the most part limited to English-language source material. It is not by accident, therefore, that many studies focus on leaders from the English-speaking world, for whom source material usually exists in abundance. Conversely, analyses of leaders whose mother tongue is not English more often than not run into problems in terms of source material, since utterances in English tend to be few and far between, official translations are hard to come by, and the option of researchers themselves engaging in translations of texts usually entails prohibitive costs. Therefore, automated coding schemes should be developed to enable the analysis of leaders in their own (non-English) language. A positive side-effect of this would be to further de-center the field of FPA away from the United States and the Anglosphere more generally.

Moreover, future comparative FPA scholarship should bring together leaders and domestic institutions. The examples provided above have shown that comparative analyses tend to focus on either one or the other, with little interconnection and thus cross-fertilization between those two areas. Most palpably, while taking into account a variety of institutional factors, QCA studies usually do not incorporate leader-specific variables in their analyses. Indeed, leaders’ characteristics are not readily available for turning them into crisp or fuzzy sets. Rather, this would require separate analyses in which the specific manifestation of certain characteristics of leaders (i.e., those which the theoretical literature qualifies as a possible condition for the outcome under examination) are established. Based on those scores, leader-specific factors could be represented by crisp or fuzzy sets, thus be incorporated into a QCA study. For instance, one could hypothesize that a prime minister’s need for power (as one of the seven traits contained in LTA) is one condition for

his or her willingness to give away key foreign policy portfolios to junior coalition partners, with the assumption being that the higher the leader's need for power, the less willing he or she is to cede control of, say, the foreign ministry to another party. Such integrated perspectives promise to offer a fuller picture of the domestic drivers of foreign policy by interacting explanatory factors pertaining to both leaders and domestic institutions.

## Notes

- 1 For details on case-study methods as well as on process tracing (and structured focused comparisons), see the chapter by Ruffa, Chapter 59, in this *Handbook*.
- 2 'Conceptual complexity' is not to be confused with 'integrative complexity' mentioned above, since the former refers to an essentially stable 'trait' while the latter addresses a context-specific 'state' (Suedfeld et al., 2005: 247).
- 3 While variation across 'time, audience, and topic' (Hermann, 2005: 180) is not categorically ruled out, the clear majority of the empirical studies using LTA (implicitly or explicitly) assume that traits and styles are essentially stable and, very importantly, sample the source material for their empirical analysis accordingly (i.e., without differentiating between issue areas or over time).
- 4 The latest norming group overview can be obtained through the Profiler Plus website (profilerplus.org).
- 5 Both the Operational Code Approach and Integrative Complexity can be used in a similar fashion. For details on how those approaches and quantitative content analysis have been integrated, see Schafer and Walker (2006) and Suedfeld et al. (2005).
- 6 The Spanish coding schemes are already available from profilerplus.org; the German coding scheme will be made available there as well.
- 7 For a more detailed discussion of QCA, and set theoretic methods more generally, see the chapter by Duşa, Chapter 57, in this *Handbook*.

## REFERENCES

- Allison, Graham T. (1971) *Essence of Decision: Explaining the Cuban Missile Crisis*. Boston: Little, Brown and Company.

- Allison, Graham T., and Morton H. Halperin (1972) Bureaucratic Politics: A Paradigm and Some Policy Implications. *World Politics* 24(Supplement), 40–79.
- Allison, Graham T., and Philip Zelikow (1999) *Essence of Decision: Explaining the Cuban Missile Crisis*, 2nd edition. New York: Longman.
- Beasley, Ryan, and Juliet Kaarbo (2014) Explaining Extremity in the Foreign policies of Parliamentary Democracies. *International Studies Quarterly* 58(4), 729–740.
- Boogaerts, Andreas (2018) Beyond Norms: A Configurational Analysis of the EU's Arab Spring Sanctions. *Foreign Policy Analysis* 14(3), 408–428.
- Brummer, Klaus (2014) Die Führungsstile von Präsidenten der Europäischen Kommission. *Zeitschrift für Politik* 61(3), 327–345.
- Brummer, Klaus (2016) 'Fiasco Prime Ministers'. Beliefs and Leadership Traits as Possible Causes for Policy Fiascos. *Journal of European Public Policy* 23(5), 702–717.
- Brummer, Klaus, Sebastian Harnisch, Kai Oppermann, and Diana Panke (eds) (2019) *Foreign Policy as Public Policy?* Manchester: Manchester University Press.
- Callahan, Patrick (1982) The CREON Project. In Patrick Callahan, Linda P. Brady, and Margaret G. Hermann (eds) *Describing Foreign Policy Behavior*. Beverly Hills: Sage, 31–51.
- Callahan, Patrick, Linda P. Brady, and Margaret G. Hermann (eds) (1982) *Describing Foreign Policy Behavior*. Beverly Hills: Sage.
- Cuhadar, Esra, Juliet Kaarbo, Baris Kesgin, and Binnur Ozkececi-Taner (2017) Personality or Role? Comparisons of Turkish Leaders Across Different Institutional Positions. *Political Psychology* 38(1), 39–54.
- Dawisha, Karen (1981) Soviet Decision-Making in the Middle East: The 1973 October War and the 1980 Gulf War. *International Affairs* 57(1), 43–59.
- East, Maurice, and Charles F. Hermann (1974) Do Nation-Types Account for Foreign Policy Behavior? In James N. Rosenau (ed.) *Comparing Foreign Policies. Theories, Findings, and Methods*. New York: John Wiley & Sons, 269–303.
- East, Maurice, Stephen A. Salmore, and Charles F. Hermann (eds) (1978) *Why Nations Act*. Beverly Hills: Sage.
- Foster, Dennis, and Jonathan W. Keller (2014) Leaders' Cognitive Complexity, Distrust, and

- the Diversionary Use of Force. *Foreign Policy Analysis* 10(3), 205–223.
- George, Alexander L. (1969) The 'Operational Code': A Neglected Approach to the Study of Political Leaders and Decision-Making. *International Studies Quarterly* 13(2), 190–222.
- George, Alexander L. and Andrew Bennett (2005) *Case Studies and Theory Development in the Social Sciences*. Cambridge, MA, and London: MIT Press.
- He, Kai, and Huiyun Feng (2015) Transcending Rationalism and Constructivism: Chinese Leaders' Operational Codes, Socialization Processes, and Multilateralism after the Cold War. *European Political Science Review* 7(3), 401–426.
- Hermann, Charles F., Maurice A. East, Margaret G. Hermann, Barbara G. Salmore, and Stephen A. Salmore (1973) *CREON: A Foreign Events Data Set*. Sage Professional Paper/International Studies Series. Beverly Hills and London: Sage.
- Hermann, Margaret G. (1987) Assessing the Foreign Policy Role Orientations of Sub-Saharan African Leaders. In Stephen G. Walker (ed.) *Role Theory and Foreign Policy Analysis*. Durham: Duke University, 161–198.
- Hermann, Margaret G. (2005) Assessing Leadership Style: Trait Analysis. In Jerrold M. Post (ed.) *The Psychological Assessment of Political Leaders. With Profiles of Saddam Hussein and Bill Clinton*. Ann Arbor: University of Michigan Press, 178–212.
- Hermann, Margaret G. (2008) Content Analysis. In Audie Klotz and Deepa Prakash (eds) *Qualitative Methods in International Relations*. London: Palgrave Macmillan, 151–167.
- Hill, Christopher (2016) *Foreign Policy in the Twenty-first Century*, 2nd edition. Basingstoke: Palgrave Macmillan.
- Horowitz, Michael, Allan C. Stam, and Cali M. Ellis (2015) *Why Leaders Fight*. Cambridge: Cambridge University Press.
- Holsti, Ole R. (1976) Foreign Policy Formation Viewed Cognitively. In Axelrod, Robert (ed.) *Structure of Decision. The Cognitive Maps of Political Elites*. Princeton: Princeton University Press, 18–54.
- Hudson, Valerie M. (2005) Foreign Policy Analysis: Actor-Specific Theory and the Ground of International Relations. *Foreign Policy Analysis* 1(1), 1–30.
- Hudson, Valerie M. (2012) The History and Evolution of Foreign Policy Analysis. In Steve Smith, Amelia Hadfield, and Tim Dunne (eds) *Foreign Policy. Theories, Actors, Cases*, 2nd edition. Oxford: Oxford University Press, 13–34.
- Jervis, Robert (2013) Do Leaders Matter and How Would We Know? *Security Studies* 22(2), 153–179.
- Kaarbo, Juliet (1996) Power and Influence in Foreign Policy Decision Making: The Role of Junior Coalition Partners in German and Israeli Foreign Policy. *International Studies Quarterly* 40(4), 501–530.
- Kaarbo, Juliet (2003) Foreign Policy Analysis in the Twenty-First Century: Back to Comparison, Forward to Identity and Ideas. *International Studies Review* 5(2), 156–163.
- Kaarbo, Juliet (2015) A Foreign Policy Analysis Perspective on the Domestic Politics Turn in IR Theory. *International Studies Perspectives* 17(2), 189–216.
- Kaarbo, Juliet, and Ryan K. Beasley (1999) A Practical Guide to the Comparative Case Study Method in Political Psychology. *Political Psychology* 20(2), 369–391.
- Kaarbo, Juliet, and Ryan K. Beasley (2008) Taking It to the Extreme: The Effect of Coalition Cabinets on Foreign Policy. *Foreign Policy Analysis* 4(1), 67–81.
- Keller, Jonathan (2005) Constraint Respecters, Constraint Challengers, and Crisis Decision Making in Democracies: A Case Study Analysis of Kennedy versus Reagan. *Political Psychology* 26(6), 835–867.
- Kille, Kent J., and Roger M. Scully (2003) Executive Heads and the Role of Intergovernmental Organizations: Expansionist Leadership in the United Nations and the European Union. *Political Psychology* 24(1), 175–198.
- Laver, Michael, and Norman Schofield (1990) *Multiparty Government. The Politics of Coalition in Europe*. Oxford: Oxford University Press.
- Leites, Nathan (1951) *The Operational Code of the Politburo*. New York: McGraw-Hill.
- Malici, Akan, and Johnna Malici (2005) The Operational Codes of Fidel Castro and Kim Il Sung: The Last Cold Warriors? *Political Psychology* 26(3), 387–412.
- Manners, Ian (2002) Normative Power Europe: A Contradiction in Terms? *Journal of Common Market Studies* 40(2), 235–258.

- Mello, Patrick (2012) Parliamentary Peace or Partisan Politics? Democracies' Participation in the Iraq War. *Journal of International Relations and Development* 15(3), 420–453.
- Oppermann, Kai, and Klaus Brummer (2014) Patterns of Junior Partner Influence on the Foreign Policy of Coalition Governments. *British Journal of Politics and International Relations* 16(4), 555–571.
- Oppermann, Kai, and Klaus Brummer (forthcoming) Who Gets What in Foreign Affairs? Explaining the Allocation of Foreign Ministries in Coalition Governments. *Government and Opposition*, doi.org/10.1017/gov.2018.19.
- Rabini, Christian, Klaus Brummer, Katharina Dimmroth, and Mischa Hansel (2018) 'Wir schaffen das': Grasping the Role of Decision Makers in German Foreign Policy. Paper presented at the 59th Annual Convention of the International Studies Association, 4–7 April 2018, San Francisco.
- Ragin, Charles C. (1987) *The Comparative Method. Moving Beyond Qualitative and Quantitative Strategies*. Oakland: University of California Press.
- Ragin, Charles C. (2000) *Fuzzy-Set Social Science*. Chicago and London: University of Chicago Press.
- Ragin, Charles C. (2008) *Redesigning Social Inquiry. Fuzzy Sets and Beyond*. Chicago, and London: University of Chicago Press.
- Rathbun, Brian C. (2004) *Partisan Interventions: European Party Politics and Peace Enforcement in the Balkans*. Ithaca: Cornell University Press.
- Rihoux, Benoît, and Charles C. Ragin (eds) (2009) *Configurational Comparative Methods. Qualitative Comparative Analysis (QCA) and Related Techniques*. Los Angeles: Sage.
- Rosenau, James N. ([1966] 2006) Pre-theories and Theories of Foreign Policy. In James N. Rosenau (ed) *The Study of World Politics. Theoretical and Methodological Challenges, Volume 1*. London and New York: Routledge, 171–199.
- Schafer, Mark (2000) Issues in Assessing Psychological Characteristics at a Distance: An Introduction to the Symposium. *Political Psychology* 21(3), 511–527.
- Schafer, Mark, and Stephen G. Walker (2006) Operational Code Analysis at a Distance: The Verbs in Context System of Content Analysis. In Mark Schafer and Stephen G. Walker (eds) *Beliefs and Leadership in World Politics. Methods and Applications of Operational Code Analysis*. Basingstoke and New York: Palgrave Macmillan, 25–51.
- Schneider, Carsten Q., and Claudius Wagemann (2012) *Set-Theoretic Methods for the Social Sciences. A Guide to Qualitative Comparative Analysis*. Cambridge: Cambridge University Press.
- Schweller, Randall L. (2004) Unanswered Threats. A Neoclassical Realist Theory of Underbalancing. *International Security* 29(2), 159–201.
- Simon, Herbert A. (1957) *Models of Man. Social and Rational*. New York: John Wiley & Sons.
- Snyder, Richard C., H. W. Bruck, and Burton Sapin ([1954] 2002) Decision-Making as an Approach to the Study of International Politics. In Richard C. Snyder, H. W. Bruck, and Burton Sapin (eds) *Foreign Policy Decision-Making (Revisited)*. New York and Basingstoke: Palgrave Macmillan, 21–152.
- Suedfeld, Peter, Karen Guttieri, and Philip E. Tetlock (2005) Assessing Integrative Complexity at a Distance: Archival Analyses of Thinking and Decision Making. In Jerrold M. Post (ed) *The Psychological Assessment of Political Leaders. With Profiles of Saddam Hussein and Bill Clinton*. Ann Arbor: University of Michigan Press, 246–270.
- Suedfeld, Peter, Philip E. Tetlock, and Carmenza Ramirez (1977) War, Peace, and Integrative Complexity: UN Speeches on the Middle East Problem, 1947–1976. *Journal of Conflict Resolution* 21(3), 427–442.
- Suedfeld, Peter, Ryan W. Cross, and Jelena Bric (2011) Two Years of Ups and Downs: Barack Obama's Patterns of Integrative Complexity, Motive Imagery, and Values. *Political Psychology* 32(6), 1007–1033.
- Suedfeld, Peter, and Susan Bluck (1988) Changes in Integrative Complexity Prior to Surprise Attacks. *Journal of Conflict Resolution* 32(4), 626–635.
- Taysi, Tanyel, and Thomas Preston (2001) The Personality and Leadership Style of President Khatami: Implications for the Future of Iranian Political Reform. In Ofer Feldman and Linda O. Valenty (eds) *Profiling Political Leaders. Cross-Cultural Studies of Personality and Behavior*. Westport and London: Praeger, 57–77.

- Van Esch, Femke, and Marij Swinkels, (2015) How Europe's Political Leaders Made Sense of the Euro Crisis: The Influence of Pressure and Personality. *West European Politics* 38(6), 1203–1225.
- Walker, Stephen G. (2003) Operational Code Analysis as a Scientific Research Program. A Cautionary Tale. In Colin Elman and Miriam F. Elman (eds) *Progress in International Relations Theory. Appraising the Field*. Cambridge, MA and London: MIT Press, 245–276.
- Walker, Stephen G., and Mark Schafer (2006) Belief Systems as Causal Mechanisms in World Politics: An Overview of Operational Code Analysis. In Mark Schafer and Stephen G. Walker (eds) *Beliefs and Leadership in World Politics. Methods and Applications of Operational Code Analysis*. Basingstoke and New York: Palgrave Macmillan, 3–22.
- Walker, Stephen G., Mark Schafer, and Michael D. Young (1998) Systematic Procedures for Operational Code Analysis: Measuring and Modeling Jimmy Carter's Operational Code. *International Studies Quarterly* 42(1), 175–190.
- Walker, Stephen G., Mark Schafer, and Michael D. Young (1999) Presidential Operational Codes and Foreign Policy Conflicts in the Post-Cold War World. *Journal of Conflict Resolution* 43(5), 610–625.
- Waltz, Kenneth N. (1959) *Man, the State and War. A Theoretical Analysis*. New York: Columbia University Press.
- Waltz, Kenneth N. (1979) *Theory of International Politics*. Boston: McGraw-Hill.



# When Talk Isn't Cheap: Opportunities and Challenges in Interview Research

Claire Greenstein and Layna Mosley

Many substantive questions and theoretical claims in political science can be usefully evaluated using various forms of qualitative analysis, from process tracing and comparative historical analysis to focus groups, ethnography, and interviews (Bennett and Checkel, 2014; Grzymala-Busse, 2011; Kapiszewski et al., 2018). Descriptions and analyses using qualitative-data-gathering techniques can help scholars to generate causal claims, build theories, provide specific examples of the more general processes assumed by formal and statistical models, and evaluate hypotheses.<sup>1</sup>

Interviews of elites or members of the mass public can play an important role: researchers might ask policymakers when and how they decide to disclose information on nuclear proliferation (Carnegie and Carson, 2019), query investors about how they evaluate the risks represented by various types of political institutions and political events (Mosley, 2003), or discuss with members of the public when they are willing to commit violence

against – and when they are willing to protect – their neighbors belonging to different ethnic groups (Fujii, 2008). These interviews often provide access to information that does not exist in other forms and that plays a key role at various stages of the research process.

Yet, despite the potential benefits of interview-based research, its use in political science has been quite limited in recent years. As political scientists are increasingly drawn to using ‘big data’ (e.g. machine-learning techniques used to code large bodies of text), and as concerns regarding causal identification privilege some questions and approaches at the expense of others (Huber, 2013; Samii, 2016), one might question the utility of interviews as a research method. In addition, emerging norms of data access and research transparency (Lupia and Elman, 2014) may create tensions for researchers who use interview-based methods: researchers must respect the guarantees of confidentiality and anonymity typically offered to interviewees – whose livelihoods and even lives sometimes



depend on human subject protections – and yet researchers must also offer enough information about interview respondents and responses to allow reviewers, editors, and readers to evaluate the conclusions they draw from interviews (Moravcsik, 2014a; Wood, 2007). Moreover, access to interviewees is not always guaranteed: potential interviewees, especially public figures, may be hesitant to speak to social scientists, and recent high-profile breaches of research ethics by political scientists may intensify their concerns (Johnson, 2018; van Noorden, 2015).

Indeed, while interviews are acknowledged as being an excellent method for identifying causal mechanisms, as well as for being the only means for assessing some types of causal claims (Kapiszewski et al., 2015; Mosley, 2013), fewer than 10% of articles published in a sample of top political science journals between 2000 and 2017 incorporate interviews, as opposed to 58% using solely quantitative methods (Pepinsky, 2018). It is worth noting, however, that while qualitative empirical approaches have overall become less common in the most prestigious journals (Pepinsky, 2018), the use of interviews in these outlets has remained steady (albeit low; see below) since the early 2000s.

And, despite the challenges associated with using interviews, many scholars continue to rely on interview data to inform all stages of their work. This is not out of loyalty to an obsolete method but rather a reminder that interview data lend themselves to developing theories, supporting causal claims, and understanding complex social phenomena in a depth that other methods cannot match. Interviews often allow social scientists to examine the causal mechanisms that underlie the correlations revealed by statistical analyses. Interviews can also facilitate the design of better survey questions, as well as of better lab and field experiments. For those interested in process tracing, interviews are a valuable data source (Bennett and Checkel, 2014; Tansey, 2007). Indeed, interviews provide rich and oftentimes surprising data that

cannot be obtained in any other way, and, when used properly, they can shape and support work that is both rigorous and innovative.

Thus, in this chapter, we offer a case for continuing – or even expanding – the use of interviews in political science research. We start by providing a brief look at the utility of interviews as a methodological tool and discussing important data collection and reporting practices. Next, having established why and how scholars use interviews, we assess the current state of interview research in political science by offering evidence regarding the prevalence of interviews and how their usage varies across different subfields, journals, and books. Finally, we present an overview of the opportunities and challenges that will likely face scholars who employ interview research in the future. While we cannot resolve the tensions and tradeoffs inherent in the use of interviews, we can and do draw attention to them. In doing so, we hope to prompt more researchers to consider how interviews might enhance their own work, and, for those with an interest in the method, we aim to provide the basic information they need to use interviews in an appropriate and effective way.

## **INTERVIEWS: WHEN AND HOW TO USE THEM**

Other works examine the use of interviews in detail, providing a range of practical advice (Kapiszewski et al., 2015; Leech, 2002; Mosley, 2013; see also Rubin and Rubin, 2011). These analyses also discuss in detail how scholars of a wide range of political phenomena can productively use interview-based data. Although we offer a brief discussion of some key issues regarding interviews, we encourage those interested in interview methods to consult these works as well.

Interviews can be used at multiple stages in the research process. First, they can help scholars to generate causal claims at the start

of the research process, as when graduate students use pre-dissertation fieldwork to establish the plausibility of potential questions and theoretical approaches. For instance, a researcher might go into the field expecting to find a link between A and B, only to find that interviewees rarely mention A in connection with B, while C surfaces much more frequently. Interviews allow scholars to discover the lay of the land and establish an evidence-based direction for a novel research agenda (Anria, 2018; Greenstein, 2018; Niedzwiecki, 2018; Petrova, 2018; Vera-Adrianzen, 2017). A series of unstructured or semi-structured elite interviews may reveal a set of common factors that inspire the formulation of a plausible causal mechanism or an overarching theoretical framework. And interviews with members of the general public can provide insight into perceptions, behavior, and attitudes that indicate a link between two seemingly disparate areas of study.

Second, interviews can be employed in the theory-testing process to assess the veracity of proposed causal mechanisms and specific causal hypotheses. This is particularly true when the issues being studied are difficult (or impossible) to quantify or have not generated much quantitative data yet. For instance, researchers examining the social impact of a new environmental policy may not be able to quantify the outcomes they are interested in studying, but they can speak to affected individuals directly and gather data on their topic of interest that way. In fact, interviews can still be useful even if the relevant social phenomena can be quantified, because in some cases the newness of a policy means that researchers might have to wait years before there are enough data points to test a theory. However, rather than putting this project on hold, they could find answers to their research question by interviewing the bureaucrats who designed this policy, the legislators who voted to pass it, the staffers in charge of implementing it, and the businesses and individuals affected by it. Such interviews could serve to either test the researchers' proposed

causal story or to illustrate the mechanisms that are assumed to operate in this theory (which are then tested using other evidence). Provided that researchers are able to reach a significant number of diverse interview respondents – and sometimes this is a serious challenge – interviews could even be conducted in stages, with the first round of interviews being used for theory generation and the second for theory testing.

Indeed, thanks to the rich, complex data that interviews yield, interviews can actually be used to test theories more directly than many other methods can. When conducting an interview, researchers can ask specific questions about the exact event or issue that interests them, regardless of whether that thing happened in the past, the present, or has yet to occur at all. In cases where such blunt or open questioning is not possible, scholars can still overcome barriers to acquiring useful data by building a relationship with interviewees that facilitates openness or by phrasing and sequencing their questions in a way that yields pertinent, if veiled, information about the topic of study. This type of one-on-one interaction with people who participated in, observed, and were affected by particular policies, events, or processes allows researchers to collect data that are immediately relevant to the research question, to avoid the use of proxy measures, and to directly assess the validity of competing theories.

Compared with many other methods, interviews may offer comparatively easy access to pertinent information. Scholars who use big data must usually rely on information that other agencies or entities have collected over time, regardless of whether those data were obtained in a rigorous manner or whether the variables in the dataset correspond perfectly to the researchers' variable of interest. Researchers who use survey data might get lucky and acquire the resources they need to conduct a survey with their ideal questions and ideal sample population, but it is far more likely that they will have to rely on results from surveys that already exist, regardless of

how well that survey's questions and answers correspond to the researcher's objectives. Archival researchers, too, often must confront frustrating gaps between their research questions and the available data, particularly when their topic of study is one that did not generate an extensive paper trail. Interviews, in contrast, allow scholars to ask knowledgeable individuals a number of direct questions about specific topics and thereby directly test, support, confirm, or discard a working theory.

In the digital era, interviews may become an even more important source of information, as individuals – especially elites – worry about communicating sensitive information via email or text. In part, this reflects concerns about the security of communication and about the likelihood of data breaches, either intentional or accidental. It also reflects an awareness that written materials may well persist in archival form. As a result, individuals involved in political processes may prefer to communicate via voice or video chat or to discuss issues in person. These methods protect privacy, especially if they do not involve recorded interchanges, but they do not leave a trail for future archival researchers. Indeed, interviews almost certainly will continue to be the only way to access certain kinds of critical information, necessary to identify causal processes and evaluate causal claims.

Interview data also can be used to test theories in mixed-methods studies. Researchers can use interview data to suggest what patterns we should discover when analyzing large-*n* data. For instance, Mosley (2003) draws on interviews with investment professionals to test the expectation, based on theories of default risk, that only some types of government policies and political events should affect the pricing of sovereign bonds. She pairs these interview-based data with statistical analyses of the correlates of sovereign-bond interest rates.

Researchers also might use later-stage interviews to gain feedback on their proposed theories and causal mechanisms. This is somewhat distinct from theory testing, as

one does not want to 'lead the witness' in terms of biasing one's findings. But, to the extent that the researcher has already gathered substantial interview data to test her theory, she might be interested in the extent to which real-world experts find her theoretical claims plausible. There is useful information to be gained by people who shaped particular events and were affected by certain phenomena, if a theory aligns with their experience. Researchers could also ask interviewees what *they* think the causal story is: 'In your experience, what role has X played in Y?' 'Some of my other respondents have indicated that X influenced Y. What do you think?' 'You've mentioned X a lot; how important is it for causing Y?' These responses do not substitute for the use of broader data for theory testing, and researchers must be aware of challenges to the accuracy of interview data (Mosley, 2013). But keeping in mind the limits of positionality, fallible memories, and respondents' desire to feel helpful or appear amiable, researchers can nonetheless use answers to these questions to gain insight from how later-stage interviewees perceive their political worlds. Moreover, sharing the findings from one's study also may be a modest means of 'giving back' to the communities in which one conducts research: participants gain a sense not only of what the researcher's specific questions are but also of the broader way in which answers to these questions inform social science research.

Third, interviews can supply data that help to justify the assumptions upon which a study is based, as when interviews offer insights into how individuals understand and interpret their experiences vis-a-vis politics (e.g. Fujii, 2010; Soss, 2015). Or, for researchers who employ formal models to analyze strategic interactions, interviews can reveal the motives of key actors – civilians who do or do not elect to join rebel groups, for instance (e.g. Baczko et al., 2018; Henshaw et al., 2019; Weinstein, 2007). Along these lines, interviews can be particularly useful when studying relatively new topics (e.g. the

effects of social media on mobilization and protest, how labor unions organize workers in the gig economy or service sector, or the ways in which gender, race, and sexual orientation intersect to affect legislators' policymaking effectiveness). In such situations, established literature may be of limited use, and the generation of informed theories and testable hypotheses may benefit from – or even depend on – the 'soaking and poking' aspect of interview-based work. Interviews offer concrete data that, in the absence of established literature, back up the researcher's claims and help justify certain theoretical and methodological choices.

Both positivist and post-positivist researchers can utilize interviews, as described above. Differences in epistemological perspectives, however, imply that different scholars will approach interviews differently. Post-positivist scholars tend to note the subjective nature of interview (and other) data. That is, interview evidence reflects the particularities of the researcher who collected it: a female graduate student from an elite US university, conducting interviews in a foreign culture, will receive a different set of answers than a male full professor from a European university, conducting interviews in his home culture. Post-positivists argue that while these differences render 'objective' analysis quite difficult, they are themselves informative. That is, they reflect how positionality affects the research process, as well as how interviewees understand their position relative to that of the researcher. Furthermore, even though the responses are necessarily influenced by the identity of the interviewer, the interview data do provide a range of viewpoints that a researcher can then use to show the diversity (or similarity) of thought on a given issue by a given set of respondents.

From a positivist perspective, on the other hand, interviews can be used to generate empirical data, which, while not free from the possibility of measurement error (Mosley, 2013), offer opportunities for objective analysis. Positivist scholars should still

pay attention to how their data-gathering strategies affect their responses (Bleich and Pekkanen, 2013), but they tend to worry less about how their gender, ethnicity, or professional status might color the responses they receive. They would assume that different researchers asking the same questions of the same individuals would receive relatively similar responses.

This is not to say that a positivist approach to interviews ignores the challenges of collecting and analyzing interview data. Certainly, information acquired from interviews may not paint the full picture. But this limitation applies to every type of data: archival evidence assumes that materials have been preserved in a representative fashion, but this is not necessarily the case. Archival materials may have been created in ways that bias the historical record; conflicts, natural disasters, and resource constraints may limit the future availability of such materials. Survey data, especially that collected by third parties, are limited by low response rates, the absence of directly pertinent questions, or poorly conceptualized items. And while focus groups may offer opportunities to understand the role of deliberation (Karpowitz et al., 2014) in group decision-making, they may offer less insight at the individual level: some participants will talk much more than others, and participants' stated (versus underlying) opinions may shift as a result of the group's conversation. One also could express concerns about the reliability and validity of work based on statistical analyses. While large-*n* analyses allow one to make claims regarding broader causal processes, they require the accurate operationalization of oftentimes complex concepts, as well as attention to the challenges of causal identification. Indeed, to the extent that the use of observationally generated quantitative data presents inferential challenges (e.g. Samii, 2016), some scholars may eschew their use in answering some questions. This, again, may represent an opportunity for using interview data instead.

Another means of addressing the measurement and inferential challenges associated with quantitative analysis is the use of mixed-methods approaches. Ideally, the advantages provided by each type of method – formal, statistical, or qualitative; observational or experimental – compensate for each method’s disadvantages. Researchers are then able to triangulate evidence, offering stronger support for a theory: interview data paired with archival data permit systematic process tracing; interview data and statistical data complement each other by offering both causal and correlative support for a given theory; interview data and formal models combine to illustrate both the experience of and the logic behind a given phenomenon.

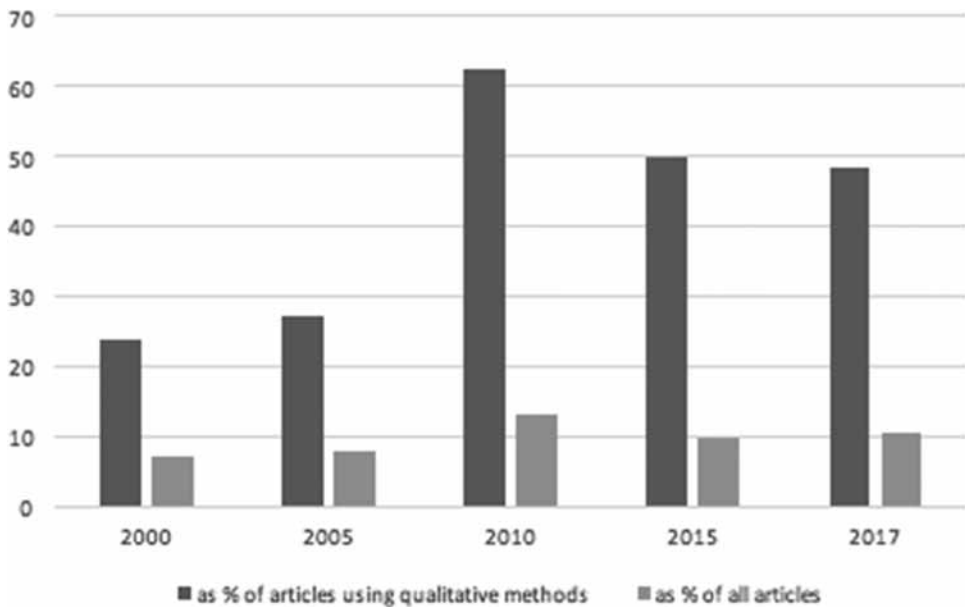
For instance, Bunte (2019) posits that domestic political coalitions help explain the ways in which governments of developing countries access credit – do they borrow from the Chinese government or from the World Bank? His cross-national time-series analyses offer some support for his claims, but they rely on somewhat rough measures of domestic interest groups; to complement this evidence, he deploys an extensive set of interview data, based on fieldwork in three Latin American countries. Taken together, these two forms of evidence offer consistent support for his hypotheses. Similarly, Brooks (2009) draws on interviews with national policymakers, private investors, and staff members of intergovernmental organizations to test her claims about the conditions under which Latin American governments reform their social security systems. Again, these interviews offer evidence that is largely consistent with the quantitative data and statistical analyses she also reports. As we show in the next section, many scholars are already using interviews in conjunction with other methods, but, unfortunately, this applies mainly to books, not to journal articles. Indeed, the relatively low incidence of interview data in top journal publications indicates that the discipline could do a better job of capitalizing on the advantages of interview data.

## THE USE OF INTERVIEWS IN POLITICAL SCIENCE

To what extent are interviews used by political scientists? We answer this question by drawing on data from Obermeier and Pepinsky’s 2018 analysis of articles appearing in six top discipline and subfield journals (*American Journal of Political Science*, *American Political Science Review*, *The Journal of Politics*, *Comparative Politics*, *Comparative Political Studies*, and *World Politics*). In their study, Obermeier and Pepinsky (2018) coded the methods – quantitative, qualitative, and experimental – employed in every article published during 2017, as well as those published at five-year intervals from 1965 to 2015. Because we are interested in the more recent use of interviews in political science research, we limit our analysis to the articles published in 2000, 2005, 2010, 2015, and 2017. Of the 1,269 total articles included in these six journals for these years, Obermeier and Pepinsky identify 297 as using qualitative methods. A significant proportion of these use a mixed-methods design, as 139 of these 297 also employ formal, quantitative, or experimental techniques.

For each article coded as employing qualitative methods, we consult the article to determine whether it uses interviews as all or part of its empirical strategy. We identify 125 articles – 42% of those using qualitative approaches but only 9.8% of the total articles published during those years – as using interviews in some capacity. By subfield, 111 of these are in comparative politics, 12 in international relations, and two in American politics.

Figure 61.1 plots these articles as a percentage of all articles in Obermeier and Pepinsky’s sample for a given year and as a percentage of the subset of articles that were coded as using qualitative methods. It is perhaps not surprising that under the broad umbrella of qualitative methodologies, interviews are an important evidentiary tool. It is perhaps more surprising to find that even as much of the discipline has turned its attention



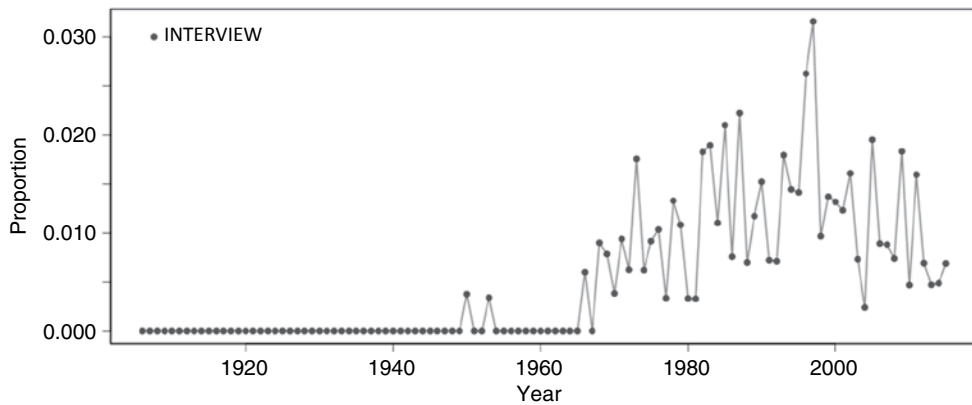
**Figure 61.1 Interviews in political science journals, 2000–17**

to survey and field experiments and the use of ‘big data’, the use of interviews has remained fairly steady – as a proportion of total articles – across this period. Among articles using qualitative methods in part or in whole, scholars used interviews much more often in the last decade – that is, in the years 2010, 2015, and 2017 – than in 2000 and 2005.

It is worth noting that, regardless of the year, interviews are used in only a small percentage of the *total* articles published in these six top journals. Their prevalence ranges from 7% in 2000 to a high of 13% in 2010. In 2017, the most recent year examined, interviews were used in 10% of the published articles. Moreover, this estimate is likely higher than what is extant in the discipline overall, given that interviews are most commonly used in the comparative politics subfield and that two of the journals in the dataset focus on comparative politics. Thus, it is reasonable to assume that the percentage of overall articles using interviews would be even lower were we to focus our analysis on journals in other subfields or on other discipline-wide outlets.

Certainly, these data paint only a partial picture. One could use a broader set of journals – such as those that receive greater numbers of submissions employing qualitative methods and/or additional journals with high impact factors – and potentially find different patterns. We therefore also use data from Wilson (2017), who relies on a longer time frame (he categorizes articles from 1906 through 2015), and who includes a larger number of top-tier journals (he codes the six journals analyzed in Obermeier and Pepinsky, 2018, as well as *International Organization* and the *British Journal of Political Science*). The dataset’s associated search engine<sup>2</sup> relies only on article titles and abstracts to identify terms used in these publications, which may mean that some articles end up being misclassified. Nonetheless, it identifies 175 articles (from a total of 25,845) containing the word ‘interview’. These results show a clear temporal trend in the use of interview data: only a few of the hits occur prior to the mid-1960s (Figure 61.2). Of course, what is less clear from this evidence is the extent to which interviews form a core part

The term 'INTERVIEW' appears 175 times out of 25,845 articles.



**Figure 61.2** Longer-term patterns in political science articles

of each article's empirical methodology; they might be used for vignettes rather than as key sources of data. One also might imagine that some pieces employing interviews do not mention these in their abstracts and therefore do not appear in the search results.

Even so, these data indicate that interviews are relatively uncommon in journal articles. While it is impossible to pinpoint exactly why this trend exists, we do know that using interview data in articles faces hurdles on both the researcher's side and the editor's side of the publication process. One challenge for researchers who rely on interviews is that it is often difficult for qualitatively-focused scholars to present information from, for instance, country case studies within the confines of journal word limits. Thus, we might expect to find interview data being used more frequently in books than in articles. Interviews also might be less frequently used in journal articles if it is the case that journal reviewers are more skeptical about the conclusions drawn from interview-based evidence, especially when these reviewers are unable – generally for human subject protection reasons – to access the full interview transcripts or recordings and so are prevented from validating the ways in which authors draw conclusions

from interview materials. More broadly, the attention given to causal identification in recent years creates additional hurdles to work using all sorts of observational data, including interviews, because such data are (rightly or wrongly) less likely to be seen as “rigorous”. These concerns may prevent authors from submitting work that relies partly or entirely on interviews. And, to the extent that scholars' records are evaluated based on their placement of articles in top discipline-wide or subfield journals, the perceived difficulties with placing interview-based work in journals may further deter scholars – especially graduate students and untenured faculty – from relying on (or even utilizing) interview-based approaches.

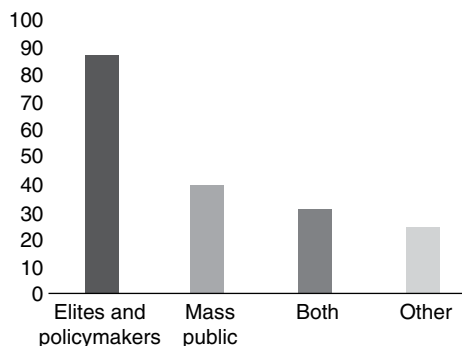
To gain a different sense of the use of interviews in political science, we also collected information on the use of interview methods in political science books published by university presses. For this analysis, we aimed not to measure the use of interview methods as a percentage of all methods or all published works – which one certainly could do, given sufficient resources to code a wide set of monographs – but rather to gain a sense of the ways in which scholars tend to deploy interviews in book-length projects. We recruited

political scientists to complete a short survey, using Facebook and Twitter to disseminate the link. Participants were recruited in two waves, in September 2018 and December 2018–January 2019. The recruitment posts were aimed at scholars who had published, or had in press, a political science book using interviews; we specified 2016–18 as the relevant time frame.<sup>3</sup> Those who opted into the study completed a short survey via Qualtrics, asking for the book's title and year of publication and about its use of interview-based evidence.

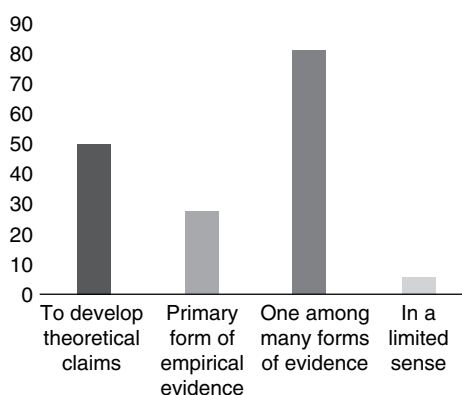
Our survey yielded 84 books. Based on our categorizations, the books fell mostly into the comparative politics subfield (n=51, 61%), although the American politics (n=9, 11%) and international relations (n=24, 29%) subfields are better represented among this set of books than in the article data above. The topics of these books are quite diverse, ranging from US state legislatures to post-conflict peacebuilding and from the use of drones to transnational labor activism.

We asked respondents whom they interviewed, as well as how they used interviews. The responses are summarized in Figures 61.3 and 61.4.<sup>4</sup> Elites and policymakers are the most common type of interviewees, used in 87% of the books reported; 39% of books rely on interviews of members of the mass public, and, of course, this implies that some books – approximately 31% – draw on both types of interviews.

We also asked survey respondents about the ways in which their books use interview evidence. The responses are summarized in Figure 61.4. Interviews are most commonly employed as part of a multi-method evidentiary strategy: over 80% of respondents categorize their use of interviews as 'one among many forms of evidence'. It also is quite common – representing half of the books described – for scholars to use interviews to develop theoretical claims. In 43% of cases, authors noted that they use interviews both to develop theory and hypotheses and as one type of data among many. What is less common, at



**Figure 61.3 Interviews in political science books, % of responses**



**Figure 61.4 Uses of interviews in political science books, % of responses**

least among those responding to our survey, is a reliance on interviews as the sole form of empirical evidence (27% of responses) or in a more limited – vignette or 'window dressing' – sense (6%). That interviews are rarely used in this very limited sense is quite striking, given that interviews are sometimes criticized as being useful for eye-catching anecdotes but for little else. The responses to our survey contradict this viewpoint, as our data indicate that scholars who use interviews are overwhelmingly using interviews to collect systematic empirical data.

Our survey of books is based only on those who responded to our call for participation.



It does appear, however, to confirm our supposition that interview-based evidence tends to appear more frequently in book-length rather than article-length publications.<sup>5</sup> This belief is further supported by data from Cambridge University Press' 2017 book catalog for politics, which consisted of books published in 2016 and 2017 (and one in 2014). According to our coding, out of 285 books offered in that year's catalog, 117 of them (41%) include interviews that were conducted by the author, and, echoing the result of the survey, these interviews were overwhelmingly used as a primary or complementary source of data, not as a source of vignettes.

Our analysis of the use of interviews suggests, then, that they are important within research that uses qualitative approaches but that qualitative-based approaches often do not feature prominently in top discipline-wide journal outlets. Our data are limited to a small set of journals (albeit with high impact factors), and we might find a greater overall prevalence of interview-focused methods were we to examine a wider set of publications. Moreover, interviews may inform the theory development reflected in articles, without being reported as a direct source of evidence. For instance, Ballard-Rosa et al.'s (2019) analysis of sovereign borrowing and domestic political institutions relies heavily on a large-n dataset of government bond issues. But the expectations developed in that piece – which link investors' assessments of political institutions with government debt managers' desires to access credit – draw significantly on interviews with investment professionals and government officials.

Yet our data likely reflect more than just an under-counting of interview-based methods in journal articles. More importantly, they reflect some of the (real or perceived) challenges associated with using interview-based approaches in article-length pieces and with finding success for those articles in the peer-review publication process. These challenges appear to be less of an impediment for book-length work. But given the overall shift

away from books and toward articles, we might worry that political science researchers are missing opportunities to deploy what is often an effective research method. In the next section, therefore, we describe some of the challenges associated with interview-based research, and we suggest some means of addressing or mitigating those challenges.

### **CHALLENGES OF INTERVIEWS: PERENNIAL PROBLEMS AND CONTEMPORARY CONCERNS**

While interviews offer researchers a flexible and versatile tool for understanding a range of political phenomena, their use is not without challenges. Some of these, such as the difficulty of acquiring proper training, the cost of interview research, and the ethics of working with vulnerable populations, are decades old. Others, such as concerns about striking a balance between confidentiality and human subjects protections versus data access and research transparency, have become more pronounced in recent years. In this section, we discuss some of these challenges; offer strategies for addressing the concrete hurdles, such as cost; and provide a look at how the discipline as a whole has addressed debates on the pros and cons of interviews, such as how to weigh privacy and human subjects protections, on the one hand, against data transparency, replication, and verification, on the other.

#### ***Reliability, Validity, and Verification***

First, as perhaps evidenced by the limited prevalence of interview evidence in many discipline-wide journals, some audiences worry about the validity and reliability of interview evidence (see also Mosley, 2013). Indeed, interviews may be brushed off as little more than anecdotes, selected to support a

researcher's framing or argument. This reaction may be explained, in part, by the use – although limited in recent years, given the data presented in Figure 61.4 – of interviews as 'window dressing'. Pithy pull quotes that catch a reader's eye, often selected because they are clever or thought-provoking, also raise concerns about representativeness, either of the entire interview with that individual or of the broader set of individuals interviewed for a project. While the desire to provide color for a project – especially one that may also use more typical forms of data – is understandable, it is also incumbent upon researchers to address concerns about how their interview data are collected, how the data are useful for crafting theories and testing hypotheses, and where potential sources of bias might exist.

While certain corners of the discipline might still protest that interviews are not systematic, comprehensive, or reliable in the way that quantitative may – or may not – be, researchers can increase confidence in their analyses by adhering to best practices regarding the collection, description, and analysis of interview data. Just as scholars who use quantitative methods are expected to refrain from p-hacking and reporting a rare specification that yields the desired results, to provide a clear description of how various concepts are operationalized and coded, and to supply the data and code used to replicate analyses, scholars using interviews should give as much information as possible (within the confines of human subjects protections) regarding their empirical strategies. While interview data will almost always be observational in nature, scholars can nonetheless address concerns about their inferential validity.

One important step in increasing the reliability of interview data is ensuring that respondents are selected as systematically and comprehensively as befits the goals of the project. Bleich and Pekkanen (2013) offer a model for this in their 'Interview Methods Appendix', which reminds scholars that when designing a study, it is important to consider what types of individuals will be

interviewed. There are many different ways to sample subjects (see Kapiszewski et al., 2018; Lynch, 2013; Martin, 2013), and one's analytical goals should drive one's choice of sampling strategies. The IRB process, a required component of interview-based research for scholars based at US universities (and sometimes beyond), also requires some reflection on how interviewees are chosen.<sup>6</sup>

Bleich and Pekkanen (2013) offer guidance on how to organize the selection of potential interviewees, as well as how to report the process by which interviewees become part of the dataset. They suggest that potential interviewees could be categorized by their profession or position within a given organization ('central banker', 'legislative staffer'), by the interests they represented in a given policy debate ('advocates for a policy' or 'opponents of a policy'), by their country or region of origin, or according to many other attributes that align with a study's theoretical claims. In some cases, a project may necessitate interviewing specific individuals rather than simply interviewing anyone who is a member of a certain category. Researchers often impose selection criteria when defining their sampling strategies; the point is to offer a transparent discussion of these strategies (as well as how effective they were in terms of efforts to recruit interviewees and problems of non-response to interview invitations).

Of course, one's project may not lend itself to specifying a comprehensive set of interviewees *ex ante*. When interviewing elites, a scholar may not always know which individuals were relevant to a given policy decision, and when interviewing members of the mass public, a scholar may rely somewhat on convenience or snowball sampling, which involves gaining the trust of a growing circle of individuals along the way. It is also particularly difficult to identify specific interviewees in advance when one's research occurs in conflict-prone or non-democratic settings or when one's research question involves sensitive topics such as anti-regime activities or sexual violence (e.g. Fujii, 2010).

But even when these concerns affect *ex ante* research design, the researcher should offer an *ex post* summary of with whom she spoke, with whom she attempted to speak but was unable to arrange an interview (the non-respondents), and how each of these individuals fits the broader aims of the empirical analyses. Beyond that, the topic and goals of each specific project should guide what additional types of interview metadata are relevant and necessary to report. For instance, a study on women's health policies should disclose what percentage of respondents were women, whereas the researcher might decide that the gender distribution of interviewees is less relevant for an analysis of how political elites advocate for party positioning on various economic issues. As a baseline, though, researchers should discuss the methods by which their sample of interviewees was recruited, including whether convenience and personal connections played a role.

Scholars also should describe, either in the main text or in a readily available supplementary appendix, the mechanics of the interviews. Assuming a semi-structured or structured approach, these metadata ought to include the questions asked of interviewees. This allows readers to assess whether respondents had incentives to misrepresent the facts or to paint past events in a more favorable light. Next, scholars should report how the interview was recorded (written notes, audio, video, or some combination), the length of each interview, and the date on which the interview occurred (one might expect that interview data could be affected by important political events). Scholars also might note how they sought to ensure the reliability and validity of interview data; for instance, did they consider using text analysis to identify systematic patterns in interview transcripts? Or did the researcher check the accuracy of interview responses – in cases where a given individual's actions in a specific instance are a matter of public or journalistic record – against third-party data?

Although we often think of interview data as being presented in the form of quotations

from the interviews, summary statistics of the mechanics and content of interviews can be helpful as well. Presenting interview data in this summary format also has the advantage of making it both easy and quick for readers to assess the reliability and validity of the data. For instance, if a medium-to-large set of individuals of a similar type are included in the sample, scholars might provide overall counts and percentages regarding the content of those interviews. For instance, what proportion of legislative staffers identified mentioned the importance of NGO or business sector lobbying? Or, when asked to describe their time horizon for making investments in developing countries, what was the distribution and mean of responses given by professional investors (Mosley, 2003)? Such summaries also allow the author to put into context specific comments or anecdotes that are drawn from individual interviews.

Establishing confidence in the reliability and validity of interview data is also helped by reflecting upon and reporting one's position relative to one's interviewees (Cammatt, 2013; Soss, 2015). We know that female scholars receive different answers than male researchers, that respondents sometimes offer different answers to co-ethnics than to perceived outsiders, and that one's identity as, for instance, a 'professor from a US university' affects both the access and answers that interviewees receive as they conduct their project (Kapiszewski et al., 2018; Su, 2017; Wedeen, 2010). It is true that the attributes of the scholar are somewhat fixed, and researchers may not know how they are being perceived or judged by interviewees. But even if we do not know exactly *how* positionality is affecting the data, it is important to acknowledge that it *is*. Thus, an engagement with these issues signals to readers that scholars are aware of context-specific challenges to reliability and validity, and, by providing data on these factors, readers will be better able to understand and assess how the data-gathering context has influenced the results (or not).

Citations of specific interview-based evidence offer another avenue for researchers to reassure readers about the credibility of the data presented. For instance, a citation might include a description of the interviewee's type and the date of the interview, thereby signaling the competence of the interviewee without compromising their identity (if guarantees of confidentiality or anonymity were given). Then, when specific passages draw on specific interviews, the interview can be referenced by number or by using such descriptors. These citations could also point to a repository (embargoed for a period of time) for interview data.

For their part, journal editors and reviewers should bear in mind the inherent tension between the protection of human subjects, on the one hand, and data access and research transparency (DA-RT), on the other.<sup>7</sup> While authors should provide some information about how they collect interview material, they are typically unable to reveal interviewees' identities or to provide reviewers with full transcripts from interviews. To do so would usually violate ethical principles for conducting human subjects research, especially when interviewees might suffer professionally or personally were information to be shared with a broader audience. Moreover, violating these guarantees would almost certainly make potential interviewees less likely to agree to be interviewed by other political scientists in the future.

These considerations reveal a fundamental tension between the discipline's move toward data access and research transparency, on the one hand, and research ethics, on the other. This tension would be difficult to resolve under the best of circumstances; doing so is rendered more difficult by the fact that, as a profession, political science has created very few formal ethical guidelines for itself. For instance, the latest (2012) iteration of the American Political Science Association's *Guide to Professional Ethics in Political Science* offers a very limited discussion – only a few sentences in length – of scholars'

responsibility to protect individuals and communities or of how political scientists ought to weight the benefits of social scientific research against the potential harms to individuals, groups, and democratic processes (e.g. Desposato, 2015). Silence on ethical issues (also see below) matters not only for interview-based research but also for field experiments, which include real-world interventions. Indeed, in 2015, the then-APSA president David Lake appointed a committee to further develop guidelines for ethical research, but given that such guidelines have not yet become part of the discipline's norms and practices (or part of what informs reviewers and editors), it is unsurprising that researchers generally struggle to balance the call for more open presentation of research materials with a concern about the confidentiality of interviewees' responses.

To a large extent, this tension is more about the data-access component – versus the research-transparency element – of the general call for scholarly openness (Bütte and Jacobs, 2015; Lupia and Elman, 2014). In most cases, although certainly not all, it is fairly easy for scholars who use interviews to be transparent about their study design and protocols without compromising the identity or the responses of study participants. The thornier question comes on the side of data access: to what extent should scholars make available, for instance, notes from or transcripts of their interviews? Granting wider access to the data might increase reviewers' confidence in the conclusions being drawn (Moravcsik, 2014a, 2014b), but it almost always violates the conditions under which interviewees consent to an interview study. Therefore, defining what DA-RT means in the interview context remains an important challenge.<sup>8</sup>

For those seeking to publish their works in political science journals, there may be a middle ground between transparency and violating human subject protections. As a condition of an article's acceptance, qualitative researchers could be asked to submit their work for verification to a trusted third party. The verifying

entity – for instance, the Qualitative Data Repository (QDR) at Syracuse University<sup>9</sup> – would be given confidential access to all materials (including interview transcripts), and the author could indicate how these underlying materials were used to generate the claims made in the manuscript. QDR could produce a verification report, shared with journal editors prior to publication, to which the authors would respond, addressing any concerns about the extent to which the interview evidence supported their claims. QDR could also hold these materials in a digital repository, where the typical provisions would include not releasing the data for 50 years. Of course, scholars would need to receive clearance from the Institutional Review Board (IRB) to take any of these steps, which may complicate matters for researchers who want to use data they collected before such norms were in place, but journals could consider making exceptions for work that uses older data.

The QDR is, in some ways, analogous to the third-party verifications of quantitative analyses that already take place in political science when the dataset contains proprietary information. In those circumstances, the author provides the verifying entity (e.g. the Odum Institute at the University of North Carolina at Chapel Hill, in the case of the *American Journal of Political Science*) with the complete dataset, as well as the required code for analysis.<sup>10</sup> But when the replication materials are posted, variables representing proprietary data are not included (a protocol that is clearly communicated both to authors and readers). Indeed, the *American Journal of Political Science*, which has required verification of the results reported in empirical articles since 2014, in 2019 submitted its first manuscript to QDR for verification and archiving. Although the article in question draws much of its data from archival rather than interview materials, it does, like most interview-based articles, use qualitative evidence to test its primary hypotheses, and it reports those data largely in a supplementary appendix (Carnegie and Carson, 2019).

It is important to note that scholars may need to anticipate the use of qualitative verification and repository mechanisms: when asking interviewees for informed consent and when filing IRB applications, they would have to indicate that materials might be confidentially shared with a replicator and perhaps embargoed in a repository. Still, despite the additional up-front work that would be necessary in order to acquire approval for these measures, such verification and depositing could allay some concerns regarding the accurate use of interview-based data.

### ***Interview Training***

Students in political science graduate programs in the United States often note that their methods-related training gives less attention to qualitative than to formal and quantitative methods. As new political scientists are often expected to have fairly extensive training in statistical and formal methods, less time is left for coursework in techniques such as archival analysis and interviewing (not to mention thematic courses in students' major and minor subfields). Students who intend to engage in fieldwork in other countries also may find it challenging to develop foreign-language skills, particularly given most programs' assistantship requirements and the pace at which they are expected to produce publishable material (Kapiszewski et al., 2015).

Still, there is certainly the potential to incorporate interview methods courses into graduate programs – after all, political methodology now exists as a distinct subfield alongside substantive subfields in many US political science departments. New journals devoted solely to articles on political methodology also have arisen; in addition to *Political Analysis*, which published its first issue in 1989, scholars can now publish exclusively methodological articles in *Political Science Research and Methods* (started in 2013) and *The Journal of Experimental Political Science* (started in 2014). However, 'methods' is

often a shorthand for *quantitative* methods, and classes in qualitative methods – including interviewing – remain rare in political science departments.

Graduate students therefore may be less inclined to consider the use of interview-based approaches – meaning that the trends described in the first section of this chapter could be reflecting supply as well as demand – when they design their own research projects. Alternatively, students who do decide to employ interview methods may receive little formal training, being told instead to ‘just talk to people’ (Mosley, 2013). Of course, some of the skill in conducting interviews and using interview evidence is developed during the interviewing process. But there is much that students can be taught and many issues that they should be reminded to consider when designing and planning an interview-based project. For instance, some interviewing strategies work better than others (and some things that work for, e.g., a tenured male academic, may be less effective for a female graduate student), and there are best practices that greatly enhance the quality and usability of the data collected, too. These practices are not mysteries; they are strategies that professors can and should communicate to graduate students.

Given that there are better and worse ways to conduct interview research, it is detrimental to the discipline as a whole to advise researchers to ‘just go talk to people’, leaving them to work out the rest through trial and error. It is very unlikely that a student who wanted to use Bayesian statistics or machine learning would be advised to figure out how to do things once they had collected data for their job-market paper or that a graduate student interested in strategic interaction would be expected to write a formal model of crisis bargaining without any graduate coursework. Interviewing is a methodological tool like any other, and political science students should have the opportunity to learn how to do it in a political science department.

Of course, some political science departments do offer courses on interviewing, and

we strongly encourage such offerings. But most graduate students who wish to use interviews are left somewhat to their own devices. Sometimes they look to anthropology, sociology, public health, or related fields for interview-related training, which, while beneficial, still might not adequately prepare graduate students for their fieldwork. Much political science interviewing involves discipline-specific considerations, as may be the case when interacting with past policymakers, who may have incentives to provide overly favorable accounts of their actions and decisions, or when asking opposition party members about anti-regime protests in non-democratic contexts. Similarly, the ways in which interview data are collected, recorded, reported, and interpreted vary across disciplines; a political science graduate student may require additional training in how best to develop a rapport with interviewees, given that she may not spend large amounts of time living and working at the fieldwork site. Thus, we emphasize the need for including interview training in political science departments.

That said, given the time and course constraints facing both graduate students and faculty, it might be easier to provide interview training as part of a broader course on qualitative or field research methods instead of as a stand-alone course. Regardless, in addition to preparing graduate students to conduct their research, offering some form of interview training helps signal that not only do faculty expect emerging scholars to embrace principles of theory development, hypothesis testing, and causal inference, but also that they acknowledge that methodological choices should be driven by the question at hand rather than the reverse. Implementing these curricular and attitudinal changes may be more difficult in some places than in others, but doing so would improve the quality of interview research and enhance the discipline as a whole.

To the extent that interview training remains unavailable in many political science departments, however, graduate students (as

well as faculty) who aspire to add interview-based components to their empirical analyses should be encouraged to seek training from other sources. These include not only standard course offerings in other disciplines but also short courses (offered by university-based social science data institutions, as well as at the APSA annual meeting) and summer institutes (e.g. The Institute for Qualitative and Multi-Method Research's summer program includes a module on interviewing).

### ***Time and Money***

Another potential obstacle to conducting interview research is the cost of traveling to speak with people in person. The time and money – as well as the impact on the earth's climate – that must be expended in order to travel to meet with potential respondents can be overwhelming, especially for graduate students and other scholars without generous research budgets. Scholars based in the United States, for instance, may find that the best times (given their teaching schedules) for research-related travel also feature the highest airfares.

These challenges are especially daunting for graduate students, who may have to finance their research trips personally. Graduate students in comparative politics are particularly likely to face this problem, given that the subfield values fieldwork of some sort or another. Dissertation committees, journal referees, and hiring committees often reward researchers who can show that they have visited the countries they study and met with individuals involved with the topics at hand. This fieldwork is expected to consist of several months, or at least several weeks, in a given country, which often means that students must take a leave of absence from their home university for a semester or travel during the summer. Graduate students therefore may find that in the absence of grant funding, they need to simultaneously finance fieldwork and take leave from income-generating teaching.

Moreover, graduate students may face contending pressures to conduct qualitative research abroad but also to limit (compared with prior generations) the time it takes to complete their PhD. This pressure, coupled with more limited area-studies training, may disadvantage graduate students in political science (relative to graduate students in other social science and humanistic disciplines) when competing for fieldwork funding (Agarwala and Teitelbaum, 2010). Although untenured faculty generally receive more compensation and research funding than graduate students, they are still likely to face challenges in finding material resources and time for extended research trips. Even tenured faculty may find it hard to travel for interview research, given their administrative, advising, and/or teaching responsibilities – as well as their desire, sometimes, to limit travel in order to minimize time away from partners, children, and aging relatives.

Field experience is, of course, invaluable; it often allows scholars to discover things and to develop a sense of causal processes that cannot be gained via secondary sources. In some contexts, interviews cannot be arranged prior to one's arrival at the field site; if one needs to identify possible interviewees once on the ground or build trust with community leaders in order to gain access (Reno, 2013), then travel is almost certainly necessary. And, while the interview itself is a key source of data, so is the surrounding metadata: how does an interviewee introduce themselves to the interviewer, or how does an interviewee relate to other colleagues at the office? To what extent is the interviewer welcomed into the community or received with suspicion, given their nationality or perceived professional background?

Nevertheless, not only *can* researchers with limited funds still conduct interview research, but they *should* also be encouraged to do so if interviews are a suitable method for developing or evaluating their research question. In some instances, creative thinking can help to reduce travel costs. For instance,

a researcher studying foreign direct investment might consider attending a conference targeted at professionals who work on investment promotion. The conference sessions themselves can provide useful background information, and the researcher can set up meetings with (and perhaps surveys of) individuals from a variety of countries, who will all be in the same place for the meeting (e.g. Bauerle Danzman, 2017). And when academic conferences are held in the same location as potential interviewees – for instance, for scholars of the US Congress or of international financial institutions when APSA is held in Washington, DC – it may be possible to schedule in-person meetings.

Technology also can enable further use of interviews. Phone, Skype, and FaceTime interviews can be excellent alternatives, particularly when time and money are tight. Certainly, making explicit the process of informed consent and developing clear expectations about whether the interaction is recorded are especially important in these settings, and researchers may need to work harder to develop an interpersonal rapport with interviewees. Still, these sorts of distance interviews are almost always better than no interviews. They also may be used to ask follow-up questions once the researcher has returned home from a face-to-face interview, had time to process the data, and identified a point that needs further clarification.

Such workarounds may not be appropriate or feasible for all research questions, of course. Depending on the country and the type of individual being interviewed, potential interviewees may have limited phone and/or internet access. Elite respondents might have their contact information online, but arranging meetings with them may require navigating their administrative staff. By contrast, mid- and lower-level government staffers may be difficult to identify and locate from afar. The same may be true for members of the general public, although social media could facilitate the identification of such individuals (Côté, 2013). Conducting

interviews in a second or third language also may be harder over the phone, and the absence of body language can complicate attempts to translate across cultural divides. These problems should not be taken lightly, and, in some cases, they may preclude gathering satisfactory data without in-person interviews. However, many researchers will find that they can maintain the integrity and quality of their research without always having to expend the time and money that traditional, in-person interview research entails.

## ETHICAL CONSIDERATIONS

When designing any empirical research project, political scientists should consider the ethical issues it presents: how, for instance, would providing information to some individuals about candidates' voting records affect not only individual voting behavior, but also an election outcome (see Desposato, 2015)? Or how might priming individuals to focus on racial resentment in the context of a survey have effects for future political and social behavior? Interviews also might retraumatize participants who are asked to recount violent events or inadvertently endanger respondents from non-democratic political systems if answers to sensitive questions are later shared with a broader audience. These concerns are not to be taken lightly, and it is the responsibility of researchers to seriously consider the potential harm that their work could cause to interview participants (Wood, 2007).

Often, these types of ethical considerations are first outlined and addressed in the context of acquiring formal approval for a research project. Researchers at academic institutions in the United States must typically acquire approval for interview work from their school's IRB, which reviews all research involving human subjects. Other countries, as well as some subnational entities, also have IRBs or their equivalents,



meaning that interview-based research projects may require review by multiple bodies. It is often at the point of completing human subjects research certification, or in the process of creating documents for IRB review, that scholars consider the requirements for informed consent, as well as how best to protect participants' confidentiality and/or anonymity (Brooks, 2013).

University-based IRB protocols are designed to ensure that studies meet federal ethical guidelines; for academic institutions, it is the threat of withdrawal of federal funding or other government sanctions that typically motivates the IRB process. In this context, researchers are asked to document how they plan to minimize or eliminate the potential harm that interviewees could incur as a result of agreeing to participate in a research project. IRBs vary significantly in how they approach interview-based projects: for instance, some require the scholar to list a specific set of individuals to be interviewed as part of the approval process. Other IRBs allow scholars to specify types of interviewees, without providing specific names *ex ante*.

It is important to remember, however, that the IRB process does not exhaust the limits of a researcher's ethical obligations. IRBs are designed primarily to protect the university and the researcher from legal liability, not to ensure the absolute wellbeing of research participants. Additionally, despite the best intentions of IRB members, these individuals cannot possibly be familiar with the specific research context in which every proposed research project will occur. This, in turn, necessitates a certain amount of transparency on the part of the researcher; if researchers are not forthcoming, IRBs may not be fully aware of the potential hazards to subjects involved.

Thus, it is the duty of each individual researcher (or, in some cases, each member of a larger research team) to familiarize themselves with the context in which the interviews will be taking place and to do their due diligence in safeguarding their respondents' wellbeing. The precautions necessary to

ensure respondents' safety may change over the course of a study; topics that were banal at the start of a research project can become flashpoints as a result of unexpected events, or a researcher can realize that they have misjudged the sensitivity of their research topic and need to adjust their questions or data-security measures accordingly. In such situations, it is incumbent upon the researcher to move beyond the protective measures that were included in the original, IRB-approved research protocol and ensure that respondents are not subjected to physical or emotional harm by granting an interview. This is particularly true when the research involves interviewing members of vulnerable groups – for instance, victims of wartime sexual violence or members of an underground opposition movement in a repressive society.

While the protection of human subjects should be of primary concern to researchers, such protections sometimes appear to be in conflict with emerging trends toward data access and research transparency. As we note above, transparency regarding the interview process helps to increase confidence in interview-based analyses, but data access raises thorny ethical and practical questions. Recording and/or sharing interview notes or transcripts is at odds with many human subjects protections, whether interviewees are elites or members of vulnerable populations, and, even if the interview recordings or transcripts can be made available, they will be devoid of potentially important contextual variables, given the protections typically promised to interviewees, as well as the potential for positionality and interviewer effects. This, in turn, leads to an increased likelihood that the data will be misinterpreted, which could cause embarrassment, distress, or harm to the original respondent. Thus, it is understandably difficult to acquire IRB approval and respondents' consent to share recordings, notes, and transcripts.

Although the difficulty of sharing – much less replicating – interview data has not grown over time, the perceived magnitude

of the problem has, because political science is placing more importance on making the research process as transparent as possible. The discipline increasingly, and often for good reason, puts a premium on providing verification or replication materials. Although this is a laudable effort, it does conflict directly with the ethics of interview research, which privilege respondents' privacy and safety above all else. In some cases, the need to protect interviewees' confidentiality simply makes it impossible for researchers to satisfy both the discipline's emerging norms and the standards of ethical research.

Researchers can still provide some transparency about their data without releasing full interview transcripts, as we note above. Again, we recommend that scholars report summary statistics about respondents either in tables in the text or appendices. As the number of interviewees increases, it might even be possible to supply some fairly detailed data (about the interview sample as a whole) while still protecting respondents' confidentiality. The data could consist of information that is useful in any study, such as the occupations of interviewees or their geographic location, or it could be worth reporting only because it is relevant to the topic at hand, such as an overview of respondents' migration histories, average number of years that informants spent in parliament, or interviewees' religious identity. Of course, researchers must record these characteristics from the start; trying to collect this kind of data after the fact is tricky, if not simply impossible. Still, if scholars are systematic and intentional about the way they solicit, collect, and publish interview data, there are ways to provide a measure of much-needed transparency and still abide by the ethical rules of human subjects research.

Depending on the context, it also can be difficult to ensure that respondents are providing truly informed consent when they agree to allow their words to be made publicly available. This is less likely to be a problem when interviewees are elites who

share the same language and culture as the researcher, but explaining academic research and publishing practices across linguistic, cultural, and/or educational divides can be a fraught process. In some cases, scholars may technically have a respondent's consent to cite and provide free access to the interview data, but their sense of a respondent's level of understanding will compel them to withhold identifying information for certain quotations or to keep certain segments of an interview private (e.g. Wood, 2007).

Even when individual respondents are amenable to having their responses made publicly available, such openness can threaten to undermine other interviewees' willingness to offer information. Seeing the details of supposedly private meetings and conversations referenced in scholarly pieces or reading about one's profession or political strategies portrayed in an unflattering light can cause future interviewees to think twice about granting access to interested scholars. It also is possible that what seems to interviewees like harmless information to reveal at one point in time may seem more harmful in the future, as individual or national political circumstances change. Similarly, if interview data are used and misinterpreted by scholars who were not present for the interview but rather accessed the transcripts online, a formerly willing interview respondent may decide not to participate in future research. This may not occur frequently, but the risk exists.

The potential damage to researchers and their colleagues can be even greater when researchers share data without respondents' consent. Such privacy breaches need not be the result of ill intent; researchers may incorrectly file an interview, forget to mark whether a respondent agreed to be quoted by name, or have a laptop stolen. This need not be a catastrophic event; depending on the topic and the respondents, a privacy violation of this sort may cause no harm and attract no attention. However, it is distinctly possible that such violations of research ethics will threaten not only one's own scholarly enterprise, but also

access for future researchers. This underscores the importance for researchers of carefully documenting consent for each interviewee, of keeping files secure through encryption or strong password protection, and of regularly considering how changing circumstances in the interviewee's political context may change the meaning of informed consent.

The issue of proper training becomes even more critical in this light – an emphasis on proper technique, best practices, and familiarity with IRB requirements, as well as broader ethical considerations and guidelines, will help ensure that potential interview respondents remain willing to discuss and share their experiences with scholars.

## CONCLUSION

This chapter has discussed certain unique characteristics of interview research, the challenges associated with interview-based approaches, and ways to improve and expand the use of this tried-and-true method. As we have shown, even in a world in which most of the methodological attention is on causal identification and the use of big data of various forms, interviews remain both relevant and useful. Consequently, graduate programs should help their students to acquire proper interview training, departments and journals should encourage scholars to utilize interviews in their research, and researchers should familiarize themselves with the advantages and drawbacks of this unique method.

As the data in the second section of this chapter demonstrate, interviews remain a small but integral part of political science research. In light of increasing privacy concerns and the hesitancy of elites to leave digital footprints, they may come to play an even more central role in projects that seek to make and test causal claims. Certainly, the discipline would benefit richly from ensuring that both current and future researchers are aware of the utility of interviews for their

projects, have the training they need to conduct interviews professionally, and understand the pitfalls and positives of interview research. With better training and a more professional approach to conducting interviews, political scientists may find that interviews are appropriate for a wider range of projects than they had considered. This, in turn, may make researchers more likely to incorporate interviews into their own work, which would increase the use of original data, expand the wider availability of interview data, and allow scholars to make more valid and robust causal claims.

Our discussion of the issues surrounding interview research both acknowledges and demonstrates that it may take much longer to responsibly gather, analyze, and present interview data than it does to write an article using a pre-existing quantitative dataset. However, given the ways in which interviews can contribute to the development of theory, the testing of hypotheses, and the richness of the discipline, there are strong reasons for researchers to invest in interview research and for journal editors to set clear standards that will facilitate the publication of interview-based work in political science outlets. Interviews may not be trendy, statistically complex, or methodologically innovative, but, as this chapter shows, they are an invaluable resource for scholars in every subfield.

## Notes

- 1 We thank Elizabeth Osman for assistance in gathering the Cambridge University Press catalog data reported in this chapter.
- 2 [https://mcwilson215.shinyapps.io/comparative\\_trends/](https://mcwilson215.shinyapps.io/comparative_trends/)
- 3 This study was deemed exempt by the IRB at the University of North Carolina at Chapel Hill (Study #18-2544), and it was approved and declared exempt (Protocol H18360) by the Institutional Review Board at the Georgia Institute of Technology.
- 4 In many cases, respondents chose more than one category to describe their use of interviews, so the percentages in Figures 61.2 and 61.3 sum to more than 100.

- 5 One also might code dissertation abstracts for the use of interviews; this would offer a better gauge of the use of interviews among doctoral students.
- 6 It is worth noting that even among US universities, the requirements of the IRB with respect to specifying interviewees and interview questions vary dramatically across institutions.
- 7 See [https://www.isr.umich.edu/cps/project\\_dart.html](https://www.isr.umich.edu/cps/project_dart.html) for a joint political science journal editors' statement on transparency. Also see <https://www.dartstatement.org/> for additional resources.
- 8 These issues have been discussed extensively as part of the Qualitative Transparency Deliberations, available at <https://www.qualtd.net/>. Also see Büthe and Jacobs (2015).
- 9 The Qualitative Data Repository is part of the Institute for Qualitative and Multi-Method Research (IQMR) at Syracuse University.
- 10 <https://ajps.org/ajps-replication-policy/>

## REFERENCES

- Agarwala, Rina and Emmanuel Teitelbaum. 2010. 'Trends in Funding for Dissertation Field Research'. *PS: Political Science and Politics* (April): 283–293.
- American Political Science Association. 2012. *A Guide to Professional Ethics in Political Science, Second Edition*. Washington, DC: American Political Science Association. Available at [www.apsanet.org/portals/54/Files/Publications/APSAEthicsGuide2012.pdf](http://www.apsanet.org/portals/54/Files/Publications/APSAEthicsGuide2012.pdf) (Accessed on 20 January 2020).
- Anria, Santiago. 2018. *When Movements Become Parties: The Bolivian MAS in Comparative Perspective*. Cambridge: Cambridge University Press.
- Baczko, Adam, Gilles Dorronsoro and Arthur Quesnay. 2018. *Civil War in Syria: Mobilization and Competing Social Orders*. New York: Cambridge University Press.
- Ballard-Rosa, Cameron, Layna Mosley and Rachel Wellhausen. 2020. 'Contingent Advantage: Sovereign Borrowing, Democratic Institutions and Global Capital Cycles'. *British Journal of Political Science*, forthcoming.
- Bauerle Danzman, Sarah. 2017. 'Leveraging WAIPA to Facilitate Private Sector Linkages'. *World Association of Investment Promotion Agencies Research Note*, World Association of Investment Promotion Agencies, April.
- Bennett, Andrew and Jeffrey Checkel, eds. 2014. *Process Tracing: From Metaphor to Analytic Tool*. Cambridge: Cambridge University Press.
- Bleich, Erik and Robert Pekkanen. 2013. 'How to Report Interview Data'. In *Interview Research in Political Science*, ed. Layna Mosley. Ithaca: Cornell University Press, 84–108.
- Brooks, Sarah M. 2009. *Social Protection and the Market in Latin America: The Transformation of Social Security Institutions*. Cambridge: Cambridge University Press.
- Brooks, Sarah M. 2013. 'The Ethical Treatment of Human Subjects and the Institutional Review Board Process'. In *Interview Research in Political Science*, ed. Layna Mosley. Ithaca: Cornell University Press, 45–56.
- Bunte, Jonas. 2019. *Raise the Debt: How Developing Countries Choose their Creditors*. Oxford: Oxford University Press.
- Büthe, Tim and Alan Jacobs. 2015. 'Symposium: Transparency in Qualitative and Multi-Method Research'. *Qualitative & Multi-Method Research* 13(1): 2–64.
- Cammett, Melani. 2013. 'Using Proxy Interviewing to Address Sensitive Topics'. In *Interview Research in Political Science*, ed. Layna Mosley. Ithaca: Cornell University Press, 125–143.
- Carnegie, Allison and Austin Carson. 2019. 'The Disclosure Dilemma: Nuclear Intelligence and International Organizations'. *American Journal of Political Science* 63(2): 269–285.
- Côté, Isabelle. 2013. 'Fieldwork in the Era of Social Media: Opportunities and Challenges'. *PS: Political Science & Politics* 46(3): 615–619.
- Desposato, Scott, ed. 2015. *Ethics and Experiences: Problems and Solutions for Social Scientists and Policy Professionals*. New York: Routledge.
- Fujii, Lee Ann. 2008. 'The Power of Local Ties: Popular Participation in the Rwandan Genocide'. *Security Studies* 17(3): 568–597.
- Fujii, Lee Ann. 2010. 'Shades of Truth and Lies: Interpreting Testimonies of War and Violence'. *Journal of Peace Research* 47(2): 231–241.
- Greenstein, Claire. 2018. 'Pressures, Promises, and Payments: Explaining Governments' Reparations Decisions after Domestic Human Rights Abuses.' PhD Dissertation, University of North Carolina at Chapel Hill.

- Grzymala-Busse, Anna. 2011. 'Time will Tell? Temporality and the Analysis of Causal Mechanisms and Processes'. *Comparative Political Studies* 44(9): 1267–1297.
- Henshaw, Alexis, Ora Szekely and Jessica Trisko-Darden. 2019. *Insurgent Women: Female Combatants in Civil Wars*. Washington, DC: Georgetown University Press.
- Huber, John. 2013. 'Is Theory Getting Lost in the Identification Revolution?' <http://themonkeycage.org/2013/06/is-theory-getting-lost-in-the-identification-revolution/> (Accessed 15 October 2019).
- Johnson, Jeremy. 2018. 'Protecting the Community: Lessons from the Montana Flyer Project'. *PS: Political Science & Politics* 51(3): 615–619.
- Kapiszewski, Diana, Lauren M. MacLean and Benjamin L. Read. 2015. *Field Research in Political Science: Practices and Principles*. Cambridge: Cambridge University Press.
- Kapiszewski, Diana, Lauren M. MacLean and Benjamin L. Read. 2018. 'Reconceptualizing Field Research in Political Science'. In *Oxford Research Encyclopedia of Politics*, ed. William Thompson. New York: Oxford University Press, DOI: 10.1093/acrefore/9780190228637.013.722.
- Karpowitz, Christopher, Tali Mendelberg and J. Baxter Oliphant. 2014. 'Gender Equality in Deliberation: Unpacking the Black Box of Interaction'. *Perspectives on Politics* 2(1): 18–44.
- Leech, Beth L., ed. 2002. 'Symposium: Interview Methods in Political Science'. *PS: Political Science and Politics* 35(4): 663–688.
- Lupia, Arthur and Colin Elman. 2014. 'Openness in Political Science: Data Access and Research Transparency: Introduction'. *PS: Political Science and Politics* 47(1): 19–42.
- Lynch, Julia. 2013. 'Aligning Sampling Strategies with Analytic Goals'. In *Interview Research in Political Science*, ed. Layna Mosley. Ithaca: Cornell University Press, 31–44.
- Martin, Cathie Jo. 2013. 'Crafting Interviews to Capture Cause and Effect'. In *Interview Research in Political Science*, ed. Layna Mosley. Ithaca: Cornell University Press, 109–124.
- Moravcsik, Andrew. 2014a. 'One Norm, Two Standards: Realizing Transparency in Qualitative Political Science'. *The Political Methodologist* 22(1): 3–9.
- Moravcsik, Andrew. 2014b. 'Trust, but Verify: The Transparency Revolution and Qualitative International Relations'. *Security Studies* 23(4): 663–688.
- Mosley, Layna. 2003. *Global Capital and National Governments*. New York: Cambridge University Press.
- Mosley, Layna., ed. 2013. *Interview Research in Political Science*. Ithaca: Cornell University Press.
- Niedzwiecki, Sara. 2018. *Uneven Social Policies the Politics of Subnational Variation in Latin America*. Cambridge: Cambridge University Press.
- Obermeier, Katharina and Thomas B. Pepinsky. 2018. 'Dataset on Methodology in Comparative Politics', Harvard Dataverse Network, <https://doi.org/10.7910/DVN/UCBNEH> (Accessed 15 October 2019).
- Pepinsky, Thomas B. 2018. 'The Return of the Single Country Study', SSRN, <https://ssrn.com/abstract=3197172> (Accessed 15 October 2019).
- Petrova, Bilyana. 2018. 'Determinants of Inequality in Latin America and Eastern Europe'. PhD Dissertation, University of North Carolina at Chapel Hill.
- Reno, William. 2013. 'The Problem of Extraterritorial Legality'. In *Interview Research in Political Science*, ed., Layna Mosley. Ithaca: Cornell University Press, 159–178.
- Rubin, Herbert J. and Irene S. Rubin. 2011. *Qualitative Interviewing: The Art of Hearing Data*. 3rd edition. Thousand Oaks, CA: Sage Press.
- Tansey, Oisín. 2007. 'Process Tracing and Elite Interviewing: A Case for Nonprobability Sampling'. *PS: Political Science & Politics* 40(4): 765–772.
- Samii, Cyrus. 2016. 'Causal Empiricism in Quantitative Research'. *Journal of Politics* 78(3): 941–955.
- Soss, Joe. 2015. 'Talking Our Way to Meaningful Explanations: A Practice-Centered View of Interviewing for Interpretive Research'. In *Interpretation and Method*, eds Dvora Yanow and Peregrine Schwartz-Shea. New York: Routledge, 193–214.
- Su, Phi Hong. 2017. '"There's No Solidarity": Nationalism and Belonging among Vietnamese Refugees and Immigrants in Berlin'. *Journal of Vietnamese Studies* 12(1): 73–100.

- Van Noorden, Richard. 2015. 'Political Science's Problem with Research Ethics'. *Nature*, June 29, 2015. <https://www.nature.com/news/political-science-s-problem-with-research-ethics-1.17866> (Accessed 15 October 2019).
- Vera-Adrianzen, Fiorella P. 2017. 'Reparative Justice in Post-conflict Peru: The Effects of Reparations on Individuals and Communities'. Conference Paper Presentation, Lima: Latin American Studies Association.
- Wedeen, Lisa. 2010. 'Reflections on Ethnographic Work in Political Science'. *Annual Review of Political Science*, 13(1): 255–272.
- Weinstein, Jeremy M. 2007. *Inside Rebellion: The Politics of Insurgent Violence*. New York: Cambridge University Press.
- Wilson, Matthew Charles. 2017. 'Trends in Political Science Research and the Progress of Comparative Politics'. *PS: Political Science and Politics* (October): 979–984.
- Wood, Elisabeth Jean. 2007. 'Field Research'. In *Oxford Handbook of Comparative Politics*, eds Carles Boix and Susan C. Stokes. Oxford: Oxford University Press, 123–146.

# Focus Groups: From Qualitative Data Generation to Analysis

Virginie Van Ingelgom

## INTRODUCTION

The idea that interviewing several people at the same time might be more advantageous than interviewing them separately was discovered a long time ago.<sup>1</sup> In a focus group, a set of people are invited by a researcher to discuss a political topic or a set of social issues and are queried concerning their ideas, beliefs, or perceptions. The researcher plays the role of a moderator and a discussion is takes place between the participants. Robert K. Merton's 'focused interview' was only one among several suggestions about the reasons and the ways to interview several people at the same time (Merton and Kendall, 1946). Merton's model became successful in the field of applied research and consulting, where its technical aspects were refined during the 1970s and the 1980s (Copsey, 2008). In the 1990s, social sciences rediscovered the method and a variety of uses and techniques now exist.<sup>2</sup>

This chapter presents the focus group method as a strategy of data collection. Focus

groups may be common to many disciplines, but its uses vary greatly. If interviewing has been a fundamental technique of data collection in political science and international relations, focus groups are still less commonly used than face-to-face interviews or even oral history (Kapiszewski et al., 2015). This chapter deals with the question of how this method is specific in its contributions to political science and international relations.<sup>3</sup> There is indeed a strong relationship between discipline, theory, and practices, as opposed to a tendency to generalise a one-size-fits-all approach to using focus groups. For example, the number of participants involved or the role of the researcher can vary greatly depending on the objectives of the research and the subfield or research. Whereas marketing research will mobilise larger focus groups with 10 to 12 participants, an in-depth political sociologist fieldwork project on citizens attitudes towards European integration will prefer gathering four to six participants. This chapter does not provide a practical

how-to guide for organising and facilitating focus groups.<sup>4</sup> Instead, it unpacks the methodological choices for using focus groups in political science and international relations, which is the primary purpose of this chapter. Therefore, the image of this method that is drawn here is necessarily selective.

The method presented in this chapter embodies most of the challenges and specificities of qualitative research. However, what chiefly characterises focus groups is the richness of the data collected as the purpose is to obtain group data, and not only data associated with particular individuals (Kapiszewski et al., 2015: 201). However, this richness is costly in comparison to what can be achieved with more standard research methods, such as face-to-face interviews or survey questionnaires. Using focus groups not only requires more resources than other qualitative methods, it also requires a large series of decisions to be made concerning the different technical aspects involved from design to moderation. More importantly, analysing focus group data leads to specific challenges, at least for research purposes. Thus, it seems quite important for researchers to get a clear idea of the characteristics and consequences of this method before using it. This chapter aims to help the reader decide if the method is appropriate to her own research project, which is the secondary purpose of this chapter.

By reviewing all the difficulties of focus group research, the chapter will come across most questions faced by qualitative researchers, both epistemologically and technically. By focusing on its specificities, it will present the trade-offs, pros, and cons of using focus groups but also of using different types of focus group. More precisely, the topics covered by the chapter will be organised as follows: first, the chapter identifies the domains and research questions where using focus groups is the more useful in political science; then, the chapter discusses the design of focus groups including sampling, questioning, and moderating, as well as issues related to their practical organisation; and finally, yet

importantly, the chapter puts a strong emphasis on analysis – from general principles to specific challenges of analysing focus groups before turning to some concluding remarks.

## **INTRODUCING FOCUS GROUPS TO POLITICAL SCIENTISTS**

Beyond the actual definition of the focus group, the scientific uses of focus groups as a strategy of data collection are quite diverse. Indeed, focus groups can be used in very different epistemological frameworks. The success of the focus group as a strategy of data collection is in part due to its appropriateness for research questions and the role accorded to it in the research design. Therefore, before one goes into its practical dispositions, it is important to understand the uses of focus groups and their appropriateness in political science.

### ***What is a Focus Group?***

Focus groups fall within the qualitative research tradition in political science and international relations. ‘Collective interviews’, ‘group interviews’, ‘discussion groups’, or ‘focus groups’ all refer to interviews conducted with several respondents at the same time. In order to define more precisely what is meant by the term focus group in this chapter, the three-point definition provided by David L. Morgan – an author who was essential in developing this method for use in social sciences – is a useful starting point. He defines a focus group ‘as a research technique that collects data through group interaction on a topic determined by the researcher’ (Morgan, 1996: 130).

Three elements stand out in this relatively inclusive definition. First, the focus group is a research method designed to collect data. Thus, focus groups are intended for research interests – in other words, for collecting discursive data aimed for systematic analysis.



Second, in the focus group, the source of data lies in the interaction within the discussion group; the relations that characterise them are not reduced to the relationship between the facilitator and interviewee but require the interactions of a collective discussion between the participants. Third, and finally, the focus group supposes the intervention of the researcher in the creation of the group discussion in order to collect data, as they will pose questions and facilitate the discussion – even if the participants shape the direction and the emphasis of the discussion. Thus, data are provoked and collected by a researcher on themes that they have chosen.

This definition enables one to exclude a series of configurations similar to focus groups, such as collective interviews used outside a research context (e.g. marketing or training purposes).<sup>5</sup> It also enables one to distinguish this method from other procedures including multiple participants but ones that do not allow the emergence of interactive discussions between them – such as group experiments aimed to record actions rather than discourses.<sup>6</sup> This definition also excludes direct observation of naturally occurring political discussions that cannot be described as interviews because the researcher does not intervene in the creation of data. Collective interviews, therefore, cannot be conflated with ordinary conversation, such as what might occur in everyday life.<sup>7</sup> The framing of the discussion is always made clear by the presence of a moderator who imposes the subject on the participants and makes sure – in a more or less direct manner – the discussion is kept alive. The researcher selects the focus. Originally, in Robert K. Merton's 'focus interview', interviewees were involved in a particular concrete situation – for example, watching a movie – that was the focus of the interview (Merton and Kendall, 1946). In contemporary political science, collective interviews are organised around topic(s) rather than common experience (Duchesne, 2017). The focus group also differs on these different points from the citizen conferences initiated by

political actors as a way of including citizens in the political decision-making process and reinforcing deliberation and public debate – often called 'mini-publics' (Grönlund et al., 2014; Jacquet, 2018).

### ***The Different Approaches to Focus Groups in Political Science***

Apart from this common definition, there are many practices of focus groups across disciplines or even researchers. Focus groups became prominent in the social sciences in the 1980s, having been developed and standardised over a few decades within the market research profession (Morgan, 1997). The adoption of this strategy of data collection in social studies was due to two factors (Duchesne et al., 2013). First, researchers on the margins between the social sciences and marketing pushed for the focus group to be used as a method that was likely to collect – in a more advantageous way than the individual interview – a substantial number of opinions in a short period of time (Krueger, 1994). Second, researchers in the social sciences saw this as a way to depart from an essentialist approach to opinions (Wilkinson, 1998). The focus group method enables the analysis of the co-construction of meaning thanks to the recording of interactions between participants. Thus, in political science, the practice of focus groups stems from a historical and classical tension between positivist and constructivist approaches.<sup>8</sup> Nowadays, and as discussed more broadly in the field, many researchers seem to adopt elements from both traditions (Soss, 2006).

This dual heritage is still evident today. On the one hand, the method is easily accepted and promoted in its canonical form by social scientists who use statistical methods, despite their concern about the group dynamic as a potential source of bias compared to the expression of individual opinions. On the other hand, researchers who emphasise the interpretive nature of data analysis, and who

see in it an instrument adapted to innovation and to methodological experimentation, also particularly value focus group research (Morgan, 1993; Barbour and Kitzinger, 1999). Thus, focus groups as a research tool – and as a research choice – gives rise to potentially contradictory methodological debates stemming from the distinct disciplinary backgrounds and epistemological assumptions of researchers (Barbour, 2007: 2). These contradictory uses are, at the same time, at the heart of the richness of this methodology. Thus, focus groups are a very flexible research method and therefore have a wide variety of applications: exploratory research, explanatory research, and evaluation research, for example (Hennink, 2014).

Exploratory research is one of the most common applications of focus groups in political science. The group dynamic makes focus groups an ideal method of data collection to explore a topic about which little is known. David M. Farrell and Michael Gallagher (1999) emphasised this point in their focus group research on voters' attitudes towards electoral systems – a topic that is not the subject of intense debate in the electorate. The authors wanted to explore what UK voters know and do not know about electoral systems. Once the basics of the four different systems were clarified, the participants of the focus groups demonstrated that there was clear evidence of support for change, though with little agreement over the preferred system. Mobilising focus groups discussion, Farrell and Gallagher's research reveals a certain ambivalence over the importance attached to constituency representation.

The exploratory use of focus groups also reveals a related advantage in that it has been applied in situations where there was a substantial difference in perspective between researchers and participants (Barbour, 2007; Morgan, 1996: 133). Here the advantage comes from the group dynamic that enables the participants to talk in their own terms as they are discussing together. This argument has been developed in the context of studies

on the risks of HIV within the male homosexual and bisexual community (Barbour 2007). However, it seems that it can be usefully applied to topics in political science as well. Indeed, a minimum of reflexivity regarding the profile of a given political scientist and their generally very specialised perspective calls into question their ability to consider the political issue in terms that correspond to the perceptions of less competent or politically interested people. As Pamela J. Conover et al. (1991: 805) underline, the group dynamic assures that 'participants talk to one another in their own language, rather than simply reacting to the questions and language of an interviewer in a one-to-one situation'.

This exploratory research application of focus groups should not undermine the potential of focus group as an effective tool for explanatory research. There are numerous examples of focus groups as a primary method in research about politics (Akachar et al., 2017; Billig, 1992; Damay and Mercenier, 2016; Delmotte et al., 2017; Duchesne and Haegel, 2004a, 2010; Duchesne et al., 2013; Frazer, 1988, 1989; Gamson, 1992; Hopf, 2002; Jarvis and Lister, 2012; Marsh et al., 2007; Stocker et al., 2016; White, 2011). Focus groups provide an opportunity for the researchers to gain access not only to certain political beliefs or attitudes, but also to identify the underlying context in which they develop – subsequently enabling an understanding of *how* participants think about political issues and not only of *what* they think. For example, in public opinion research, 'focus groups are valuable in revealing the process of opinion formation, in providing glimpses of usually latent aspects of this process, and in demonstrating the social nature of public opinion' (Delli Carpini and Williams, 1994: 62). By forcing participants to 'think out loud', focus groups enable an observation of the process of opinion-formation in action and in interaction with one another.

As the principal asset of focus groups is to provide a tool for participants to discuss

and debate, focus groups have proven useful for studying (de)politicisation and, more precisely, to understand how citizens ‘talk politics’ or resist the idea of discussing politics (Conover and Searing, 2005; Gamson, 1992).<sup>9</sup> In this respect, focus groups – when designed appropriately – can generate data on how citizens get involved or not in political discussion (Duchesne, 2017). Focus groups are particularly useful for studying topics that are considered either sensitive or difficult for people, as the dynamic of group discussion helps individual participants to get access to more ideas or to express things that would otherwise be too difficult. Political issues are not only complex, but also contentious.

Moreover, the focus group is a research tool that gives a voice to the research participant by giving them an opportunity to define what is relevant, salient, or important for understanding their experience or perception. Here, the non-directive approach of a focus group is essential in order to assess the salience of a political issue.<sup>10</sup> Robert K. Merton and Patricia L. Kendall (1946: 545) already argued in this line in the 1940s, asserting that the non-directive approach:

gives the subject the opportunity to express himself about matters of central significance to him rather than those presumed to be important by the interviewer ... it uncovers what is on the subject's mind rather than opinion of what is on the interviewer's mind. Furthermore, it permits subject's responses to be placed in their proper context rather than forced in the framework which the interviewer considers appropriate.

Finally, focus groups provide an effective tool for evaluative research, in particular for the evaluation of public policies, to examine the effectiveness of a program or policy instrument (e.g. Hsiao-Li Sun, 2012; Shek, 2017). The idea is to uncover the strengths and weaknesses of policy instruments in an evaluative perspective. Traditionally, public health policies have been largely evaluated through the use of focus groups (Basch, 1987). For example, focus groups have been used to examine how the LGBT community processed media messages about HIV/AIDS

and how they influenced peoples' perceptions of AIDS (Kitzinger, 1994).<sup>11</sup>

### ***Combining Focus Groups with Other Methods***

Originally, in Robert K. Merton and Patricia L. Kendall's ‘Focused Interview’ in the 1940s (1946), focused interviews were used to clarify the results of experiments – watching film series in order to study the role of mass communications in the dynamics of attitude formation and change (Delli Carpini and Williams, 1994). In this tradition, focus groups continue to be used not only with experiments, but also with participant observation, face-to-face interviews, and surveys (Morgan, 1988: 30–6). In particular, studies using both focus groups and opinion surveys, drawing on qualitative and quantitative data, are very common. Thus, one specificity of focus group research is that it has been widely used alongside other methods in mixed methods studies (Hennink, 2014: 17).<sup>12</sup> David L. Morgan cited more than 60% as the number of empirical publications using focus groups along with another method (Morgan, 1996: 130). As a result, the combination of methods has been the object of close attention in manuals on focus groups, methodological reflections anchored more broadly in a concern for triangulation (Barbour, 2007: 46–7; Caillaud and Flick, 2017). David L. Morgan (1993, 1996: 134–6) pursued this reflection in depth, proposing a conceptual framework that identified four different ways of combining qualitative and quantitative methods – and collective interviews and opinion surveys in particular. For example, focus groups can be used before conducting a survey to identify salient issues in the current political debates on which to develop survey questions, refine a question's wording, or specify relevant response categories in order to increase data validity (Conover et al., 1991). They can also be used as a microscope to understand surveys' answers (Van Ingelgom, 2014).

### ***When not to use Focus Groups in Political Science?***

If the approaches are multiple – exploratory, explanatory, or evaluative – focus groups are not always the most appropriate method to use and can lead to poor quality research. Most of misuses of the focus group method result from the collective nature of data collection. To this regard, when compared to traditional research methods, focus groups must be placed on a continuum between ethnography's emphasis on naturalness and unobtrusiveness on the one hand, and the greater control provided by experiments and in-depth interviews on the other (Delli Carpini and Williams, 1994).

On the one hand, compared with ethnographic approaches such as participant observation, the setting of a focus group is less natural, in the sense that the researcher creates the situation to be observed in order to collect her data. If naturalness is of primary importance, then the focus group is not the most appropriate tool as the setting will remain artificial. On the other hand, compared to experiments, the logic of the focus group – even in a very positivist approach – is not in isolating specific cause and effect relationships. More precisely, to assess the adequacy of the method, one needs to consider the collective and interactive dimension of the data collected. The group dynamic is at the heart of most methodological problems.

To this regard, focus groups are less suitable for eliciting personal experiences or individual narratives from each participant. Even if participants may share their individual story with other participants, who can be strangers, there is insufficient time to provide a personal narrative about a specific topic. Moreover, due to the dynamic of discussion, individual narratives will be fragmented, incomplete, or even confused as other participants interrupt or question each other. Of course, this is less of a concern for focus groups with fewer participants (Barbour,

2007). If the researcher is attempting to collect data on individual narratives, then they should carry out individual interviews. In other words, a focus group is not simply a means for obtaining accounts of individuals, rather it is 'a means to set up a negotiation of meanings through intra- and inter-personal debates' (Crang and Cook, 1995: 56).

Finally, the choice of focus groups as a research tool is based on the conviction that individual attitudes are not given, but instead result from a process of construction that occurs using speech in a collective and sometimes even contradictory context (Duchesne and Haegel, 2004b; Duchesne et al., 2013). Thus, focus groups assume – contrary to surveys – that attitudes, opinions, and perceptions are developed in part in interaction with other people and opinions cannot be observed in a vacuum as individuals do not form opinions in isolation. At the heart of the method is the analysis of shared meanings and disagreements. Thus, obviously, focus groups are not an appropriate method to measure attitudes (Barbour, 2007: 19).

### **DESIGNING AND CONDUCTING FOCUS GROUP RESEARCH**

Even if focus groups are tailored to particular research questions, any researcher aiming to conduct focus group faces the same set of important choices. Who to recruit as participants? How to recruit participants? How many groups to run and how many people to gather in each group? How to stratify participants in each group? There are no definitive answers to these questions and certainly no one-size-fits-all approach to using focus group in political science and international relations. Nonetheless, there are general guidelines and important considerations when conducting focus group research that can improve methodological rigour and, thereafter, the quality of data produced (Hennink, 2014).

## ***Research Design and Sampling***

Like any other strategy of data collection, focus groups call for due attention in developing an appropriate and rigorous research design. As, 'the essential purpose of focus group research is to identify a range of perspectives on a research topic, and to gain an understanding of the issues from the perspective of the participants themselves' (Hennink, 2014: 2), effective sampling is key to the success of focus groups and to determining their comparative potential (Barbour, 2007: 3). By providing access to a wide range of political views, perceptions, or experiences, the group setting increases the likelihood of unexpected results emerging from focus group analyses (Delli Carpini and Williams, 1994).

As the focus group is a qualitative methodology, participants are not selected by means of systematic random sampling. The purpose of qualitative sampling is, indeed, to reflect the diversity within the population under scrutiny or to select a group in order to observe within-group dynamics rather than aspiring to recruit a representative sample (Kuzel, 1992). As underlined by Richard A. Krueger and Mary Anne Casey (2009: 66), the purpose of focus groups research is 'not to infer but to understand, not to generalise but determine a range, and not to make statements about the population but to provide insights about how people in the groups perceive a situation'. Thus, the objective is never to aggregate individual attitudes from a systematically random sample in order to generalise about a wider population. If this were the aim, then focus groups would inevitably be negatively evaluated, as even dozens of focus groups gathering hundreds of participants would never be representative of the entire population (Stanley, 2016).

The logic behind sampling is instead theoretical, as the researcher theorises on the dimensions that are likely to be relevant in order to give rise to differing political beliefs, perceptions, or experiences. Thus, 'sampling relates to the anticipated use of

the selected criteria in making comparisons once the data have been generated. In other words, purposive sampling allows for the data to be interrogated purposefully, that is, in order to carry out systematic comparison' (Barbour, 2007: 58). Sampling is central and theoretically driven; comparison and segmentation are at the core of a focus groups research design.

## ***Comparison and Segmentation***

Since the groups are the main unit of analysis, it is a classical rule to convene these to facilitate comparison. The question of the comparison is central when it comes to group composition and to the number of focus groups that are ultimately conducted. Indeed, the question of how many groups to hold reflects the comparative research design. The number of groups is determined by the comparisons that the researchers wishes to make and thus by the sub categories the researcher wants to target. The more segmented groups are – for example, by education, country, or age – the more groups will be necessary to complete the research design.

However, it is fundamental to consider that each participant may potentially meet several of the desired criteria in terms of diversity – for example, education, age, political affiliation. Then comparisons can be made based on fewer groups than sampling might suggest, as intra-group comparison is also an important tool of analysis. Moreover, the number of groups to be conducted is not only guided by the comparative research design, but also by the principle of 'saturation' (Glaser and Strauss, 1967). Saturation is 'the point at which information collected begins to repeat itself. After reaching the point of saturation, any additional data collection becomes redundant, because the purpose of qualitative research is to uncover diversity and context, rather than a large number of participants with the same type of experience' (Hennink, 2014: 43).

*An Example of a Comparative Research Design: Citizens Talking About Europe (Duchesne et al., 2013: 160–195)*

The *Citizens Talking About Europe* (CITAE) project was set out to gather and record discussions between citizens in Paris, Brussels, and Oxford on the subject of Europe. In analysing what makes sense for the participants in these discussions, the aim was to understand how citizens (de)politicise European issues. The CITAE project convened 24 focus groups, each consisting of four to eight participants. Groups were constituted according to a two-fold criterion of social and national differences (double-layer design). The sampling strategy builds on survey research that shows the profound effects of national as well as social differences on attitudes to European integration. Therefore, the groups were constituted in such a way as to differentiate social and national characteristics. Thus, the social homogeneity pursues both a methodological and a theoretical objective. From the methodological perspective, it was essential to ensure a minimum of shared comprehension, linguistic rapport, and relatively easily communicable social experiences between the participants in a group. At the same time, it was important to avoid striking social differences of the kind that generate domination effects. Three social categories according to occupation – workers (and/or unemployed, or casually employed), employees, and managers – were distinguished. The research team also added a fourth category, party activists, with the idea of gathering competing partisan views of the subject. Finally, the CITAE research team convened two of each category of groups in order to control the effects of group dynamics and to reach saturation.

**Double-layer design**

	<i>Belgium</i>	<i>France</i>	<i>UK</i>	
Working class	2	2	2	Saturation?
White collars	2	2	2	
Managers	2	2	2	
Activists	2	2	2	
	Saturation?			

Researchers have reported on the results of as many as 66 focus groups (Liebes and Katz, 1990), 37 focus groups (Gamson, 1992), and as few as six focus groups (Damay and Mercenier, 2016; Delmotte et al., 2017). A well thought out comparative research design should, however, lead the researcher to keep the number of focus groups to a bare minimum as the researcher should also consider time constraints – recruitment, transcription, and analysis of focus groups are all very time consuming.

When it comes to the theoretical sampling criteria – also called segmentation – of the groups, in political science research in particular, social stratification is made necessary

by the characteristics of focus group method itself. Speaking in public, especially on political subjects, is socially determined. There is the matter of greater or lesser ease of expression, and of confidence in the ability to say what one is thinking and to convince others about what one believes. There is also the matter of the use of words, the structures of language, and cultural references. Thus, one of the golden rules when selecting the participants for a given focus group is to secure social homogeneity in order to minimise domination. Education and occupation of participants are a classical proxy in order to achieve social homogeneity.

Of course, many focus groups discussions have a dominant participant who monopolises the discussion, is always the first to respond to a question, or provides lengthy comments (Ritchie and Lewis, 2003). Nevertheless, social homogeneity determined by the education and/or the occupation of participants – and attentive moderation – will limit the risk of domination and ensure that the other participants are not completely overshadowed or inhibited.

### ***Recruitment Strategy and Selection of Participants***

Whether working on quantitative or qualitative data, the quality of recruitment is a decisive factor in the reliability of the survey. In the case of focus groups, the issue is twofold. Because, by recruiting, one builds not only a sample according to criteria that will weigh on the comparative analysis, but also a group whose interactions will be decisive for the quality of the material (Duchesne and Haegel, 2004b: 45). Although collective interviews are not more complicated than individual interviews in terms of recruitment, they are in terms of their implementation. For example, if an error of selection occurs in an individual interview, the interview in question can always be set aside in a later phase of analysis. This is less the case for collective interviews. Such an error may hinder, or even compromise, the group discussion and the interpretation of the overall results for that group. Of course, the interactions that take place in the context of these collective interviews always involve a certain ‘social alchemy’ and are, as a result, at least partly unpredictable. However, the recruitment should enable the researchers to anticipate what will take place as much as possible, notably by controlling the social homogeneity of each group.

Obviously, one considers that a good discussion dynamic would be one that enabled each participant to express themselves and thus to understand others and to be understood. This means that participants must

speak the same language, each are able to identify and ‘gauge’ their participants as quickly as possible after meeting them, and to situate themselves in relation to the other participants. From this perspective, the recruitment aimed to avoid bringing together people that are too far removed from each other socially in order to facilitate discussion. Convening smaller groups is thus also important. With groups of four to seven participants, participants are in position to talk to each other and everyone has an opportunity to secure the floor (Duchesne, 2017).

The process of recruiting a sample that fits the desired research design can be time-consuming. Thus, the first question to consider is whether pre-existing groups or constructed groups best correspond with the research question. A common method is to use a key informant, who is sometimes paid, to be an organiser and who convenes a group of acquaintances – for example, friends, neighbours, or co-workers. Familiarity facilitates the dynamic of group discussion. However, by delegating the composition of the groups to key informants, researchers risk weakening the social stratification of the groups and/or, ending up with only very competent participants.<sup>13</sup> The issue of political competence is often at the core of many classical political science questions and thus should be carefully thought of. As William Gamson remarks (1992), interviewees who use personal contacts tend to recruit from amongst their acquaintances people whom they believe to be the most likely to talk in public and who, accordingly, are mainly found to be amongst the most educated. Thus, interpersonal recruitment does not allow a choice of all participants and the certainty to recruit lay citizens, which can be fundamental to the study of (de)politicisation, for example (Duchesne, 2017). Familiarity between the participants may also lead to many implicit discussions that are difficult to interpret for the researcher.

Artificial groups – defined as groups gathering strangers, differentiated according

to theoretical sampling criteria – could be preferred because they typically avoid all of the aforementioned problems. In order to have maximum control over the selection process, researcher can choose to construct groups that are lacking any kind of pre-existing sociability – in other words, groups in which the participants were not previously acquainted with each other or the researcher. A common and tempting general strategy is delegating recruitment to specialists – for example, opinion polling companies (Baglioni and Hurrelmann, 2016).<sup>14</sup> Nevertheless, such companies tend to work with files of volunteers who often end up being almost focus group specialists, something researchers in political science might also want to avoid. Researchers may be keen to make contact with a public that most often eludes opinion surveys and eludes focus groups when recruited by professional research agencies. Researchers could want to reach a population that is most often excluded (or self-excluded) in opinion surveys, those whose feeling of competence – particularly political competence – is low, thus predisposing them to avoid this kind of participation. As Sophie Duchesne (2017: 371) underlines, ‘it is well-documented that the more remote people feel from the political field, the less likely they are to accept to participate in interviews, in general, and, in particular, in interviews on political topics’ (Gaxie, 1978). Having to speak publicly in front of strangers on potentially controversial political subjects while being recorded for two hours or more is clearly a particularly unusual and potentially harrowing experience that might discourage certain potential respondents.<sup>15</sup> Having in mind the comparative potential also increases the likelihood of including groups that might be otherwise overlooked (Barbour, 2007). For instance, Marsh et al. (2007: 59) mobilised focus groups to research political participation – and more precisely how young people understand and experience politics. By engaging with 12 focus groups, including 65 participants, these researchers contrasted

responses from a ‘politically active’ group with those from a ‘politically inactive’ group. Consequently, using a financial incentive to attract these types of participants is sometimes considered indispensable.

For the same reasons, the subject of the discussion could not be communicated to the potential participants in order not to discourage citizens who are not interested in (or even repelled by) political issues from participating. Selecting a venue regularly used by participants is also essential in order to limit the cost of participating as well as to help participants to feel more quickly at ease and comfortable. For example, Heidi Mercenier realized her focus groups in community and youth centers (Delmotte et al. 2017). It is worth noting that a venue could also have particular associations for participants. For example, a focus group held in a university building may lead participants to feel the need to show expertise on the matter in hand or to feel as if they were back in school. If the question of knowledge or expertise is central to the research question, perhaps a university building should be avoided. Here, reflexivity on the settings and the relationship between the researchers and the participants is fundamental.

Depending on the research objectives, recruiting participants who, on the one hand, do not know each other and, on the other hand, present specific characteristics, necessitates a sometimes costly recruitment system. Researchers could want to obtain detailed socio-political information from respondents in order to select those who ultimately participate in the focus groups and suit best their research objectives and their comparative research design. A questionnaire could therefore be given to all those who replied to advertisements or any other recruitment strategy<sup>16</sup> in order to select the final participants.

In practice, a variety of ways exist to recruit participants: community gatekeepers, informal networks, advertisements, leaflets, posters, formal services, etc. William Gamson (1992: 16) set up recruiting tables at community events such as festivals, fairs, and flea markets,



whereas Tamar Liebes and Elihu Katz (1990) approached members of different communities and asked them to recruit additional participants. Jonathan White (2011: 230) recruited taxi drivers across 10 cities in the UK, Czech Republic, and Germany directly at the taxi rank at a time that coincided with low customer demand. Multiplying the recruitment strategy is, most of time, a good way to go, particularly in order to recruit different profiles.

### ***Discussion Guide and Moderation***

Setting up collective interviews also involves a series of choices linked to procedures of the organisation of the discussion itself. Due to the extent that the researcher is active in the construction of the research, a high degree of reflexivity is necessary, at least in retrospect.

It is important to distinguish first between structure that controls questioning and structure that controls the group dynamics. In managing group dynamics, a less structured approach allows participants to talk as much or as little as they please, while a more structured approach means that the moderator will encourage those who might otherwise say little and limit those who would otherwise dominate the discussion. When it comes to the focus group questions, a less structured discussion means that the group can pursue its own interests, while a more structured approach means that the moderator imposes the researchers' interest, as embodied in the questions that guide the discussion. It is crucial to allow participants to speak freely and not to force them to redirect the flow of the discussion onto the topic because this allows the groups to ignore the topic and to address matters that interest them.

A key factor that makes groups more or less structured is, thus, the number of questions. In focus group research, the aim of the discussion guide is twofold: to foster productive discussion and to elicit useful data that meet the research objectives. Hennink (2014: 48) underlines that 'a discussion guide is a

pre-prepared list of topics or actual questions used by a moderator to guide the group discussion'. Therefore, 'the moderator uses the guide as a resource to maintain the balance between the researchers focus and the group's discussion' (Morgan, 1997: 48). The discussion guide is thus a checklist to ensure that the different dimensions of the topic under scrutiny are covered throughout the discussion. While the non-directive approach is one of the main characteristics of the focus group method, in academic research, the discussion guide is often structured around pre-designed questions and probes. This structuration helps to enhance the comparability in data analysis as the study population is often divided into subgroups.

Concretely, the first few moments in a focus group discussion are critical. In this brief time, the moderator must give enough information so people feel comfortable with the topic, create an open atmosphere, provide the ground rules, and set the tone of the discussion. The best introduction is often the honest admission that researchers are there to learn from the participants, and so the goals for the session should be expressed in those terms. The discussion begins with broad questions to launch the dynamic among group participants so that they begin to feel comfortable and competent. A central objective is to get each participant to give some meaningful response or opening statement. One should easily be able to respond. The initial question should be something that all the participants would be interested in. Beginning with a general question that emphasises the participants' interests lets the researcher hear the participants' point of view rather than starting with the researcher's interests. Thus, there is an opportunity to discover new ways of thinking about the issues. As importantly, they have to be worded with the appropriate level of vocabulary and syntax. This is of course challenging when the sample encompasses people with quite varied social – and possibly national – backgrounds. Typically, follow-up questions are used when necessary to clarify participants' opinions.

On the guiding and moderation of the discussion, there is also a fair amount of variation in existing research. The CITAE team opted to conduct a discussion that was largely non-directive but structured by managing the dynamic of the group discussion (Duchesne et al., 2013: 160–95). The conduct of the sessions was based on a schedule of scripted questions put to the groups in a rigorously standardised manner in every group. Sessions lasted about three hours (a long period by comparison with other focus group studies) and were structured around five questions, taking about 30–45 minutes each. This left open large time slots for discussion between participants, following Robert K. Merton and Patricia Kendall's (1946) advice to 'not introduce a given topic unless a sustained effort is made to explore it in some detail'. Of course, longer-lasting groups face the challenge of arousing the interest of the participants for three hours.

Vignettes and imagined scenarios have been successful as a means to prompt group discussion (Gamson, 1992; White, 2011). In William Gamson research, the moderator was showing the participants Jonathan White presented a card at the beginning of the group. Tamar Liebes and Elihu Katz (1990) had their groups begin by watching an episode of *Dallas*, and then used the show to structure the discussion that began with very general questions and became more specific as the discussion progressed. In the end, the objectives of the research should define the length and the depth of the focus group. However, the researcher should always keep in mind that not leaving enough time for discussing collectively a topic, a question, or an idea is a mistake and that the group dynamic is essential and the discussion should be prompt.

## ANALYSING FOCUS GROUPS

Focus group data are typically analysed using conventional methods for qualitative

analysis.<sup>17</sup> However, analysing focus group data provides its own set of challenges. Of course, the method of analysis selected depends on the purpose of the study and the approach to analysing focus group data varies from study to study. Again, it is not the intention here to provide a 'how to' guide for analysis focus group data, but rather to underline some specificities of analysing focus group data in political science research.

### **Data Collection**

The analysis of collective interviews differs from the outset by the potential diversity of the available material. The data stream begins with field notes and recordings taken during each focus group, continues with the debriefing following the groups, and goes into transcripts. Thus, focus group data include documented observations made by the research team throughout the sessions. Indeed, a note-taker should be part of the team and attend the group discussion to develop a written summary of not only the key issues raised, but also of the group dynamic (Hennink, 2014: 83–5). Field notes cover the arrival period, the break, and the period after the discussion closed and before the participants' departure. These notes include all comments made and all conduct and action that the team members present were able to observe and recall. The debriefing notes are essential and constitute as such the first step of the interpretative analysis.

Transcription remains the essential element of the corpus. Indeed, many analytical approaches require data to be transcribed to produce a written record of the discussion for analysis (Hennink, 2014: 87). In the case of face-to-face interviews, for the most part, the methods of analysis proposed in the manuals rely on text analysis. In the case of focus groups, this reduction to the text is more difficult to operate. The degree of

precision of the transcription needed obviously depends on the type of analysis one intends to conduct. For example, content analysis or thematic analysis requires a verbatim transcript, whereas conversational analysis has additional transcription norms, because understanding how participants express themselves is as important (Hennink, 2014). In this latter case, the transcription needs to include details on hesitations, pauses, and word emphasis but also on body language, when relevant. In any case, it is thus important that the transcripts reflect as

finely as possible what occurs in the collective interviews.

Research that requires a careful record of who was speaking to whom will require either a video recording or an assistant that takes notes throughout the group interview. The key difference between audio and video recording is the intrusiveness of the latter. As all the efforts that went into recruiting is wasted when a 'technical problem' causes a loss of data, it is advised to pay equal attention to the quality of the recording. Ensuring the quality of the recorded data is crucial.

*Example of Transcribing Conventions: Talking About the Royal Family (Billig, 1998: 24)*

The main signs used in the transcriptions are as follows:

= to be found at the end of one speaker's utterance and at the beginning of another's: no audible gap between the two speakers. This would be expected if the second speaker were interrupting the first, who then gives way.

[ two or more speakers making utterances at the same time.

(.) short pause. If the pause is of sufficient duration to be measurable, then the time is indicated in seconds: i.e. (2) indicates a pause of two seconds.

... omitted or inaudible material.

{ } clarificatory addition, often referring to tone or gesture i.e. {laughing}.

underlining of words in the text indicates emphasis through raised voice; capitalisation indicates particular emphasis or volume.

Moreover, as focus group research is conducive to a diversification of the modes of moderation, it also contributes towards producing complementary material, which the group itself elaborates at the request of the moderator. This complementary material could also be analysed. As part of the CITAE research, Sophie Duchesne also analyses the cards produced by the participants to answer the question 'who profits from Europe?' in order to understand and characterise the social gap existing in the reactions to European integration (Duchesne et al., 2013).

### ***General Principles of Focus Group Analysis***

As stressed by Sue Wilkinson (1998: 169), 'focus groups are distinctive ... primarily for the method of data collection (i.e. informal group discussion), rather than for the method of analysis'. Thus, focus groups analysis follows the general principles of qualitative data analysis. Some of them are however more important to keep in mind when analysing focus groups.

First, as with other research methods in social sciences, focus groups can benefit from

the classic advice that decisions about how to collect the data will depend on decisions about how to analyse the data. The degree of structure in the collective interviews and the segmentation of the group composition points to the obvious influence that the research design has on the subsequent analysis of the data. For example, although using a more structured approach in collecting data does not force the use of a group-by-question grid in the analysis, using a less structured approach certainly limits the value of this option.

Second, the analysis should follow two important general principles: the analysis is systematic and continuous. Systematic analysis means that the analysis strategy is documented and that all the material is analysed. The interactive nature of the data reinforces the need for the systematisation of data analysis. Researchers must continually be careful to avoid the trap of selective perception. Data analysis is also sequential and continuous; data collection and analysis are concurrent. As already mentioned, the debriefing notes taken after each focus group are the very first step of interpretative analysis. The principle of saturation is anchored also in this concurrence of data collection and analysis as the number of focus groups to conduct is determined through iterative process until information saturation is reached (Hennink, 2014: 43).

Third, the analysis can focus on the group taken as a whole or on the interactions between the participants. Concretely, this leads to choosing between focusing the analysis on the content of the opinions collectively expressed – or, on the contrary, on the way in which they are elaborated (Duchesne and Haegel, 2004b: 83). In the first case, one will initially characterise each of the groups, synthesise the content of what was expressed, and then conclude by comparing the groups. While in the second case, one will have to individualise participants to understand better what is happening at each moment of the discussion. Although the group is the main unit of analysis, most of the time the analysis needs also to take into account individuals

within the group (Barbour, 2007: 129). However, the analysis should never be at the level of individuals only.

Finally, as Monique M. Hennink (2014: 87) underlines, ‘approaches to data analysis can broadly be divided into those that break-up data into segments or themes for analysis (e.g. content analysis, thematic analysis) and those that do not break-up data to analyse the whole narrative (e.g. discourse analysis, conversation analysis)’. As useful as this categorisation is, in fact, this distinction should rather point to the two steps of analysing focus groups rather than be seen as two alternatives. These two steps are interpretation (mainly based on an analysis of alliances and sensitive moments) and data reduction (mainly either by categorisation or condensation and systematic comparison). There is, indeed, a need to retain something of the richness of the data collected, while ensuring that the amount of data generated are analysed systematically and comparatively (Frankland and Bloor, 1999). Those two steps are described in the next sections.

### ***Interpretation: Alliances, Sensitive Moments, and Experiential Knowledge***

The main challenge of analysing focus group data is reflecting and utilising the interaction between participants and taking into account group dynamics (Kitzinger, 1994). This could be very challenging as the discussions in focus groups are marked by the fragmented and disorganised nature of the reasoning of participants. This is particularly true at times when the discussion becomes heated, and when the participants tend to talk at the same time. The transcripts consist of many sequences that are difficult to understand for anyone who merely reads them – that is, for anyone who looks for meaning only in the content and the sequence of utterances. Participants react, gradually, to what is said and to what they hear. In order to not to

take quotes out of context, interpretative analysis is thus an essential first step. The interpretative phase helps researchers to begin the process of incorporating the data.

Before trying systematically to analyse and compare the data, it is necessary to go through a stage of interpretive construction of the meaning of the utterances exchanged. Relating utterances to the development of relations between the participants in order to understand meaning has been developed by Michael Billig (1992). From conversations between close relatives about the British Royal family, he shows how arguments are adapted to the reputation of the interlocutors: someone considered to have strong opinions will be led to retain his role, to construct his discourse, and to adapt his responses in order to always have the last word. By contrast, the CITAE groups were made up of strangers without any prior reputations to maintain (Duchesne et al, 2013).<sup>18</sup> However, their interventions could certainly be interpreted as reactions of participants to each other and impressions of group members that were developed over the course of the session. These reactions tended to be strengthened after a while in terms of more or less explicit alliances. The interpretive analysis therefore provides a framework from which one could analyse how participants constructed their interventions and adapted their positions or opinions according to interactions with those whom they wanted to express an agreement or disagreement. In the same vein, the relationship established between both participants and the moderator is part of the data construction and should be taken into account in the analysis. For all of this, making syntheses and summaries of the groups proves to be a useful step in the interpretative analysis.

When it comes to using interaction between participants to generate data, interpreting the data from focus groups requires distinguishing between what participants find interesting and what they find important (Morgan, 1997). When participants discuss a topic at length, this is a good indication that

they find it interesting, but it does not necessarily indicate that they find it important. In addition, specific attention is necessary for researchers to understand the strategies that participants adopt to deal with perspectives that are difficult to express – both because of their emotional charge and their complexity.

In particular, Jenny Kitzinger and Clare Farquhar argue for the analytical potential of ‘sensitive moments’, defined by their emotional charge and indicated by ‘explicit comments from the research participants, hesitation and awkwardness, reactions of surprise or shock, individual defensiveness or tentative collective exploration’ (Kitzinger and Farquhar, 1999: 156). The simple idea behind this concept is to be particularly attentive in the analysis to moments ‘where something is happening’ in the group discussion. In order to deal with those ‘sensitive moments’, participants may employ humour in expressing their points of view rather than speaking about them directly (Liamputtong, 2012: 175). Of course, they can also simply remain silent. Silences are thus also a valuable resource for analysis in terms of group interaction, as well as discourse content. Indeed, what is not said can be as important as what is said (Barbour, 2007: 141). Blake Poland and Ann Pedersen (1998) propose a useful analytical distinction between ‘silences of estrangement’, where issues have no salience for participants, and ‘silences of familiarity’, where issues are not mentioned as they are taken for granted.

Finally, Jenny Kitzinger and Clare Farquhar (1999: 171) stress that ‘above all, when groups delved into and began to address sensitive issues in depth, this provided invaluable information about the place of personal experience within public discourse’. Thus, eluding experiential knowledge is also a useful tool for interpretation (Stanley, 2016). William Gamson found that experiences in the form of anecdotes have a specific place in discussions. Participants frequently made their point by ‘telling a story’ that sought to make a broader point about

how the world works (Gamson, 1992: 122). Those stories are sometimes from television or newspapers, but the majority of anecdotes are about themselves or at least someone they know personally – such as family, friends, and co-workers.

### **Data Reduction: Systematic Comparison and Coding**

Michael Billig's research, *Talking Of the Royal Family*, is based on 63 focus groups recorded in over 2,800 pages of typed transcript. The longest transcript of a single interview was 150 pages (Billig, 1998: 18). The CITAE project includes 2,400 pages of typed transcript for 24 focus groups conducted in three different countries (Duchesne et al., 2014). Thus, data reduction is often a necessary step in the analysis of focus groups but should always come after a careful phase of interpretation of the data.

Rosaline Barbour (2007) underlines that three basic factors influence the salience given to a topic: (1) how many groups mentioned the topic? (2) How many people within each of these groups mentioned the topic? (3) How much energy and enthusiasm did the topic generate among participants? On this last point, the interpretative analysis is of course of precious help. The two former factors call for systematic comparison. Systematic comparison, focusing on both inter- and intra-group differences, helps qualitative data analysis to overcome the limits of purely descriptive accounts (Barbour, 2007: 131). It helps also to avoid the danger of selective recall of, and attention to, some data and neglect of other contradictory data (Frankland and Bloor, 1999). Pre- or post-questionnaires and field notes are useful here to record details on participants and to further explore intra-group differences and similarities.

Since qualitative analysis is essentially comparative, indexing data facilitates comparative analysis by gathering all data on a particular topic under one heading (Frankland

and Bloor, 1999). It also helps when managing the data – field notes and debriefing notes should also be coded. Commonly, researchers are interested in the themes that emerge from the data. This kind of analysis is referred to as thematic analysis – or frame analysis – and is processed through systematic coding. The process of coding involves researchers reading and re-reading the text and assigning codes, which relate to the content of the data and are of interest to the researcher's analytic framework. In a first stage, codes are likely to be quite broad, and then to become narrower and more focused as the work continues. Coding fulfils the essential function of linking the different parts of the corpus to each other. It enables the systematic comparison between different excerpts of the text that are attached to a given code applied throughout the entire corpus. The use of Computer-Assisted Qualitative Data Analysis Software (CAQDAS), such as Atlas.ti and NVivo, enables the application of code to text to be reviewed, tested, and revised (Vila-Henninger, 2019). Because of their search and retrieve facilities, CAQDAS enables researchers to find patterns and divergences in large datasets.<sup>19</sup> Such packages facilitate the qualitative analysis and ensure that the analysis incorporates the entire data set.

Once coded, the transcripts can be systematically analysed, both qualitatively and quantitatively. Qualitative analysis involves going back to an interpretative analysis, but only on a selected portion of the data collected. For example, relying on CITAE dataset, Virginie Van Ingelgom (2013, 2014) retrieved and analysed all the statements made by participants stating in the pre-questionnaire that their country membership to the European Union was neither good nor bad. By comparing those statements, she was able to construct a typology that identifies three mechanisms of depoliticisation: ambivalence, indifference by distance, and indifference by fatalism. Quantitative analysis includes the examination of the frequency of certain kinds of statements by groups or

by individuals. In the same study, Virginie Van Ingelgom was able to demonstrate that those ambivalent participants were not less competent or able to discuss European affairs than others were. However, counting should still be understood in line with qualitative epistemology. Counting is important in identifying patterns in the data and not in using numbers in ways to attach significance to the actual values (Barbour, 2007: 131). In general, researchers present findings through a combination of direct quotes taken from the group discussions, interpretative summaries, and comparative tables.

## CONCLUSION

While interviewing has traditionally been an essential technique for gathering data in political science and international relations, focus groups are not yet a core mode of data collection in this field. However, more and more researchers are mobilising this technique of data collection in an innovative way and very often in combination with other techniques, such as opinion polls. Yet, while organising a collective discussion may look easy or straightforward to an outsider, it involves much more than sparking a conversation between participants, at least when focus groups are used for research purposes, as this chapter has demonstrated.

The purpose of focus groups is indeed to collect and to analyse group data for research objectives on a topic focus constructed by the researcher. They are a very demanding technique, but they provide a rich and detailed set of data about participants' perceptions, beliefs, and frames in their own words. Focus groups represent a flexible research tool, as they may be quite general or very specific; highly structured or non-directive; stemming from positivist to interpretative traditions. Of course, the most important differences among interviews with individuals, oral histories, and focus groups are the number of participants

involved, the role the researcher plays, and the structure of the interaction (Kapiszewski et al., 2015: 193–4). These three elements define the type of analysis that can be undertaken but stress at the same time the trade-offs between these three techniques of data collection. Each type of interviews has characteristics, advantages, and drawbacks<sup>20</sup> – so does the focus group. The aim of this chapter was to give the reader the key to decide if focus groups are an appropriate technique for data collection for their research. Indeed, as for any research technique of data collection, the decision to use it must be based on the appropriateness of the method for obtaining answers to specific research questions (Stewart and Shamdasani, 2015: 177).

Thus, as useful as any methodological chapter as this one could be, the main and concluding advice to be given to any researcher that intends to engage with focus groups research should be to get familiar with the very convincing work that mobilises this technique of data collection. This chapter has built on those examples of innovative and rigorous research. They are the best illustrations of when and how to use focus groups in political science and international relations and therefore should be on any reading list.

## Notes

- 1 Myself, I also discovered this idea a long time ago thanks to Sophie Duchesne. Any errors in this chapter are mine and only mine. She was so generous to teach me everything she knew about this very demanding methodology. I would like to thank her for her continuous commitment to ambitious research, as I would like to thank all colleagues and students that I have had the chance to work with in order to generate data through focus groups in the last fifteen years. A special thank to the students who I have been teaching this method to over the years, in particular those of the ECPR Winter and Summer Methods Schools.
- 2 This chapter concentrates on focus groups as used by political scientists and IR specialists, even if political parties and administrations increasingly employ the techniques to uncover attitudes

- towards political campaigns or public policies (Savigny, 2007). Note also that focus group research has evolved recently by the introduction of virtual focus groups (Bloor et al., 2001: 74–88; Lobe, 2017; Stewart and Shamdasani, 2015: 157–76). Addressing specific issues related to virtual focus groups is, however, beyond the scope of this chapter.
- 3 Based on the results of a survey of a US political science faculty, Kapiszewski et al. (2015: 190–1) detail in their chapter on interviews, oral histories, and focus groups that 81% of field research projects reported made significant use of interviews. This percentage varies across subfields: 92% for comparative politics, 84% for IR scholars, 84% for Americanists, and 50% for political theorists. Overall, according to this survey, focus groups represent 13% of field research projects. Their survey also underlines that while only 6% of the reported fieldwork projects carried out in the 1960s and 1970s in the United States employed focus groups, 15% of the projects from 2000 onwards did so (Kapiszewski et al., 2015: 201). To our knowledge, the use of focus groups in international relations is very rare. However, the chapter will include examples of research conducted on attitudes towards the EU in order to offer researchers in international relations relevant examples.
  - 4 In order to find a practical how-to guide, the reader can refer to the existing plethora of focus group handbooks (Barbour, 2007; Bloor et al., 2001; Hennink, 2014; Liamputtong, 2011; Morgan, 1997; Stewart and Shamdasani, 2015).
  - 5 Note that, in our view, the label ‘focus groups’ is very inclusive. Thus, collective interviews for research purposes also qualify as focus groups. The main differences in this case will be the kind of moderation that is used, and in particular, how active the facilitator is in the discussions.
  - 6 Note that this definition will thus exclude original Merton’s ‘focused interview’ from the current practices of focus groups. Robert K. Merton (1987: 550) underlined this point in his *The focused interview and focus groups*, when declaring: ‘the truth of the matter is that there can’t be many people in the field of social science and certainly none in the related field of marketing research who know less about focus groups than I’.
  - 7 The obvious methodology here is ethnography. Nina Eliasoph, Camille Hamidi, and Katherine Kramer Walsh spent years observing people discussing political issues (Cramer Walsh, 2004; Eliasoph, 1998; Hamidi, 2010).
  - 8 For a stimulating illustration of this tension, see Liam Stanley’s (2016) piece in *Politics*.
  - 9 For a well-detailed argumentation on how to use focus groups for studying (de)politicisation, see the recent chapter by Sophie Duchesne (2017).
  - 10 According to Robert K. Merton and Patricia Kendall (1946), an effective focused interview has several interrelated characteristics: non-directedness, depth, range, and specificity. More recently, Sophie Duchesne (2017: 376–8) renewed the arguments in favour of adapting the non-directive approach to focus groups when developing questions and moderating.
  - 11 To this regard, and in extension, the method is also associated more closely with political marketing. In this perspective, political parties have increasingly employed focus groups to gauge citizens’ attitudes towards particular policies since at least the mid 1990s (Savigny, 2007). However, as those focus groups are used outside a research context, they are not included in this chapter as it deals with focus groups as a strategy of data collection for research purposes.
  - 12 On mixed method research design, see Harbers and Ingram, Chapter 58, this *Handbook*. The aim of using qualitative and quantitative methods in combination is to gain a broader understanding of the research issue that no single method alone can provide, each approach illuminating different dimensions of an empirical object under scrutiny.
  - 13 As such having competent participants is not a problem, but it could be depending on the objectives of the research. For example, if ones want to understand how the European integration process is perceived, gathering only participants highly competent on European matters and very interested in the topic will lead to misleading conclusions.
  - 14 Another tempting strategy is recruit students (Bruter, 2005). If this could be justified for realising a pilot or for exploratory research, strong reflexivity is needed as students are a very specific target group when it comes to political issues and politicisation.
  - 15 In order to record and being able to transcribe precisely the discussion, it is common to use video recording and/or audio recording, but this has an obvious cost as it can be perceived as intrusive by some participants.
  - 16 While pre-interview questionnaires might produce some interesting biographical data, the risk is also to atomise the group as extended silences are a bad way to begin discussion – and could take up valuable time better devoted to collective, rather than individual, activities (White, 2011: 239).
  - 17 On qualitative data analysis see Harbers and Ingram, Chapter 58, this *Handbook*.
  - 18 Sophie Duchesne and her team developed their own way of analysing interaction, drawing on



Michael Billig (1992) and Jenny Kitzinger and Clare Farquhar (1999). Each of the 24 discussions analysed was the subject of an interpretive narrative account by Sophie Duchesne, Elizabeth Frazer, or Florence Haegel. Using the video recording, the transcription, the questionnaires filled in by the participants, and the observational notes written by the team after each group session, she constructed an account of the discussion, and everything that happened around it, by responding to the following questions: what happened between the participants? What conflicts were avoided and what conflicts were engaged with in the discussion? What agreements or what consensus was found and how? How did alliances between group members develop? What divisions did they reveal? What were the subjects of discussion, explicitly and implicitly? What resources did the participants mobilise? The final document took the form of a narrative about the participants and their conduct within the group. The template is detailed in Duchesne et al. (2013: 190–2). For a detailed explanation on the way to mobilise alliances in the study of (de)politicisation, see Duchesne (2017).

- 19 Note that CAQDAS are, of course, also used for interpretative coding. Silences, sensitive moments, disagreements, experiences, and so on, can be coded as well.
- 20 For a systematic comparison of those advantages and drawbacks, see Kapiszewski et al (2015).

## REFERENCES

- Akachar, S., Celis, K., & Severs, E. (2017). Hoop en verraad: wat moslimjongeren verwachten van vertegenwoordigers met een etnische minderheidsachtergrond. *Res Publica* 59(4), 463–483.
- Baglioni, S., & Hurrelmann, A. (2016). The Eurozone crisis and citizen engagement in EU affairs. *West European Politics* 39(1), 104–124.
- Barbour, R. (2007). *Doing focus groups*. London: Sage.
- Barbour, R., & Kitzinger, J. (eds.) (1999). *Developing focus group research: Politics, theory and practice*. London: Sage.
- Basch, C.E. (1987). Focus group interview: An underutilized technique for improving theory and practice in health education. *Health Education Quarterly* 14, 411–448.
- Billig, M. (1992). *Talking of the royal family*. London: Routledge.
- Bloor, M., Frankland, J., Thomas, M., & Robson, K. (2001). *Focus groups in social research*. London: Sage.
- Bruter, M. (2005). *Citizens of Europe? The emergence of a mass European identity*. New York: Palgrave Macmillan.
- Caillaud, S., & Flick, U. (2017). Focus groups in triangulation contexts. In R. Barbour & D.L. Morgan. *A new era in focus group research* (pp. 155–178). Houndmills: Palgrave MacMillan.
- Copsey, N. (2008) Focus groups and the political scientist. *European Research Working Paper Series*, 22.
- Conover, P.J., Crewe, I., & Searing D.D. (1991). The nature of citizenship in the United States and Great Britain: Empirical comments on theoretical themes. *Journal of Politics* 53, 800–832.
- Conover, P.J., & Searing, D.D. (2005). Studying 'everyday political talk' in the deliberative system. *Acta Politica* 40(3), 269–283.
- Cramer Walsh, K. (2004). *Talking about politics: Informal groups and social identity in American life*. Chicago: The University of Chicago Press.
- Crang, M. & Cook, I. (1995). *Doing ethnographies*. Norwich: Geobooks.
- Damay, L., & Mercenier, H. (2016). Free movement and EU citizenship: A virtuous circle? *Journal of European Public Policy* 23(8), 1139–1157.
- Delli Carpini, M.X., & Williams, B. (1994). The method is the message: Focus groups as a method for social, psychological and political inquiry. In M.X. Delli Carpini, L. Huddy, & R. Shapiro. *Research in micropolitics: New directions in political psychology* (pp. 57–85), Vol. 4. Greenwich: JAI Press.
- Delmotte, F., Mercenier, H., & Van Ingelgom, V. (2017). Belonging and indifference to Europe: A study of young people in Brussels. *Historical Social Research* 42(4), 227–249.
- Duchesne, S. (2013). Social gap: The double meaning of 'Overlooking'. In S. Duchesne, E. Frazer, F. Haegel, & V. Van Ingelgom. *Citizens' reactions to European integration compared: Overlooking Europe* (pp. 65–95). Houndmills: Palgrave MacMillan.

- Duchesne, S. (2017). Using focus groups to study the process of (de)politicization. In R. Barbour & D.L. Morgan. *A new era in focus group research* (pp. 365–387). Houndmills: Palgrave MacMillan.
- Duchesne, S., Frazer, E., Haegel, F., & Van Ingelgom, V. (2013). *Citizens' reactions to European integration compared: Overlooking Europe*. Houndmills: Palgrave MacMillan.
- Duchesne, S. & Haegel, F. (2004a). La politisation des discussions, au croisement des logiques de spécialisation et de conflictualisation. *Revue Française de Science Politique* 54(6), 877–909.
- Duchesne, S., & Haegel, F. (2004b). *L'enquête et ses méthodes: l'entretien collectif*. Paris: A. Colin.
- Duchesne, S., & Haegel, F. (2010). What political discussion means and how do the French and the (French speaking) Belgians deal with it. In M. Wolf, L. Morales & I. Ken'ichi. *Political Discussion in modern democracies. A comparative perspective* (pp. 44–61), London: Routledge.
- Eliasoph, N. (1998). *Avoiding politics: How Americans produce apathy in everyday life*. Cambridge: Cambridge University Press.
- Farrell, D.M., & Gallagher, M. (1999). British voters and their criteria for evaluating electoral systems. *The British Journal of Politics and International Relations* 1(3), 293–316.
- Frankland, J., & Bloor, M. (1999). Some issues arising in the systematic analysis of focus group materials. In R. Barbour & J. Kitzinger. *Developing focus group research: Politics, theory and practice* (pp. 144–155). London: Sage.
- Frazer, E. (1988). Teenage girls talking about class. *Sociology* 22(3), 343–358.
- Frazer, E. (1989). Feminist talk and talking about feminism: teenage girls' discourses of gender. *Oxford Review of Education* 15(3): 281–290.
- Gamson, W.A. (1992). *Talking politics*. Cambridge: Cambridge University Press.
- Glaser, B., & Strauss, A. (1967). *The discovery of grounded theory: Strategies for qualitative research*. New-York: Aldine de Gruyter.
- Grönlund, K., Bächtiger, A., & Setälä, M. (2014). *Deliberative mini-publics. Involving citizens in the democratic process*. Colchester: ECPR Press.
- Gaxie, D. (1978). *Le Cens Caché: Inégalités culturelles et ségrégation politique*. Paris: Le Seuil.
- Hamidi, C. (2010). *La Société civile dans les cités: engagement associative et politisation dans des associations de quartier*. Paris: Economica.
- Hennink, M.M. (2014). *Focus group discussions*. Oxford: Oxford University Press.
- Hopf, T. (2002). Making the future inevitable: Legitimizing, naturalizing and stabilizing. The transition in Estonia, Ukraine and Uzbekistan. *European Journal of International Relations* 8(3), 403–436.
- Hsiao-Li Sun, S. (2012). *Population policy and reproduction in Singapore. Making future citizens*. Abingdon: Routledge.
- Jacquet, V. (2018). The role and the future of deliberative mini-publics: A citizen's perspective. *Political Studies*, 67(3), 1–19.
- Jarvis, L., & Lister, M. (2012). Disconnected citizenship? The impacts of anti-terrorism policy of citizenship in the UK. *Political Studies* 61(3), 656–675.
- Kapiszewski, D., MacLean, L.M., & Read, B.L. (2015). Interviews, oral histories, and focus groups. In D. Kapiszewski, L.M. MacLean, & B.L. Read. *Field Research in Political Science: Practices and Principles* (pp. 190–233). Cambridge: Cambridge University Press.
- Kitzinger, J. (1994). The methodology of focus groups: The importance of interaction between research participants. *Sociology of Health and Illness* 16, 103–121.
- Kitzinger, J., & Farquhar, C. (1999). The analytical potential of 'sensitive moments' in focus group discussions. In R. Barbour & J. Kitzinger. *Developing focus group research: Politics, theory and practice* (pp. 156–173). London: Sage.
- Krueger, R.A. (1994). *Focus groups: A practical guide for applied research*, 2nd edition. London: Sage.
- Krueger, R.A., & Casey, M.A. (2009). *Focus groups: A practical guide for applied research*, 4th edition. Thousand Oaks: Sage.
- Kuzel, A.J. (1992). Sampling in qualitative inquiry. In B.F. Crabtree & W.L. Miller (eds). *Research methods for primary care* (pp. 31–44), Vol. 3. Thousand Oaks: Sage.

- Liamputtong, P. (2011). *Focus group methodology. Principles and practice*. London: Sage.
- Liebes, T., & Katz, E. (1990). *The export of meaning*. Oxford: Oxford University Press.
- Lobe, B. (2017). Best practices for synchronous online focus groups. In R. Barbour & D.L. Morgan. *A new era in focus group research* (pp. 227–250). Houndmills: Palgrave MacMillan.
- Marsh, D., O'Toole, T., & Jones, S. (2007). *Young people and politics in the UK: Apathy or alienation*. Basingstoke: Palgrave MacMillan.
- Merton, R.K. (1987). The focused group interview and focus groups: Continuities and discontinuities. *Public Opinion Quarterly* 51, 550–566.
- Merton, R.K., & Kendall, P.L. (1946). The focused interview. *American Journal of Sociology* 51, 541–557.
- Morgan, D.L. (1988). *The Focus Group Guidebook (Focus Group Kit, Book 1)*. Thousand Oaks: Sage.
- Morgan, D.L. (1993). *Successful focus groups: Advancing the state of the art*. Newbury Park: Sage.
- Morgan, D.L. (1996). Focus group. *Annual Review of Sociology* 22, 129–152.
- Morgan, D.L. (1997). *Focus groups as qualitative research*. Thousand Oaks: Sage.
- Poland, B. & Pedersen, A. (1998). Reading between lines: Interpreting silences in qualitative research. *Qualitative Inquiry* 4(2), 293–312.
- Ritchie, J. & Lewis, J. (2003). *Qualitative research practice: A guide for social science students and researchers*. London: Sage Publications.
- Savigny, H. (2007). Focus groups and political marketing: science and democracy as axiomatic? *The British Journal of Politics and International Relations* 9(1), 122–137.
- Soss, J. (2006). Talking our way to meaningful explanations: a practice-centered view of interviewing for interpretative research. In D. Yanow & P. Swartz-Shea. *Interpretation and Method: Empirical research methods and the interpretative turn* (pp. 127–149). Armonk: Sharpe.
- Shek, D.T.L. (2017). The use of focus groups in programme evaluation: Experience based on the project P.A.T.H.S. in a Chinese context. In R. Barbour & D.L. Morgan. *A new era in focus group research* (pp. 129–154). Houndmills: Palgrave MacMillan.
- Stanley, L. (2016). Using focus groups in political science and international relations. *Politics* 36(3), 236–249.
- Stewart, D., & Shamdasani, P.N. (2015). *Focus groups: Theory and practice*. Thousand Oaks: Sage.
- Stocker, G., Hay, C., & Barr, M. (2016). Fast thinking: Implications for democratic politics. *European Journal of Political Research* 55(1), 3–21.
- Van Ingelgom, V. (2013). When ambivalence meets indifference. In S. Duchesne, E. Frazer, F. Haegel, & V. Van Ingelgom. *Citizens' reactions to European integration compared: Overlooking Europe* (pp. 96–123). Houndmills: Palgrave MacMillan.
- Van Ingelgom, V. (2014). *Integrating indifference. A comparative, qualitative and quantitative approach to the legitimacy of European integration*. Colchester: ECPR Press.
- Vila-Henninger, L.A. (2019). Turning talk into 'Rationales': Using the extended case method for the coding and analysis of semi-structured interview data in Atlas-ti. *Bulletin of Sociological Methodology*, 43(1), 28–52.
- Wilkinson, S. (1998). Focus groups in health research: Exploring the meaning of health and illness. *Journal of Health Psychology* 3(3), 329–348.
- White, J. (2011). *Political allegiance after European integration*. Houndmills: Palgrave MacMillan.

# Interpretive Approaches in Political Science and International Relations

Xymena Kurowska and  
Berit Bliesemann de Guevara<sup>1</sup>

When I first started writing about diversity work as institutional plumbing over 10 years ago, I did not expect I would be in touch with plumbers asking if I can use their images of leaky pipes. Research: it is where we end up. (Ahmed, 2018)

The inclusion of a chapter on interpretive approaches in the *Handbook on Research Methods in Political Science and International Relations* (PS&IR) illustrates the growing interest in the practice of interpretive research. It demonstrates the realisation that attention to meaning and context – the objective of interpretive research – enhances our analyses of politics. The resort to local knowledge in interpretive analysis provides means to address puzzles of contemporary politics that positivist tools cannot unpack. Through the introduction of interpretivism, the traditional bipartite division of social-scientific methods into quantitative and qualitative ones gives way to a tripartite modality that consists of quantitative-positivist, qualitative-positivist, and qualitative-interpretive methodologies (Yanow, 2003). The position of interpretive

approaches in PS&IR remains, however, ambiguous, despite the long interpretive tradition in the social sciences and humanities (and in policy analysis more specifically), as well as the wealth of recent publications explaining interpretive approaches in PS&IR (Bevir, 2000; Schwartz-Shea and Yanow, 2012; Lynch, 2014; Yanow and Schwartz-Shea, 2014; Bevir and Rhodes, 2016; Rhodes, 2017; Heinelt and Münch, 2018; Bevir and Phillips 2019; Steele et al., 2019) and the established status of interpretivism in adjacent disciplines, such as sociology, social anthropology, and human geography. Arguably, this ambiguous status reflects the difficulty of integrating the philosophy and methodology of research, which the opening quote illustrates: in interpretive research, the insight comes in the process of growing engagement and a cross-field search for apt metaphors, where the researcher is a supplicant who learns with others, rather than an expert who tests on others. Interpretive knowledge production relies on an iterative

back-and-forth between theory and lived experience, that of the researcher and of her interlocutors.

In this chapter, we first address persistent misunderstandings that have affected the reception of interpretivism in PS&IR and have often led to an unwarranted application of positivist assessment criteria to interpretive research projects. We briefly discuss the overlaps with and differences to poststructuralism and critical agendas more broadly and focus in some detail on how interpretivists understand theory and method. In contrast to formalised methods, interpretive approaches, crucially, do not use a method template. Rather, method is always specific to, and may change with, the context of the research problem. It cannot be mandated *a priori*. Second, we describe the aims and features of interpretive research and illustrate them through references to existing interpretive scholarship. These examples demonstrate the importance of systematicity in interpretive analysis, which follows the principle of a situated method selection and prioritises flexibility and the shifting of research strategies in accordance with contextual demands. In the third part, we discuss examples of interpretive research in action. We examine the question of reflexivity as the interpretivist counterpoint to positivist objectivity, the extensive contribution of Lee Ann Fujii to the development of interpretive methodology, and the specific usefulness of art-based methods for interpretivist research. As practicing fieldworkers, our emphasis is on interpretive data generation through fieldwork; we do, however, point to some research practices in interpretive policy analysis (IPA) as a starting point for our readers to explore other interpretive approaches that would deserve an elaboration on their own.

## **CLEARING THE GROUND: REVISITING THE BINARIES**

The difference between interpretivist and positivist traditions has primarily to do with

ontology and epistemology rather than methods *per se*. Positivists are philosophical dualists: they situate themselves as separate from the world they observe. They can then study this world through standard procedures that are independent from the context under study. Such a ‘view from nowhere’ is a necessary condition for producing warranted positivist knowledge (Nagel, 1989). Interpretivists are philosophical monists who reject the possibility of transcending context.<sup>2</sup> Centrally, interpretivists reject the correspondence theory of truth, that is, the notion that we can establish an unequivocal correspondence between the truth or falsity of a statement and the real state of affairs. In their view, such an operation is impossible because there is no ‘direct’ access to reality: the ‘view from nowhere’ is an illusion, all the more dangerous as it masquerades as truth. We rather get to ‘know’ reality, or make sense of it, by continuously interpreting it, most pertinently through language as the most common means of representation. Yet language cannot represent reality or truth in a total sense because it is socially mediated and temporally and geographically conditioned (Lynch, 2014: 16); it never quite encompasses reality. While it may be used to frame and manipulate, language effects cannot be fully controlled, so it is misleading to see language as a precise tool in the hands of either the researcher or any other social actor. Quite the opposite: we may find ourselves and the social actors we study ‘positioned’ by language, when larger discourses define who we are. The classic illustration of this condition in International Relations literature is Carol Cohn’s research into the world of nuclear defence intellectuals. Cohn shows how these experts distance themselves from the harm that the nuclear weapons they advocate can cause by constructing a pseudo-neutral ‘technostrategic’ lingo. While being replete with gendered/sexualised metaphors and related emotions, such lingo constructs a reality based on a claim to objectivity that pretends to be purely rational and is immune

to critique (Cohn, 1987). Yet Cohn is also taken by surprise when she catches herself increasingly using that lingo as a result of her participatory observation. This reflexive moment triggers a major shift in her central research question from ‘How can they think this way?’ to ‘How can anyone think this way?’ (Cohn, 2006), including the researcher.

Demonstrating the normative background of allegedly neutral or law-like concepts takes centre-stage in interpretive research. Oren (1995) reveals, for example, how the claim of a democratic peace, that is, the claim that democracies do not wage war against other democracies, which is widely regarded as one of the few laws in International Relations, is a historically specific and value-laden agenda. Klein (1994) shows how deterrence is a social global practice that changes in meaning, since the process of threat construction that underpins deterrence depends on the conception of the underlying values to be defended. Tannenwald (1999) traces the non-use of nuclear weapons as an evolving normative prohibition that over time has stigmatised nuclear weapons as unacceptable weapons of mass destruction. What transpires in such research is that every fact has a normative grounding.<sup>3</sup> Interpretivists insist that ‘facts’ arise within the context of a prior set of beliefs or theoretical commitments (Rorty, 1989). Any claims to neutrally describe an external reality are suspicious. The fact–value binary is not only untenable, it also tends to impose meaning rather than prevent bias: distance is both a stance and a cognitive-emotional orientation (Kondo, 1986: 75). Interpretivism shows in this context that ‘there are no “real” entities, only culturally mediated social facts, and [...] social science is always perspectival and entwined with the pursuit of moral or material goods’ (Schaffer, 2016: 2). This position must not be confused, especially in the age of post-truth, with relativism or ‘anything goes’. It is instead about making sense of power and ethics as constituents of meaning, an objective that remains at the core of the interpretive

research agenda (Lynch, 2014). Considering power relations as contingent social and historical arrangements derives from the constructionist sensibility behind interpretive research: interpretivists see reality as socially produced. Such a sensibility goes hand in hand with a baseline criticality – famously introduced by Cox (1981: 129) – between critical and problem-solving theory: interpretivists are critical researchers as they do not take the world as they find it but inquire how the social order came about. They deal in reconstructions, however, rather than post-structuralist deconstruction. In other words, they reconstruct the ways in which problems are socially understood, but they tend to fall short of a substantive poststructuralist critique that deconstructs the logic behind social representations of such problems (Bacchi, 2015) and the concomitant exclusions that such representations bring about.

In view of its conceptual commitments, interpretivism in PS&IR is then better thought of as an approach with a matching methodology, rather than a set of methods. Methodology here denotes a logic of method selection and application according to the conceptual setup of the project. Because of the contrasting ontological and epistemological holdings of positivist and interpretivist approaches, methods cannot be mixed arbitrarily (Yanow, 2003). Methods applied within one research project need to be compatible with one another, as well as respond to the overall research problem, since epistemology is ‘a skin, not a sweater’ (Marsh and Savigny, 2004: 165). There is therefore doubt among interpretivists about both mixed-method designs and analytical eclecticism, which advocate forsaking paradigmatic differences in favour of pragmatic combinations of methods (Sil and Katzenstein, 2010). For one, such pragmatism may end up in a ‘neither/nor’ research design, that is, one which is consistent neither with positivist nor interpretivist evaluation criteria, possibly giving rein to self-confirmatory research. More fundamentally, the logic of method application and

interpretation of results as a rule unfolds from within a paradigmatic position. Failing to disclose such a position does not translate into methodological pluralism. Calls for pragmatic complementarity may in fact subordinate interpretivist concerns to positivist models under the cover of methodological pluralism.

The emphasis on consistency does not mean, however, that methods should be boxed for exclusive use within specific approaches. Rather, what appears as a single method can be understood and used differently within positivist and interpretivist projects. The notion of data collection versus data generation is a useful way to think about this distinction, as it also reflects understandings of what constitutes empirics. In very general terms, positivist data collection assumes the existence of observer-independent information that is picked up in accordance with a prior research design. It requires a preceding construction of variables and hypothetical causal chains, which are tested on new material/information. In interpretivist data generation, data are not to be found but are made in concrete settings as a result of encounters with humans and non-humans. Data generation reflects the premise of the emergent research design, that is, one which crystallises in the process of research through learning in the field. If positivists have a prior idea of what constitutes their data, for interpretivists it is rather the evolving research question that renders things data. Sources of data are context-specific and they may include numbers, in the sense of the meaning that numbers evoke.<sup>4</sup> Here, certain established dichotomies begin to crumble.

Stone (2016), for example, proposes an interpretive theory of quantitative analysis that collapses the typecast distinction between qualitative and quantitative analysis as different styles of perceiving the world. Quantitative analysis gets its meaning from narrative: counting is in fact metaphor-making, numbers do not speak for themselves unless authors narrate their studies for audiences, and counting is itself an act of

power imposing a certain order on the world (Stone, 2016: 169). Bevir (2014) develops a distinctive narrative form of explanation that unpacks political action interpretatively and shows that the division of labour between approaches that 'explain' and those that aim at 'understanding' is a misconstruction. Ginger (2014) builds an interpretive framework for content analysis, which is usually seen as a positivist method producing numeric data for statistical treatment. Ethnography can similarly be applied both 'for information' through positivist data collection and 'for meaning' through interpretive data generation (Wedeen, 2010). Positivist ethnography explicitly aspires to adjudicate truth claims – what Jessica Allina-Pisano (2009) refers to as 'peeling the onion'. In her pledge for realist ethnography, she suggests that an understanding of the ways people think about their world 'can be a necessary condition for the collection and use of reliable empirical data about them' (2009: 55). International Relations saw some controversy over the early reduction of ethnography to an empiricist data-collection machine (Vrasti, 2008), so it is crucial that the field makes room for ethnographies based on different ontologies.

Interpretive sensibility can help further methodological pluralism in this context, as it rejects methods determinism. What defines such sensibility, following Geertz's (1993) famous statement, is not a matter of methods but of 'thick description'. Thick description is about sorting out structures of meaning by distinguishing frames of interpretation at play and how their co-existence produces friction. Any type of data in interpretive research, whether coming from a mass survey, the analysis of documents or an ethnographic interview, should be treated as evidence of meanings embedded in action (Wagenaar, 2011; Bevir and Rhodes, 2016: 18). For this purpose, method as tool is a misleading metaphor. It presupposes that the researcher has a clear sense of the product that the use of the tool is supposed to manufacture, yet method is not a 'system for offering more or less

bankable guarantees' that guides us to our destination (Law, 2004: 9). Excessive focus on procedure diminishes the ability to generate data through exposure and surprise and may lead to instantiating the initial assumptions from before the research. Good scholarship is characterised instead by making intelligent connections between subject matter and method (Hellmann, 2009), so method is never something outside of the material. It is performative rather than representational: the use of method actively shapes knowledge production, rather than being an application of a neutral procedure. In this sense, methods are practices of world-making: they are acted upon within research but also represent political choices and create new entities (Aradau and Huysmans, 2014: 598).

Finally, the understanding of the role of theory and the distinction between theory and empirics in interpretive research also needs clarification. Generally speaking, positivists treat research and theories they test or build as a replication of the social-political world. Interpretivists see research findings as resulting from intersubjective, meaning-focused processes that themselves interact with, and potentially shape, the studied world (Schwartz-Shea and Yanow, 2012: 40). Interpretive research action unfolds in tacking back and forth between the theoretical and the empirical, the abstract, and the concrete. This means that interpretivists do not in a direct sense test theory. They arrive at theoretical arguments via a recursive and iterative back-and-forth between theoretical frameworks and the experience of data generation. Theory and empirics feed on and fold into each other, and interpretivists seek to articulate this process. Burawoy (1998), for example, talks about interpretive theory reconstruction in contrast to positivist theory testing. Such reconstruction starts with an existing theory to be elaborated and deepened by probing 'negative cases' and inconsistencies. Theory is a scaffolding that keeps the researcher steady, guides the dialogue with participants, and provides the means

for thematising the researcher's participation (Burawoy, 1998: 5). Fieldwork constitutes in this context 'a sequence of experiments that continue until one's theory is in sync with the world one studies' (Burawoy, 1998: 17). Wilkinson (2014), in contrast, takes a more decentering perspective on theory in fieldwork. She objects to 'decanting empirical material into theoretical containers', that is, cleaning up data so that they fit into pre-existing theoretical categories. She recommends instead that an interpretive researcher temporarily and consciously neglect theory while in the field, in order not to impose meaning (2014: 394). In a more radical move informed by a transformation of his theoretical framework while in fieldwork, Zirakzadeh (2009) suggests learning how to discard theories. This is not theory rejection; the suggestion rather gets to the heart of the inescapable requirement for any interpretive social scientist to, as Taylor (1971) puts it, confront one's language of explanation with that of the self-understanding of one's subject. Such confrontation does not mean accepting the language of the participants at face value. But it requires learning with and from them, a process that theory may obstruct if treated as prior to experience. It is by tacking between concrete details and conceptual abstractions that we can refine and undermine, negate and create novel explanations about politics.

## **PREMISES AND COMPONENTS OF INTERPRETIVIST EXPLANATION**

The objective of interpretive research is to make sense of meaning and meaning-making practices of contextually situated social actors, the researcher being among them (Schwartz-Shea and Yanow, 2012). The core premises of interpretive research include the focus on the polysemy of such meaning and a dialectics between embeddedness and variability of social practice, the analytical use of concepts that derive from context, abductive



logic of inquiry, and reflexivity. In the following, we will discuss these premises in turn.

Interpretivists concentrate on meanings as opposed to laws, rules, and correlations between social categories or deductive models because they see meanings as constitutive of action and therefore necessary to make sense of, if we are to explain political action (Bevir and Rhodes, 2006). Meanings are holistic but not necessarily rational or conscious. Explaining an action requires locating it in the wider web of meaning, which differs from both the quantitative-positivist explanation, based on correlation, and the qualitative-positivist explanation, based on causation. For an interpretivist, it is not sufficient, for example, to explain a voter's party choice by either correlating it with an attitude to social justice as detected through a survey or by hypothesising a link between a voter's identity (determined through context-independent variables) and their party choice, and then testing the hypothesis through structured interviews. What interests her is how a participant makes sense of other relevant meanings that link their self-understanding to the vote. How do people name their circumstances, for example, including their place in a community, in a way that makes certain choices sensible and contextually appropriate for them? How is social justice contextually understood, beyond the slogan itself and as experienced in daily life, and how does this connect to other ideas that form a political stance? As we will develop below, the fact that most participants use a given term does not mean consistency of meaning across such users. Even though they use similar terms to describe social justice, for example, they may still understand its meaning differently. Such difference can only be captured and linked to political action through methodologies attuned to meaning. An interpretivist will then be curious how respondents to a survey interpret the issues they are asked to express opinions about, as well as how they make sense of the questions themselves, beyond the categories assigned by the researcher (Walsh, 2009). The

responses will depend on such interpretations, a connection that a survey cannot identify and explore. Similarly, structured interviews constructed to elicit data for the researcher's prior hypotheses may miss the contextual logic and vocabulary of political action that fall outside the hypothesised parameters.

If interpretivist political scientists see political life as action laden with meanings, they differ in how they conceive of such meanings. The most significant divide is between the hermeneutical and poststructuralist traditions that diverge in their views on the nature of political subjectivity. Those working in the poststructuralist tradition see subjects as constituted in discourse. As such, subjects are produced by language and so is their action and their sense-making (Bacchi, 2015). In other words, poststructuralist scholars see meanings as derived from quasi-structures governed by hegemonic discourse and thus reject a strong notion of autonomy. Discourse analysis in International Relations tends to adopt such positions (Milliken, 1999; Hansen, 2007; Epstein, 2008; for policy analysis, cf. Howarth and Griggs, 2012). Bevir and Rhodes (2006) take issue with the ensuing take on the individual as a product of a fixed ideology, with little capacity to change it. From within the hermeneutical tradition, they suggest the concept of 'situated agency' that reflects the contextual construction of the subject and their interaction in the world, with a mediating rather than determining role of theory and ideology. The 'knowability' of such a process can only be assumed from within the contextualised perspective of the subject, whether situational member or researcher (Schwartz-Shea and Yanow, 2012). The implication is that the 'observer' can only 'understand' by participation, by learning how (not) to put things to use within the field. Oftentimes, as we know from ethnomethodology, blunders and unintended interventions lead to such insight. Shehata (2006) notes, for example, that his intrusive presence and the extent to which he challenged social-class taboos as

an Egyptian-American scholar working in an Egyptian sweatshop contributed greatly to how he came to understand the operation of social class in Egypt. This is a good illustration of how it is 'by mutual reaction', that is, by engagement rather than detachment, that we discover the properties of the social order (Burawoy, 1998: 14). That is also the interpretive answer to the question of what value close personal observations bring to the study of politics. Such proximity requires, however, an acknowledgment that the practitioner's point of view is as important as the researcher's, and this challenges the claim to control over the research process (Fenno, 1986). We tend to neglect that social actors know more about what they are up to and its consequences than social scientists give them credit for (Giddens, 1990: 309). The interpretive interactive data generation with participants is a form of co-construction of knowledge that collapses the distinction between the observer and the observed or subject and object of inquiry.<sup>5</sup>

The logic of inquiry that best suits such a practice of knowledge production is abduction (Schwartz-Shea and Yanow, 2012; Lynch, 2014). Abduction differs from induction in that theoretical influences are acknowledged in the process of data generation, but it also differs from deduction. Abduction does not test theories brought to the local environment; research is instead about a continuous iterative recursiveness between the theoretical frameworks, the socio-political and institutional parameters that the researcher is situated in, and the lived experience of the situational members and the researcher. Abduction works by making sense of the present perplexity by resorting to past experience, making analogies from case to case, and probing concepts in new situations. It thus relies on a 'processual merger of creativity, experimentation, testing and adaptation' (Hellmann, 2009: 641). As a form of hunting for clues to make a diagnosis under conditions of uncertainty, abduction is by default inconclusive. It is a mode of inference

that concentrates on experience and practical consciousness. It looks for congruent categories without assuming their completeness or primacy (Onuf, 2013: 98–107). Abduction is thus guided by a radical premise, as seen from traditional research designs in PS&IR, that, '[n]othing new can ever be learned by analysing definitions, that is, by restricting the intellectual operation to the deductive and inductive modes of inference, because in these cases the possible knowledge is already included in the premises' (Rytövuori-Apunen, 2009: 644).

Questioning the definitional approach implies a distinct perspective on working with concepts in interpretive research. The meanings of key concepts and the concepts themselves 'emerge' *in situ* as the researcher learns what is meaningful to situated members, rather than being defined *a priori* and brought to the field to be tested (Schwartz-Shea and Yanow, 2012: 38–40). In other words, instead of prior identification of concepts and their definition and operationalisation into variables, interpretivists learn the key concepts while in the field. This also applies to researchers working with text, as they learn meaning by analysing how concepts are used and positioned. Zirakzadeh (2009) provides an instructive account of having to abandon the concept of 'modernisation' in his research on ETA (Euskadi Ta Askatasuna), a Basque nationalist liberation organisation, in Pamplona. As his fieldwork progressed, politics no longer appeared to take place amid a large-scale transformation of a premodern society into a modern one but involved street-level episodes and small, local organisations (Zirakzadeh, 2009: 106). 'Modernisation' failed to capture the multi-dimensional and programmatically eclectic local politics. He then also decided to abandon an associated survey he had designed before the fieldwork, as it revealed to have little to do with how nationalists and local residents in general understood their political disagreements (Zirakzadeh, 2009: 104). As he grew despondent in the field, an unexpected

event ‘abducted’ his attention towards the local ‘asociaciones de vecinos’ (community or neighbourhood associations) representing the heterogeneity of local politics. This ‘hook’ made him redesign the research ‘on the fly’ around a locally meaningful concept to capture participants’ understandings and choices *in situ*. Howard Becker (1998) formulated this strategy as letting the case define the concept, in contrast to bounding the case by a conceptual fiat. The research on meaning within such a strategy is about working back from use rather than testing a definition from outside of the context. Given the contextual use of cases in interpretivist projects, where the field is constructed over time through activities that take an episodic and fluid character (Amit, 2000), interpretivists prefer to speak of ‘casing’ (Soss, 2018) rather than case selection. In contrast to a prior identification of cases, strategically in qualitative-positivist research or through random sampling in quantitative-positivist research, casing relies on situational access and cultivating research relationships. It reflects a developmental approach that corresponds to learning in the field, avoiding a ‘lock-in and tunnel vision’ that case selection risks (Soss, 2018: 26).

Such risk is particularly substantial given the basic premise that meaning is polysemic. Although participants’ understandings and sense-making are at the forefront of interpretive research, we should be sceptical about shared meaning. Making sense of difference in interpretation is instead a crucial analytical entry point. Soss (2006) discusses this commitment in relation to his study on adopting the stigmatised identity of ‘welfare’ clients. The breakthrough in his research happened in the realisation of how differently his participants understood the term ‘number’, as in the formulations: ‘you are just a number’ and ‘I felt like a number’. For some, it evoked the angry feeling of being insignificant; for others, it was a welcome sign of anonymity and therefore privacy of their welfare claims. In his research on ‘democracy’ in

Senegal, Schaffer (1998) demonstrates the different meanings attached to this term in the Senegalese society. Senegalese elites invoked the word ‘democracy’ in ways similar to the usage of many political scientists, as a democratic system in which elections are contested and outcomes uncertain. Lower-class Senegalese used the Wolof equivalent, *demokaraasi*, to mean ‘equality’ or the attainment of ‘collective economic security via mutuality’ (Schaffer, 1998: 85). Such differences translate into crucial political distinctions. Yet they could hardly be captured by surveys or structured interviews, which are framed in the language of the researcher, who decides on the range of meanings attributed to democracy. Insight about difference in meaning can be obtained through conversational interviewing that follows the lead of the participant, while situating them in the wider context. Schaffer (2006) resorts to ‘ordinary language’ interviewing for this task: because the meaning of the word is in its use rather than in a dictionary, we need to study how people use the word in different contexts. The various uses of the word may not, however, fit together neatly. As in the examples above, consistency of language does not necessarily mean uniformity of meaning or equivalence across groups. To construct a coherent account of participants’ understandings and ascertain their sources and consequences, it is useful, Soss (2006) recommends, to start with the following questions: what frameworks do your participants use to make sense of their situation? How are individual comments part of a whole, and how are they commentaries on one another?

To be granted access to people’s experiences and lives for the purpose of research is a privilege that needs to be negotiated and accounted for, however. Interpretivists see reflexivity as a way of becoming more grounded in this reality. Reflexivity brings out the researcher’s role in the construction of the research problem and thus exposes how knowledge is marked by its origins. It involves ‘a keen awareness of, and theorizing

about, the role of the self in all phases of the research process' (Schwartz-Shea, 2006: 102). Reflexivity shows the intertwining of accessing, generating, and analysing data in the production of academic claims to knowledge. In his account of life on the shop floor in Alexandria, Shehata (2006) demonstrates at length how researcher identities simultaneously generate dynamics of situational inclusion and exclusion. Being male and perceived as Muslim enabled entrée to some circumstances in his Egyptian factory studies, while shutting off access to other potentially research-related settings, notably Christian Egyptians and women. Bringing to light one's position in the space of knowledge production, together with the interests and confines it may impose on research practice, helps increase trustworthiness, and as such, reflexivity constitutes the interpretive counterpoint to positivist objectivity (Schwartz-Shea and Yanow, 2012: 100). Strengthening the ethical integrity of research is one important function of reflexivity, but reflexivity may also help theorising informal observations. Becoming more aware of one's own sense-making can turn unplanned observations in the field into valuable sources of data and insight (Fujii, 2016: 1150). As 'there is never nothing going on', any mundane moment in the field can be revealing about the context and a broader social order: Shehata's (2006) realisation in the process of research, that social class and organisational and cultural expectations precluded his wearing sandals, was, for example, simultaneously an evidence-generative and data-analytical moment.

## INTERPRETIVE RESEARCH ACTION

In what follows, we discuss some examples of interpretivist research strategies. Given the wealth of works in this tradition, this cannot be more than a glimpse of some interpretive research in action. We showcase the range and diversity of interpretivist sensibility

through selected examples, some of them fieldwork-based, others working with document-analytical strategies, others again with arts-based methodologies. With regard to fieldwork-based interpretivist research, we first sketch the methodological innovation of Lee Ann Fujii (2008, 2010, 2011, 2015, 2016, 2018) in her interpretive work on violence. Fujii's oeuvre illustrates a comprehensive research agenda that attends to methodological detail at every level of interpretive explanation. Subsequently, Xymena Kurowska (2014, 2019) probes the notion of reflexivity by discussing the polysemy of meaning during her own 'research-at-home' among Polish border guards. Second, we discuss examples from the wealth of approaches to discourse analysis in the field of interpretive policy analysis (IPA), which illustrate interpretivist approaches that work with different types of texts, while not necessarily involving fieldwork. Especially, we draw on examples from IPA's focus on policy narratives, metaphors, frames, and myths (cf. Münch, 2016), and hear from some of Berit Bliesemann de Guevara's (2015, 2016) work on urban legends of international intervention and myths in international politics. Finally, Berit Bliesemann de Guevara describes her use of arts-based methods for interpretive scholarship (Gameiro et al., 2018; Bliesemann de Guevara and Arias Lopez, 2019), arguing that if employed in observing the principles of interpretivist methodology, such creative methods hold particular potential to unearth the polysemy of meaning highlighted in hermeneutical approaches.

### *Examples of Interpretivist Research Based on Fieldwork*

Fujii's (2011) research on the Rwandan genocide harnesses localised interpretations that add contextual nuance to our knowledge about violent conflict. Her research strategy brings out the importance of reconstructing

the local context and local meanings for making sense of violence. It relies on local stories, however they are not taken at face value but instead situated within the context of their social milieu and the research encounter itself. Her central research question, ‘how do ordinary people come to commit genocide against their neighbours?’, addresses the deadlock of much ethnicity-based literature, which cannot explain the different pathways that lead to mass violence or the different forms that participation in such violence takes over time and place. Fujii (2011) shows how social ties and immediate social context explain the processes through which ordinary people came to commit mass murder in Rwanda. It was not, she explains, the primordial hatred, as the ‘ethnic conflict’ literature would have it, but rather local ties that facilitated recruitment to violence. Once involved, participants associated such violence as being part of their identity, which contributed to their ongoing participation. Specifically, Hutu leaders used family ties to recruit male relatives, while ties among members of the killing groups helped to initiate reluctant or hesitant members into committing violence with the group. Yet ties of friendship also attenuated murderous actions, leading Hutu killers to help save Tutsis in specific contexts. Which ties became salient depended on the context. Key to Fujii’s understanding of these local mechanisms were two components: the contextual employment of Granovetter’s (1985) concept of ‘social embeddedness’, and a field-specific set of questions, namely: ‘who were the peasant killers known as Interahamwe? How did they come to participate in the genocide? And why would some Interahamwe try to help the very people they were supposed to kill?’ (Fujii, 2008).

There are at least three methodological contributions that crystallise in Fujii’s research process and help thinking about interpretive research in action. First is the relational approach to interviewing, where data are co-generated in research relationships (Fujii,

2018). Relational interviewing engages participants in a two-way dialogue, which is shaped by the particular context in which it occurs, as well as by the interests, beliefs, and backgrounds that each party brings to the exchange. It is, Fujii claims, through these interactions that data emerge. Their value lies not in their factual accuracy but in what they convey about the speakers’ worlds and how they experience, navigate, and understand them. The basic premise of such an approach is its open design: the researcher learns during encounters with and in the field. Mistakes are gifts in this process, as they reveal flawed assumptions and lead to final conceptualisations over the course of the research process, rather than prior to it. Fujii is adamant, however, about the limitations to such relationships and prefers to call them ‘working relationships’ (Fujii, 2018: chapter 2) rather than rapport. The latter implies familiarity and intimacy, which is not only rare but usually illusive. The researcher neither has the power to cast the locals in the role of research subjects, if they are not interested in participation, nor controls how they themselves are received in the local setting – although such reception affects data generation. Second, Fujii (2010) points to the analytical value of what she calls ‘meta-data’, that is, spoken and unspoken thoughts and feelings that participants do not always articulate in interviews, but which emerge in other ways. She speaks of five types of such data – rumours, inventions, denials, evasions, and silences – which are integral to the processes of data generation and analysis. Meta-data are important as they indicate the operating rules of the community and also reveal features of the research relationship that shape data generation. In violent conflict environments, participants may want to diminish their complicity in atrocities. Yet they may also want to embellish some components of their belonging. Third, Fujii introduces the notion of ‘accidental ethnography’ in political science (Fujii, 2015). Accidental ethnography involves paying systematic attention to the unplanned moments that take

place outside of the scripted research situation. Such moments point to the larger political and social worlds that help one understand the context in which both the researcher and the researched are embedded.

Engaging in accidental ethnography often relies on reflexivity by the researcher. However, and as also indicated by Fujii in her caution about rapport, reflexivity should not be regarded as an automatic resolution of interpretive concerns. Finlay (2002) points to the rhetorical function of reflexivity as a claim to authority and credibility. Pillow (2003) is uncomfortable with validated and thus comfortable reflexive strategies, which involve reflexivity as 'researcher know thyself', reflexivity as recognition of the other, reflexivity as truth, and reflexivity as transcendence. She suggests instead that we engage in 'reflexivities of discomfort', which interrupt uses of reflexivity as a tool of methodological power (clarity, honesty, humility). Rose (1997) formulates a more radical critique against what she sees as a problem of 'transparent reflexivity'. Transparent reflexivity emerges when the researcher claims being able to ascertain both how their own subjectivity translates into data and how the process of knowledge production can be charted within the broader landscape of power.

In this respect, Kurowska's (2019) research-at-home among Polish border guards who had been trained by their German colleagues to become 'European' border guards presented certain dilemmas. An indigenous researcher, that is, one 'native' to the site, arguably possesses a distinct advantage of cultural competence that facilitates insight. We usually associate a variety of interrelated advantages to insidership, such as the value of shared experiences, the value of greater access, the value of cultural interpretation, and the value of deeper understanding and clarity of thought for the researcher (Labaree, 2002: 103). It should therefore follow that Kurowska, being Polish, had a distinct advantage for gleaning meaning about the lived

experience of transformation in the region. And yet, in fieldwork encounters, it soon transpired that this may not be straightforwardly the case. One of the revealing moments happened when a former border-guard trainer winked at the researcher as she was leaving his office and said, 'we are all Germans, no?' There was no way of knowing the meaning of this particular wink. But there was irony that pervaded her interviewees' own recognition of being hailed into becoming European border guards. The acknowledgement of power relations that were hardly in their favour was an occasion to use, play, ridicule, and distinguish themselves in a way that did not deny the 'civilisation shift' but re-appropriated it in local ways. The Polish border guards claimed superiority of having deliberately accepted the initial patronising of their German colleagues to see 'what's on the other side [of the border]' (Kurowska, 2014: 558). Their keen sense of tension between emancipation and paternalism may have exceeded the researcher's range of reflexivity. On this occasion, she may have been the most uncomfortable, although in an unclear way, with the interviewee putting her in the same (ironised) category of 'an aspiring German'.

Such encounters first bring out polysemy of meaning, which is not easily captured even by a 'native' researcher, and, second, point to the intersubjective character of interpretive knowledge production: participants may show the researcher her own co-optation into normative discourses, including scholarly frames. The kind of reflexivity that emerges in such situations is intersubjective, rather than stemming from the researcher's introspection. It relies on the notion that the shifting self is always articulated through specific social interactions and that what we research is our relations with the researched (Rose, 1997: 313). This reverses the hierarchy of expertise in interpretive research as compared with the positivist stance: while the latter posits the researcher as an expert in control of both research subjects and the research process, an interpretivist is a supplicant who

acknowledges that the research subjects have a greater knowledge of the nuances of meaning that structure their everyday lives (England, 1994: 82).

### ***Examples of Interpretive Policy Analysis***

Studying practices of naming and framing is also a core research strategy that interpretive policy analysts have developed to reconstruct meaning-making in public discourse on political ‘problems’. IPA takes issue with policy analysis’ preoccupation with policy-making as problem-solving, epitomised most schematically in the ‘policy cycle’ (Barbehön et al., 2015). Rather, IPA authors draw attention to the processes through which an issue becomes defined as (a specific type of) actionable ‘problem’ in the first place. Interpretive policy analysts, just as interpretivists in general, differ in their approaches both with regard to their poststructuralist or hermeneutical positioning and the specific methods through which they study discourses. Yet they all share an analytical interest in the reconstruction of how certain social interpretations become institutionalised as more or less collectively binding and thereby legitimate. Authors associated with the ‘argumentative turn’ (Fischer and Forester, 1993; Fischer and Gottweis, 2012) or ‘interpretive turn’ (Healy, 1986; Yanow, 1995) in policy analysis have suggested portraying policy-making as an ‘ongoing discursive struggle over the definition and conceptual framing of problems, the public understanding of the issues, the shared meanings that motivate policy responses, and criteria for evaluation’ (Fischer and Gottweis, 2012: 7).

Interpretive approaches to policy analysis do not see political problems as objectively given but focus on the processes by which certain issues become problematised, that is, discursively rendered into politically addressable problems (Barbehön et al., 2015). Problematisations can be conceived of, in

a Foucauldian poststructuralist tradition, as products of larger power/knowledge regimes that precede subjects and let certain ways of thinking about a political issue appear as apparent and factual. The researcher’s task here is to reconstruct the contingent historical processes that have led to the problematisation and that are anything but necessary or definitive – a methodology that Foucault has called genealogy. This is usually combined with a critique of the order that had allowed this specific problematisation, with all its silences and exclusions, come about. Bacchi’s (2012) ‘what’s the problem represented to be’ approach, for example, suggests a range of questions aimed at uncovering the unspoken alternatives and silences of specific problematisations, in order to reveal the larger power/knowledge regime shaping actors’ room for political imagination: which assumptions underpin the representation of the ‘problem’? What is not being problematised? How could we think differently about the ‘problem’? What effects does this representation of the ‘problem’ have?<sup>6</sup>

Hermeneutical reconstructions of problematisations have focused more strongly on those elements that structure discourse and have given actors’ discursive struggles over problem definitions more room, for example, by studying the role of frames, narratives, metaphors, myths, and categories in the process of political problematisation (Goffman, 1974; Yanow, 1992; Hajer, 1997; Stone, 2002). For Rein and Schön (1993), frames are perspectives through which an amorphous and diffusely defined problematic situation makes sense in a specific way. Different frames applied to one and the same situation will create differing social realities with different effects on subsequent ‘problem-solving’ attempts. From a frame-analytical perspective, political struggles take place not over the political aims and measures to solve a given problem but over the framing of the problem as such: is homelessness, for example, an individual problem connected to medical causes, such as mental health, or a

governance problem connected, for example, to a lack of housing or welfare due to economic austerity? Is drug trafficking supplier- or consumer-driven, and is it predominantly a legal, medical, or political problem? (Münch, 2016: 80).

Metaphors are another structuring element of policy discourse that interpretive policy analysts study. Metaphors work by transferring concepts including their meaning from one context into a different one, where its transferred meaning takes effect. They are often used to simplify complexity by placing abstract concepts into contexts of everyday life. Interpretivists analyse metaphors not as mere rhetorical instrument but as powerful discursive meaning-making strategies with tangible effects on how a problem is understood (Lakoff and Johnson, 1980; Maasen, 2000; Gronau and Schneider, 2009). Gronau and Nonhoff (2011) have shown, for example, how dressing the 2008 financial crisis into metaphors of natural catastrophe helped deflect responsibility from bankers' purposeful actions to a 'global financial system', which appeared like a *force majeure*. Other often used fields for metaphorical concept transfers in politics are medicine (metaphors of crisis, intervention, corruption as disease, failing states as patients, etc.), engineering (the state as ship, the bureaucratic machinery), and warfare (war on drugs, defence of the Euro zone) (Münch, 2016: 97–8).

For Stone (2002), metaphors are elements of larger policy narratives that can be reconstructed by looking across different texts. Narratives make sense of the social world through employment: they connect discrete events with each other and thereby do not only order them chronologically (beginnings, middles, ends) but also causally. Narratives ascribe roles such as hero, villain, and victim to actors and thus distribute responsibility and blame. In her analysis of causal stories that structure the discourses on ethnic segregation in the housing sectors in Germany and the UK, Münch (2010) shows that in both countries, the dominant explanatory narrative

blames ethnic segregation of minorities in cities on the voluntary decision of members of these minorities to retreat into ethnically defined neighbourhoods. She then explores alternative explanations of ethnic segregation, ranging from political decisions on housing and social policies, community-council governance, economic considerations of private landlords, and the conscious segregation of housing associations to the strategies of the majority population and general demographic change. She concludes that while the reasons for ethnic segregation are mainly structural, the causal stories remain focused on segregation as a result of human behaviour, since it makes the policy problem of 'segregation' politically more easily addressable than complex structures would allow. Struggles over dominant policy narratives do not only occur in textual form; often they are accompanied by rituals, symbols, and performances in conjunction with which they gain authority. Bliesemann de Guevara (2017) analyses policy narratives and performances around political decisions regarding Germany's participation in the military intervention in Afghanistan and the subsequent question of the withdrawal of troops despite ongoing violent conflict in the country. She shows how politicians employ different narratives to justify their decisions and how they use the staging of insider knowledge through troop visits in Afghanistan to give the weight of 'having been there' and 'seen with one's own eyes' to their interpretations of the 'situation on the ground'.

The analysis of policy narratives that manifest in public discourse raises the question of which sources actually count as text to be studied. In order to reconstruct the way in which UN peacebuilders make sense of the situation of intervention they find themselves in, official documents that adhere to specific templates of writing reports, evaluating projects, and briefing superiors will say a lot about formal organisational practices but reveal little about the sense-making of the individuals who author these reports. Increasingly aware



of the limits of formal interviews and document analyses, and in line with Fujii's idea of meta-data as discussed above, Bliesemann de Guevara and Kühn (2015), for example, started to pay more attention to 'private' conversations with UN peacebuilders, which were replete with anecdotes that recurred in slightly altered forms across space and time. These 'urban legends of intervention', seemingly harmless and entertaining episodes that happened to 'a colleague of a colleague', revealed a deeper level of meaning-making in peacebuilders' own roles, their interactions with 'the locals', and the failures of peacebuilding. The urban legends revolved around three themes: 'the intervened' as barbarians; 'the interveners' as plagued by western/northern hubris; and intercultural interactions fraught with cultural misunderstandings. These meta-narratives resonated in part with much older colonial tropes of encounters between 'civilised explorers' and 'barbarian colonised' and revealed the deeply engrained orientalist thinking on which international peacebuilding interventions are built. Myths are another category of particular narratives – powerful foundational social narratives or widely held beliefs – which have been studied by interpretivist PS&IR scholars as meaning-making devices that help legitimise social order and political practices (contributions in Bliesemann de Guevara, 2016).

### ***Examples of Using Arts-Based Methods in Interpretive Research***

In her reflection on myths in politics, Yanow (2016) refers to Polanyi's (1967: 306) observation that '[t]he fact that we can possess knowledge that is unspoken is of course a common-place and so is the fact that we must know something yet unspoken before we can express it in words'. In order to explore such 'tacit knowledge', arts-based research methods are particularly well suited to help unearth research participants ideas, feelings, and meaning-making. Gameiro and colleagues

(2018), for example, have developed a metaphor-centred drawing method to explore infertility and healthcare experiences among Black and Minority Ethnic (BME) women in Cardiff, whereby both the non-verbal mode of expression and the possibility to express something intangible, such as feelings or ideas, in a tangible metaphor provided ways of rendering unspoken or sensitive knowledge communicable, thus not only generating data but also sensitising the research to the polysemy of meaning that BME women attached to their talk about fertility challenges and experiences with the UK healthcare system. In their study on experiences of violent conflict in Myanmar, Julian and colleagues (2019) have adapted this method to hermeneutically explore core concepts such as 'conflict', 'peace', and 'protection', and to understand the experiential knowledge around these concepts that civilians living amidst political violence hold. Among other things, the research reveals the polysemic use of these concepts, which differs considerably between communities in areas of violent conflict, other communities in Myanmar, and international humanitarian organisations in the country – a difference that remains hidden in analyses of these organisations because they operate with preconceived concepts and also because of the sensitivity of the violent experiences that have formed local understandings. 'Peace', for example, was seen as a 'bad word' by some communities in Myanmar's Kachin state, who associated 'peace' with a ceasefire period that brought economic exploitation and a loss in autonomy to Kachin people. Key to the analysis is again the context, which gives rise to different meanings of terms in different parts of the country, depending on specific historical uses and people's experiences and associations – revealed in this research through the use of drawing as a non-verbal method that allows the 'drawing out' of those differences.

Against the background of limitations of written and spoken language to express full meaning, arts-based research methods promise a useful extension to the qualitative

methods employed by interpretivists. How we know and make sense of the world is also attached to embodied, emotional, and affective ways of knowing, which pose an even more profound challenge to the researcher than the polysemy of meaning. Feelings and abstract ideas and thoughts are difficult to objectify in words, which nonetheless form the core of much of our qualitative social-scientific methods. Arts-based methods have been shown to offer an epistemic way into these other dimensions of hermeneutical knowing. Textiles thematising experiences of violent conflict, for example, have been shown to voice, in their function as object witnesses, the difficult knowledges they bear in documentary, visual, symbolic, material, and sensory registers, some of which are specific to their textile quality. Furthermore, curating such conflict textiles as a methodology of caring for difficult knowledge has been argued to avoid interpretive closures of the political violence addressed in them and to bring about an affective force in their audiences, with the possibility of resulting in a transformation of thought and perhaps even action among the latter (Andrä et al., 2019). Yet the use of textiles in research can go further to include interaction and exchange between different groups of research participants. In a project exploring the subjectivities of former members of armed groups in Colombia, the research team led by Bliesemann de Guevara and Arias Lopez (2019) uses textile narratives to allow former armed actors to unstitch and restitch meaning around their own role in Colombia's war and postwar periods and other members of Colombian society to resonate with their stories.<sup>7</sup> Inspired by Bruno Munari's 'illegible books' (cf. Maffei, 2015), which the Italian modernist artist developed to communicate non-verbal ideas, the textile narrative method chosen in the project invites participants to sew and embroider their stories, experiences, thoughts, or feelings in cloth. These 'textile pages' authored by different types of actors are then arranged into thematic textile

books, which allow onlookers to appreciate the original textiles and their meaning, but also to deconstruct this original meaning when read in conjunction with others, and construct one's own signification or story. In exhibitions of the textiles books, visitors are furthermore invited to embroider their reactions and answers to the textile narratives. The arts-based element thus turns individual textile accounts into powerful means of non-verbal communication and exchange, with the potential to itself make a social intervention into societal meaning-making processes, in order to contribute to peace. Perhaps this example can serve as an invitation to take interpretivist approaches in PS&IR into a new, more activist direction.

## CONCLUSION

Interpretivist approaches to PS&IR see politics, from the global to the state/society and to the local, not in terms of causal laws but as contingent, shifting practices that can be examined through research for the meanings in action that are attached to them (Bayard de Volo, 2016: 244). Interpretivists go beyond the institutionalist understanding of politics to study politics 'from below' and 'from within' and, in the process, make sense of power relations in particular political settings. They are interested in how particular meanings produce and organise power in specific ways. In this chapter, we have shown that rather than constituting a set or toolbox of methods, interpretivism is a methodology that rests on four pillars: the search for polysemic meanings, hermeneutic or poststructuralist analysis that includes the very concepts it studies, an abductive logic of inquiry, and reflexivity. Individual studies will differ in exactly how these principles are employed and which ones are foregrounded, as our research action examples have shown. What all these studies nonetheless have in common is a rejection of the idea of a world

out there that can be objectively known and that the researcher is not enmeshed in. From an interpretivist perspective, the idea of academia as an 'ivory tower' does not make sense – apart from being a powerful, sense-making metaphor used in public discourse to establish or contest a certain hierarchical order of expertise and knowledge. The interpretivist scholar is part of the world she studies, and her methods will both seek to account for the meaning her research participants share with her and reflect on the limits of this sharing. In this lies the rigour of interpretivist research.

## Notes

- 1 We would like thank Sybille Münch and Chiara Ruffa for their comments on earlier drafts of this chapter. Kurowska's work on this chapter was funded through European Commission MSCA Individual Fellowship RefBORDER grant no. 749314.
- 2 On monism and dualism in PS&IR, see Jackson (2011).
- 3 The Latin etymology of the word 'fact' is the verb 'facere', which means 'to do' or 'to make', which, as Lynch (2014: 25, nt. 4) argues, connotes action or construction rather than unchanging truth.
- 4 For a comprehensive overview of a variety of interpretive research methods in use, see 'Introduction' in Yanow and Schwartz-Shea (2014), specifically Table I-1 on p. xvi.
- 5 See Kurowska and Tallis (2013) for an example of research that uses the notion of 'chiasmatic crossings' to describe relational knowledge production that amounts to contextual co-authorship in fieldwork.
- 6 Due to interpretivists' general lack of a substantive critique that deconstructs the logic behind social representations of problems, elaborated earlier in this chapter, many authors of the post-structuralist strand of 'problematization' would indeed not count themselves as 'interpretivists' (cf. Bacchi, 2015).
- 7 Newton Fund/Colciencias project '(Un)Stitching the Subjects of Colombia's Reconciliation Process', 2018–2020 (AHRC project reference AH/R01373X/1; Colciencias project reference FP44842-282-2018), hosted by Aberystwyth University, UK, University of Antioquia, Medellín, Colombia, and the Association of Victims and Survivors of Northeast Antioquia, Colombia (see: <https://gtr.ukri.org/projects?ref=AH%2FR01373X%2F1>).

## REFERENCES

- Ahmed, S. N. (2018) Tweet by @feministkilljoy, 25 September 2018, *Twitter*, <https://twitter.com/SaraNAhmed/status/1044564472944316416> (accessed 31 October 2018).
- Allina-Pisano, J. (2009) 'How to Tell an Axe Murderer: An Essay on Ethnography, Truth, and Lies'. *Political Ethnography. What Immersion Contributes to the Study of Power*. E. Schatz. Chicago, IL and London, Chicago University Press, pp. 53–73.
- Amit, V., Ed. (2000) *Constructing the Field: Ethnographic Fieldwork in the Contemporary World*. London and New York, Routledge.
- Andrä, C., B. Bliesemann de Guevara, L. Cole and D. House (2019) 'Knowing Through Needlework: Curating the Difficult Knowledge of Conflict Textiles'. *Critical Military Studies*: <https://doi.org/10.1080/23337486.2019.1692566>.
- Aradau, C. and J. Huysmans (2014) 'Critical Methods in International Relations: The Politics of Techniques, Devices and Acts'. *European Journal of International Relations* 20(3): 596–619.
- Bacchi, C. (2012) 'Why Study Problematizations? Making Politics Visible'. *Open Journal of Political Science* 2(1): 1–8.
- Bacchi, C. (2015) 'The Turn to Problematization: Political Implications of Contrasting Interpretive and Poststructural Adaptations'. *Open Journal of Political Science* 5(1): 1–12.
- Barbehön, M., S. Münch and W. Lamping. (2015) 'Problem Definition and Agenda-Setting in Critical Perspective'. *Handbook of Critical Policy Studies*. F. Fisher, D. Torgerson, M. Orsini and A. Durnova. Cheltenham, Edward Elgar, pp. 241–258.
- Bayard de Volo, L. (2016) 'Comparative politics'. *Routledge Handbook of Interpretive Political Science*. M. Bevir and R. A. W. Rhodes. London and New York, Routledge, pp. 241–255.
- Becker, H. (1998) *Tricks of the Trade: How to Think about Your Research While Doing It*. Chicago, IL, University of Chicago Press.
- Bever, M. (2000) *Interpretive Political Science*. London, Sage.
- Bever, M. (2014) 'How Narratives Explain'. *Interpretation and Method: Empirical Research Methods and the Interpretive Turn*. D. Yanow

- and P. Schwartz-Shea. Armonk, NY and London, M.E. Sharp, pp. 281–290.
- Bevir, M. and R. A. W. Rhodes (2006) 'Defending Interpretation'. *European Political Science* 5(1): 69–83.
- Bevir, M. and R. A. W. Rhodes, Eds. (2016) *Routledge Handbook of Interpretive Political Science*. London and New York, Routledge.
- Bevir, M. and R. Phillips, Eds. (2019) *Decentring European Governance*. London and New York, Routledge.
- Bliesemann de Guevara, B. Ed. (2016) *Myth and Narrative in International Politics: Interpretive Approaches to the Study of International Relations*. London, Palgrave Macmillan.
- Bliesemann de Guevara, B. (2017) 'Intervention Theatre: Performance, Authenticity and Expert Knowledge in Politicians' Travel to Post-/conflict Spaces'. *Journal of Intervention and Statebuilding* 11(1): 58–80.
- Bliesemann de Guevara, B. and B. Arias Lopez (2019) 'Biographical narratives and textile art: a research example from the Colombian peace process'. *Stitched Voices blog*, 15 August. <https://stitchedvoices.wordpress.com/2019/08/15/biographical-narratives-and-textile-art-colombia/> (accessed 13 October 2019).
- Bliesemann de Guevara, B. and F. P. Kühn (2015) 'On Afghan Footbaths and Sacred Cows in Kosovo: Urban Legends of Intervention'. *Peacebuilding* 3(1): 17–35.
- Burawoy, M. (1998) 'The Extended Case Method'. *Sociological Theory* 16(1): 4–33.
- Cohn, C. (1987) 'Sex and Death in the Rational World of Defense Intellectuals'. *Signs* 12(4): 687–718.
- Cohn, C. (2006) 'Motives and Methods: Using Multi-sited Ethnography to Study US National Security Discourses'. *Feminist Methodologies for International Relations*. B. Ackerly, M. Stern and J. True. New York, Cambridge University Press, pp. 91–107.
- Cox, R. (1981) 'Social Forces, States and World Orders: Beyond International Relations Theory'. *Millennium* 10(2): 126–155.
- England, K. V. L. (1994) 'Getting Personal: Reflexivity, Positionality, and Feminist Research'. *The Professional Geographer* 46(1): 80–89.
- Epstein, C. (2008) *The Power of Words in International Relations: Birth of an Anti-whaling Discourse*. Cambridge, MA and London, MIT Press.
- Fenno, R. F. (1986) 'Observation, Context, and Sequence in the Study of Politics'. *American Political Science Review* 80(1): 3–15.
- Finlay, L. (2002) 'Negotiating the Swamp: the Opportunity and Challenge of Reflexivity in Research Practice'. *Qualitative Research* 2(2): 209–230.
- Fischer, F. and J. Forester (1993) *The Argumentative Turn in Policy Analysis and Planning*. Durham, NC, Duke University Press.
- Fischer, F. and H. Gottweis (2012) *The Argumentative Turn Revisited: Public Policy as Communicative Practice*. Durham, NC, Duke University Press.
- Fujii, L. A. (2008) 'The Power of Local Ties: Popular Participation in the Rwandan Genocide'. *Security Studies* 17(3): 568–597.
- Fujii, L. A. (2010) 'Shades of Truth and Lies: Interpreting Testimonies of War and Violence'. *Journal of Peace Research* 47(2): 231–241.
- Fujii, L. A. (2011) *Killing Neighbors: Webs of Violence in Rwanda*. Ithaca, NY, Cornell University Press.
- Fujii, L. A. (2015) 'Five Stories of Accidental Ethnography: Turning Unplanned Moments in the Field into Data'. *Qualitative Research* 15(4): 525–539.
- Fujii, L. A. (2016) 'Politics of the "Field"'. *Perspectives on Politics* 14(4): 1147–1152.
- Fujii, L. A. (2018) *Interviewing in Social Science Research: A Relational Approach*. New York, Routledge.
- Gameiro, S., B. Bliesemann de Guevara, E. El Refaie & A. Payson (2018) 'DrawingOut – An Innovative Drawing Workshop Method to Support the Generation and Dissemination of Research Findings'. *PLoS ONE* 13(9): e0203197.
- Geertz, C. (1993) *The Interpretation of Cultures: Selected Essays*. London, Fontana.
- Giddens, A. (1990) 'Structuration Theory and Sociological Analysis'. *Anthony Giddens: Consensus and Controversy*. J. Clark, C. Modgil and S. Modgil. London, Falmer Press, pp. 297–315.
- Ginger, C. (2014) 'Interpretive Content Analysis: Stories and Arguments in Analytic Documents'. *Interpretation and Method: Empirical Research Methods and the Interpretive Turn*. D. Yanow and P. Schwartz-Shea. Armonk, NY, M.E. Sharp, pp. 331–348.

- Goffman, E. (1974) *Frame Analysis: An Essay on the Organization of Experience*. New York, Harper & Row.
- Granovetter, M. (1985) 'Economic Action and Social Structure: The Problem of Embeddedness'. *American Journal of Sociology* 91(3): 481–510.
- Gronau, J. and M. Nonhoff (2011) Von Schurken und Schlampern: Metaphorische Verdichtungen von Erzählungen internationaler Finanzkrisen. Unpublished paper prepared for the 3rd Open Session, International Relations Section of the German Political Science Association, Munich.
- Gronau, J. and S. Schneider (2009) Metaphorical Concepts in the Construction of International Legitimacy. Working Paper Series of the Committee on Concepts and Methods 37/2009, International Political Science Association (IPSA).
- Hajer, M. A. (1997) *The Politics of Environmental Discourse: Ecological Modernization and the Policy Process*, New York, Oxford University Press.
- Hansen, L. (2007) *Security as Practice: Discourse Analysis and the Bosnian War*. London and New York, Routledge.
- Howarth, D. & Griggs, S. (2012) 'Poststructuralist Policy Analysis: Discourse, Hegemony, and Critical Explanation'. *The Argumentative Turn Revisited: Public Policy as Communicative Practice*. F. Fischer and H. Gottweis. Durham, NC, Duke University Press, pp. 305–342.
- Healy, P. (1986) 'Interpretive Policy Inquiry: A Response to the Limitations of the Received View'. *Policy Sciences* 19(4): 381–396.
- Heinelt, H. and S. Münch, Eds. (2018) *Handbook of European Policies. Interpretive Approaches to the EU*. Cheltenham and Northampton, MA, Edward Elgar Publishing.
- Hellmann, G. (2009) 'Pragmatism and International Relations'. *International Studies Review* 11(3): 638–641.
- Jackson, P. T. (2011) *The Conduct of Inquiry in International Relations: Philosophy of Science and Its Implications for the Study of World Politics*. New York and London: Routledge.
- Julian, R., B. Bliesemann de Guevara & R. Redhead (2019) 'From Expert to Experiential Knowledge: Exploring the Inclusion of Local Experiences in Understanding Violence in Conflict'. *Peacebuilding* 7(2): 210–225.
- Klein, B. S. (1994) *Strategic Studies and World Order*. Cambridge, Cambridge University Press.
- Kondo, D. K. (1986) 'Dissolution and Reconstitution of Self: Implications for Anthropological Epistemology'. *Cultural Anthropology* 1(1): 74–88.
- Kurowska, X. (2014) 'Practicality by Judgment: Transnational Interpreters of Local Ownership in the Polish-Ukrainian Border Reform Encounter'. *Journal of International Relations and Development* 17(4): 545–565.
- Kurowska, X. (2019) 'When Home is Part of the Field: Experiencing Uncanniness of Home in Field Conversations'. *Tactical Constructivism: Expressing Method in International Relations*. B. Steele, H. Gould and O. Kessler. London and New York, Routledge.
- Kurowska, X. and B. C. Tallis (2013) 'Chiasmatic Crossings: A Reflexive Revisit of a Research Encounter in European Security'. *Security Dialogue* 44(1): 73–89.
- Labaree, R. (2002) 'The Risk of "Going Observationalist": Negotiating the Hidden Dilemmas of Being an Insider Participant Observer'. *Qualitative Research* 2(1): 97–122.
- Lakoff, G. and M. Johnson (1980) *Metaphors We Live By*. Chicago, IL and London, University of Chicago Press.
- Law, J. (2004) *After Method: Mess in Social Science Research*. London and New York, Routledge.
- Lynch, C. (2014) *Interpreting International Politics*. New York, Routledge.
- Maasen, S. (2000) *Metaphors and the Dynamics of Knowledge*. London and New York, Routledge.
- Maffei, G. (2015) *Munari's Books*. New York, Princeton Architectural Press.
- Marsh, D. and H. Savigny (2004) 'Political Science as a Broad Church: The Search for a Pluralist Discipline'. *Politics* 24(3): 155–168.
- Milliken, J. (1999) 'The Study of Discourse in International Relations: A Critique of Research and Methods'. 5(2): 225–254.
- Münch, S. (2010) *Integration durch Wohnungspolitik?: Zum Umgang mit ethnischer Segregation im europäischen Vergleich*. Wiesbaden, Springer VS.
- Münch, S. (2016) *Interpretative Policy-Analyse: eine Einführung*. Wiesbaden, Springer VS.
- Nagel, T. (1989) *The View from Nowhere*. Oxford, Oxford University Press.

- Onuf, N. (2013) *World of Our Making. Rules and Rule in Social Theory and International Relations*. London and New York, Routledge, 2nd edition.
- Oren, I. (1995) 'The Subjectivity of the "Democratic" Peace: Changing U.S. Perceptions of Imperial Germany'. *International Security* 20(2): 147–184.
- Paipais, V. (2011) 'Self and Other in Critical International Theory: Assimilation, Incommensurability and the Paradox of Critique'. *Review of International Studies* 37(1): 121–140.
- Pillow, W. (2003) 'Confession, Catharsis, or Cure? Rethinking the Uses of Reflexivity as Methodological Power in Qualitative Research'. *International Journal of Qualitative Studies in Education* 16(2): 175–196.
- Polanyi, M. (1967) 'Sense-Giving and Sense-Reading'. *Philosophy* 42(162): 301–325.
- Rein, M. and D. Schön (1993) 'Reframing Policy Discourse'. *The Argumentative Turn in Policy Analysis and Planning*. F. Fischer and J. Forester. London, Duke University Press, pp. 145–166.
- Rhodes, R. A. W. (2017) *Interpretive Political Science*. Oxford, Oxford University Press.
- Rorty, R. (1989) *Contingency, Irony, and Solidarity*. Cambridge, Cambridge University Press.
- Rose, G. (1997) 'Situating Knowledges: Positionality, Reflexivities and other Tactics'. *Progress in Human Geography* 21(3): 305–320.
- Rytövuori-Apunen, H. (2009) 'Abstractive Observation as the Key to the "Primacy of Practice"'. *International Studies Review* 11(3): 641–645.
- Schaffer, F. C. (1998) *Democracy in Translation: Understanding Politics in an Unfamiliar Culture*. Ithaca, NY, Cornell University Press.
- Schaffer, F. C. (2006) 'Ordinary Language Interviewing'. *Interpretation and Method. Empirical Research Methods and the Interpretive Turn*. D. Yanow and P. Schwartz-Shea. Armonk, NY and London, M.E. Sharpe, pp. 150–160.
- Schaffer, F. C. (2016) *Elucidating Social Science Concepts. An Interpretivist Guide*. New York, Routledge.
- Schwartz-Shea, P. (2006) 'Judging Quality: Evaluative Criteria and Epistemic Communities'. *Interpretation and Method: Empirical Research Methods and the Interpretive Turn*. D. Yanow and P. Schwartz-Shea. Armonk, NY and London, M.E. Sharpe, pp. 89–114.
- Schwartz-Shea, P. and D. Yanow (2012) *Interpretive Research Design: Concepts and Processes*. New York and London, Routledge.
- Shehata, S. (2006) 'Ethnography, Identity, and the Production of Knowledge'. *Interpretation and Method: Empirical Research Methods and the Interpretive Turn*. D. Yanow and P. Schwartz-Shea. Armonk, NY and London, M.E. Sharpe, pp. 244–263.
- Sil, R. and P. J. Katzenstein (2010) 'Analytic Eclecticism in the Study of World Politics: Reconfiguring Problems and Mechanisms across Research Traditions'. *Perspectives on Politics* 8(2): 411–431.
- Soss, J. (2006) 'Talking Our Way to Meaningful Explanations'. *Interpretation and Method*. D. Yanow and P. Schwartz-Shea. New York and London, M.E. Sharpe, pp. 127–150.
- Soss, J. (2018) 'On Casing a Study versus Studying a Case'. *Qualitative and Multi-Method Research* 16(1): 21–27.
- Steele, B., H. D. Gould & O. Kessler, Eds. (2019) *Tactical Constructivism: Expressing Method in International Relations*. London and New York, Routledge.
- Stone, D. (2016) 'Quantitative Analysis as Narrative'. *Routledge Handbook of Interpretive Political Science*. M. Bevir and R. A. W. Rhodes. London and New York, Routledge pp. 157–170.
- Stone, D. (2002) *Policy Paradox: The Art of Political Decision Making*. New York, Norton.
- Tannenwald, N. (1999) 'The Nuclear Taboo: The United States and the Normative Basis of Nuclear Non-Use'. *International Organization* 53(3): 433–468.
- Taylor, C. (1971) 'Interpretation and the Sciences of Man'. *The Review of Metaphysics* 25(1): 3–5.
- Vrasti, W. (2008) 'The Strange Case of Ethnography and International Relations'. 37(2): 279–301.
- Wagenaar, H. (2011) *Meaning in Action: Interpretation and Dialogue in Policy Analysis*. New York and London, M.E. Sharpe.
- Walsh, K. C. (2009) 'Scholars as Citizens: Studying Public Opinion through Ethnography'. *Political Ethnography. What Immersion*

- Contributes to the Study of Power*. E. Schatz. Chicago and London, Chicago University Press, pp. 165–182.
- Wedeen, L. (2010) 'Reflections on Ethnographic Work in Political Science'. *Annual Review of Political Science* 13: 255–272.
- Wilkinson, C. (2014) 'On Not Just Finding What You (Thought You) Were Looking For. Reflections on Fieldwork Data and Theory'. *Interpretation and Method: Empirical Research Methods and the Interpretive Turn*. D. Yanow and P. Schwartz-Shea. New York and London M.E. Sharpe, pp. 387–405.
- Yanow, D. (1992) 'Silences in Public Policy Discourse: Organizational and Policy Myths'. *Journal of Public Administration Research and Theory* 2(4): 399–423.
- Yanow, D. (1995) 'Practices of Policy Interpretation'. *Policy Sciences* 28(2): 111–126.
- Yanow, D. (2003) 'Interpretive Empirical Political Science: What Makes This Not a Subfield of Qualitative Methods'. *Qualitative Methods* 1(2): 9–12.
- Yanow, D. (2016) 'Forward'. *Myth and Narrative in International Politics: Interpretive Approaches to the Study of International Relations*. B. Bliesemann de Guevara. Basingstoke, Palgrave Macmillan pp. v–xii.
- Yanow, D. and P. Schwartz-Shea (2014) *Interpretation and Method: Empirical Research Methods and the Interpretive Turn*. Armonk, NY and London, M.E. Sharp.
- Zirakzadeh, C. (2009) 'When Nationalists are not Separatists: Discarding and Recovering Academic Theories while doing Fieldwork in the Basque Region of Spain'. *Political Ethnography: What Immersion Contributes to the Study of Politics*. E. Schatz. Chicago, IL, Chicago University Press, pp. 97–118.



# Afterword

We conclude our *Handbook's* tour of research methods in political science and international relations with appreciation: appreciation for the tremendous contributions to the volume and to the discipline of these chapters' authors, and of the many colleague scientists they reference; and appreciation also for the great progress and accomplishments in research methodology that have forged an impressive modern science of politics and international relations. From formulating interesting and

important research questions and agendas, to developing powerful positive theories, to effectively conceptualizing and accurately measuring their moving parts, and properly evaluating, precisely estimating, and substantively interpreting their empirical manifestations, modern political science and international relations is indeed an impressive, productive, and useful science, thanks to – and rooted in – these sound research methods.



# Index

- Abadie, Alberto 813, 836  
accelerated failure time model 662–3  
accountability 236–9  
Achen, Christopher H. 776, 899, 941  
Action Against Hunger 63  
active labor-market (ALM) 589  
Adams, James 219  
Adcock, Robert N. 332, 339  
additive scale models 356–9  
agent based models (ABMs) 290–1  
Ajao, O. 1068–9  
Akaike's information criterion (AIC) 603  
Albert, James H. 902, 915  
algorithmic bias 99–100  
algorithms 540–2  
    matching 810–11  
Allison, Paul 719–0  
Alvarez, R. Michael 289–90  
*American Journal of Political Science* 425, 440, 899, 999, 1180  
American National Election Study 212  
American Political Science Association (APSA) 137  
*American Political Science Review, The* 425, 440, 899, 938, 1119  
*American Voter, The* 175  
analytical narratives 132–3  
Angrist, Joshua D.  
    causality 792  
    DID 823, 828–9, 832  
    instrumental variables (IV) 758, 760  
    TSTC 622  
anomalies 20  
    filibuster 233–4  
    veto 231–3  
anonymity 1044–5  
Anselin, L. 732–3  
application programming interfaces (APIs) 388–93  
arguments, unsatisfying 51–3  
Aronow, P.M. 773, 791, 1003  
artificial intelligence (AI) *see* deep learning  
Ashworth, S. 131, 199  
Asmussen, Nicole 921, 927  
assignment mechanism 825–7  
assumptions, inconsistent 50–1  
attitudes 1045–6  
attitudes, measuring 371–3  
    conclusion 381–2  
    multilevel regression with post-stratification (MrP) 373–9  
    uses in literature 379–81  
attrition 791–4  
audiences, bargaining before 228–9  
    does it make a difference? 239–40  
autoregressive conditional heteroskedasticity (ARCH) 613  
autoregressive distributed lag (ADL) 633–4  
autoregressive, integrated, moving average (ARMA/ARIMA) 606, 647  
    model 634  
Bafumi, Joseph 916–17, 922, 926  
bagging 1083–4  
Bailey, Michael A. 920–1, 927  
bandwidth selection 839–1, 849  
Banks, Jeffrey S. 197, 249, 269–70  
Barberá, Pablo 410, 903, 927  
Bareinboim, Elias 1019, 1021–2  
bargaining 195–6, 292–3  
    during war 270–1  
Baron, David P. 249, 292  
Baron-Ferejohn (BF) model 245–6  
    critiques 246–8  
Bassi, Anna 249, 987  
Bateman, David A. 920–1  
Battaglini, M. 131–2  
*Bayesian Analysis for the Social Sciences* 901–2  
*Bayesian Data Analysis* 902  
Bayesian ideal point estimation  
    agenda 924–6  
    comparability 918–2  
    conclusion 931  
    estimation 913–17  
    hierarchical estimation 926–7  
    ideal point and text data 928–31  
    implementation 922–4  
    introduction 910–11  
    nominate 917–18  
    other data 927–28  
    spacial model of voting and RUM 911–13  
Bayesian information criterion (BIC) 603, 943  
Bayesian methods  
    Bayesian inference explained 896–8  
    introduction 895–6  
    major developments 902–5  
    new developments and discussions 905–7  
    short history 898–902  
*Bayesian Methods: A Social and Behavioral Sciences Approach* 901  
Bayesian model averaging (BMA) 954  
Bayesian modeling and inference, postmodern 969–70  
    Bayes factor calculations 976–7  
    Bayesian nonparametrics 977–81  
    classical Bayes 962–5

- conclusion 981–2
- Hamiltonian Monte Carlo 972–5
- introduction 961–2
- modern Bayes 965–9
- poster predictive checks 970–2
- Bayesian model selection, comparison, and averaging
  - conclusion 956–7
  - finite mixture models, model averaging, and stacking 952–6
  - how *not* to test competing theories 940–1
  - introduction 937
  - model selection and Bayes factors 941–6
  - motivating example: theories of Congress 938–40
  - predictive model evaluation 946–52
- Bayesian Structural Vector Autoregression (BSVAR) 592
- Bayes, Thomas 962
- Beach, Derek 105, 1138
- Beck, Nathaniel 618, 629, 722
- behavioral economics
  - accepted procedures 161–3
  - disciplinary cultural differences 167–70
  - research issues and questions 165–7
  - starting assumptions 163–5
- Beim, Deborah 283–5
- Bengio, Y. 1063, 1066
- Bennett, Andrew 105, 117, 1133, 1135, 1137, 1143
- Benoit, Kenneth 256
  - political positions from texts 505, 509, 511, 513
  - text as data 466, 477, 490–1
- Berry, Frances Stokes 314, 323
- Berry, William P. 314, 323
- Bevir, M. 1214, 1216
- bias 164, 827–8
  - see also* selection bias
- bias, algorithmic 99–100
- big relational data (BRD)
  - community detection 560–3
  - concluding remarks 571–3
  - influence and propagation 563–6
  - introduction 552–3
  - network topography 557–60
  - statistical dependency 566–8
  - storing and describing large networks 553–7
  - thinning large networks 557
  - Twitter network analysis of Kavanaugh case 568–71
- bilateral bargaining 227–8
- Billig, Michael 1204, 1205
- Bivariate Probit Model 703–4
- Blackwell, David 978–9
- blame game 236
- Blei, David M. 930–1
- Bliesemann de Guevara, Berit 1219, 1223–4, 1225
- blotto games 273
- Boehmke, Frederick J. 314, 316, 318
- Böhmelet, Tobias 707, 709
- Bolukbasi, Tolga 100, 478
- Bowers, Jake M. 455, 788, 794, 800, 810, 986, 1003
- Box, G.E. 606, 634
- bracketing 828–9
- Brader, T. 989–90
- Brandenberger, Laurence 882
- Brandt, P.T. 592, 606, 609, 612
- Breusch-Godfrey (BG) 603
- bridge sampling 943–5
- Brummer, Klaus 1161
- BSON 85
- Bueno de Mesquita, Bruce 21, 131, 199, 267, 279
- Bureaucratic Politics Model 1152
- Buttice, Matthew 377–378
- Callander, Steven 279, 314
- Calonico, S. 840, 842–3
- Camerer, C.F. 165, 168, 990
- Cameron, Charles M. 228, 229, 231, 278, 283, 284, 285
- Campbell, Donald T. 835–6, 1036
- Card, David 826, 1029
- Carroll, Royce 917–18
- Carrubba, Clifford J. 271, 283, 285
- Casella, George 967, 979
- case selection 114–16
- case study methods
  - case definition 1135–38
  - case-selection techniques 1138–39
  - comparative case studies 1141–2
  - conclusion 1144–5
  - introduction 1133–4
  - options 1142–4
  - post-positivism and large-N approaches 1134–5
  - single case studies 1139–1
- Cattaneo, Matias D. 836, 838–9, 841, 843, 845–50, 855
- causal effects 769–70
- causal heterogeneity 580
- causal inference 121–2, 296–7, 593, 1038
  - confounding 1038–40
  - contamination 1040–1
  - EITM research designs 128–35
  - practical advice 137–8
  - simple theory, non-obvious implications 126–8
  - structural estimation 135–7
  - from theoretical model to implications 123–6
  - what is EITM 122–3
- causality
  - conclusion 798–800
  - estimation 774–80
  - hypothesis testing 781–91
  - introduction 769
  - partially controlled research design 791–4
  - and research design 769–74
  - sensitivity to assumptions 796–8
  - uncontrolled research designs 795
- censoring, left or interval 676
- central bank independence (CBI) 586
- Chamberlain, G. 628–9
- change 184
- Chappell, Henry W. 150–1

- Chib, Siddhartha 905, 944, 976–7  
 Citizens Talk about Europe (CITAE) 1197  
 Claassen, Christopher 381, 951  
 Clark, Tom S. 279, 283–4, 625, 930  
 classification and clustering 522–3  
   applications of supervised methods 529–30  
   applications of unsupervised methods 530–1  
   complications and extensions 531–2  
   design and procedure 525–6  
   evolution of methods 523–4  
   supervised methods 526–7  
   text classification 523  
   unsupervised methods 527–9  
 classification and regression tree (CART) 1082–3  
 Cleveland, William S. 438–9  
 Clinton, Joshua 902, 910, 917–8, 924–5  
 coalition politics 244–5  
   beyond non-cooperative game theory 250–6  
   conclusion: moving forward 256–8  
   new institutionalism 245–8  
   non-cooperative models 248–50  
 coalitions 292–3  
 Coarsened Exact Matching (CEM) 812  
 coding 342–3  
 coffee, demand for (illustration) 759–62  
 Cohn, Carol 1212–13  
 collective action 186  
 Collier, David 105, 332, 339, 1118, 1143  
*Communist Manifesto, The* 13  
 community detection 560–563  
 comparability 517–518  
 Comparative Policy Agendas Project 470  
*Comparative Political Studies* 425  
 Comparative Research on the Event of Nations (CREON) 1153–4  
 comparative statics 133–4  
 complexity 289–90  
   conclusion 304–5  
   machine learning and estimation 295–301  
   representation 301–4  
   theory 290–5  
 compiler average causal effects (CACEs) 748–9, 754–5, 757  
 conceptualization 331–2  
   attributes and structure 334–5  
   conclusion 347–8  
   evaluation 336–7  
   term, sense and reference 332–3  
 concessions, policy 229–30  
 confounding 1038–40  
 consistency 776–8  
 contemporaneous error 618–19  
 contexts 585, 587  
 control 1012  
 Cook, Scott J. 628, 744  
 copula 672–3  
 correspondence analysis 477  
 counterfactual scenarios 136–7  
 covariate balance 848–9  
 Cowes Commission 146  
 Cox model 664–5  
*CQ Weekly* 236  
 Cranmer, Skyler J. 725–6, 882  
 credibility 583  
 Cross-Section Time-Series (CSTS) 616–17  
 CSV (Comma-Separated Values) 85–6  
 cumulative impulse response functions (CIRFs) 607–8  
 Curini, Luigi 220  
 cutoff values 849  
 Dafoe, Allan 1041–2  
 Dahl, Robert A. 333–4, 336, 342  
 Darmofal, David 566, 679  
 data access 543–4  
 Data Access and Research Transparency (DA-RT) 343  
 data collection 112–13, 1201–2  
 data, experimental 1029–32  
 data generating processes (DGPs) 599  
 data reduction 1205–6  
 data science projects  
   concluding remarks 100  
   data 81–2  
   data collection and jingle-jangle fallacies 87–91  
   data in files and streams 84–6  
   data in memory 82–4  
   data preparation 93–7  
   data stores and databases 86–7  
   data wrangling 91–3  
   responsible projects 97–100  
   toolkit 80–1  
   workflow 79–80  
 data sources  
   qualitative 339–41  
   quantitative 341–2  
 Davidson–Harel algorithm 559–60  
 deception 168–9  
*Deciding to Decide* 282  
 decision-making 164–5  
 decisions, judicial 285–6  
 deep learning  
   AI 1054–7  
   conclusion 1070–2  
   image data 1058–61  
   introduction 1053  
   multinodal data 1068–9  
   recent developments 1069–70  
   text data 1061–8  
   understanding deep 1057–8  
   understanding learning 1057  
 deep parameters 136  
 de Finetti, Bruno 895–6  
 DellaVigna, Stefano 1018, 1024, 1029–30  
 de Marchi, Scott 255–7, 293  
 democracy, measuring 903

- Denny, Matthew J. 302, 484, 487  
 design 1014–1015, 1028  
*Designing Social inquiry* 1119  
 Desmarais, Bruce A. 321, 566, 725–6, 882  
 Desposato, Scott W. 927, 999  
 dictionaries 535–7  
 Diermeier Daniel 183, 249–50, 530  
 difference-in-difference-in-difference (DIDID or 3D)  
   581  
 Differences-in-Differences (DID) ) 581–2  
   conclusion 831–3  
   estimation 829–31  
   history 824  
   introduction 822–3  
   issues 824–29  
   notation and method 823–4  
 diffusion networks 318–21  
 discussion 1200–1  
 distributional semantic models (DSM) 477–8  
 diversity, internal 184–5  
 Documenting the Now 98  
 Downs, Anthony 206, 208, 209, 220, 498  
 Draper, Daniel 688, 690  
 drivers, domestic political 1159–60  
 Druckman, James N. 993, 1041  
 Duchesne, Sophie 1199, 1202  
 Dunning, Thad 752, 825, 1020  
 duration analysis  
   conclusion 678–9  
   introduction 659–60  
   parallel durations 667–73  
   serial durations 673–8  
   single duration 660–7  
 Durkheim, Emile 174, 177  
 dyadic data analysis  
   conclusion 726–7  
   directed vs. undirected dyads 718–9  
   extensions and alternatives 725–6  
   introduction 717–8  
   multilevel analysis 720–1  
   rare events 719–20  
   spacial dependence 723–5  
   unobserved dyadic heterogeneity 721–3  
 dyadic event history analysis 317–18  
 dynamic causality 580  
 dynamic models  
   robust 362–3  
   standard 362  
 dynamics 361–3  
 economic modeling 577–8, 586–98  
   (R)DD 582  
   DID 581–2  
   Franzese 580–1  
   NHR 584–5  
   randomized controlled trial (RCT) 578–80  
 Egerod, Benjamin C.K. 509  
 election models (electoral) 293–5  
 Elman, C. 1133, 1143  
 emotions 992–3  
   *see also* sentiment analysis  
 Empirical Implications of Theoretical Models (EITM)  
   142–4  
   QRE 144–5  
   RBC 146–51  
   *see also* causal inference  
 Enns, Peter 379–80  
 ensemble Bayesian model averaging (EBMA) 953–4  
 equations, dynamic systems  
   dynamics 632–4  
   introduction 632  
   multiple equations 634–5  
   reduced form VAR 635–9  
   state-space approach 647–56  
   SVAR 639–41  
   VEC models 641–6  
 equilibrium point predictions 131–2  
 error, contemporaneous 618–19  
 errors 164  
 error, systematic measurement 518–20  
 Escobar, Michael D. 978–9  
 estimands, restrictive 1011–12  
 estimation 606, 689–91, 913–17, 1017–8  
   within known window 845  
   RD 841–3  
   small-group preference 1085–6  
 ethics 1183–6  
 ETL (Extract, Transform, Load) 79–80  
 EU enlargement 710–13  
 event history analysis (EHA) 312, 314–16  
 events, repeated 677–8  
 evidence-driven models (EDMs) 61–3  
   agent-based models (ABMs) 60–1  
   counterfactual analysis 72  
   data construction 67–8  
   malnutrition 63–4  
   refinement and validation 71–72  
   specification 68–71  
   step-by-step guide 64–5  
   theoretical grounding 65–7  
   uptake 72–4  
 experimental designs and methods  
   conclusion 995  
   introduction 985  
   new developments 993–5  
   new developments: topics 990–3  
   political science 985–9  
   types of laboratory experiments in political  
     science 989–90  
 explanations, alternative 113–14  
 exploratory data analysis (EDA) 449–54  
 exponential random graph model (ERGM) 877,  
   879–81, 883–8  
 expressed agenda model 476

- falsification analysis, RD 850–1
- Famine Early Warning Systems Network (FEWS NET) 63
- Fariss, Christopher J. 360, 363
- fast and greedy algorithm 561
- Fearon, James D. 46, 50, 52, 264, 266–8, 272, 719
- feasibility 33–4
- Feasible Generalized Least Squares (FGLS) 618
- features, altering (data) 95–7
- Fey, Mark 197–8, 269–70
- field experiments
  - causal mechanisms 1004
  - conclusion 1004–5
  - crossover and stepped-wedge designs 1001–2
  - experimental designs 1000
  - introduction 999–1000
  - multi-armed bandit and adaptive designs 1002
  - selective trials 1003–4
  - spillover 1002–3
- field experiments, theory and external validity
  - addressing external validity 1018–20
  - concerns 1011–13
  - conclusion 1033
  - external validity 1010–11
  - illustration of parametric structural model 1022–32
  - introduction 1007–8
  - place of theory 1013–18
  - running example 1008
  - theory 1008–10
- fields, specifying 108–12
- fieldwork 1219–22
- findings 17–19
- Fisher, A.R. 769, 781–2
- Fisherian framework 845–7
- Fisher, Ronald A. 963, 1016
- fit 691–2
- fixed-effects (FE) 581, 623–4
- focus groups
  - analysing 1201–6
  - conclusion 1206
  - designing and conducting 1195–1201
  - introduction 1190–1
  - political scientists 1191–5
- forecast error variance decompositions (FEVD) 609
- forecasting 577–8, 584, 593
  - see also* prediction
- Foreign Policy Analysis (FPA)
  - comparing leaders and domestic institutions 1154–61
  - conclusion 1161–3
  - introduction 1148–49
  - landmark scholarship 1153–4
  - multi-faceted levels 1149–52
- forests, random 1083–4
- formalization 45–8
- frailty 666–7, 670
- Franzese, Robert J. Jr 323, 586, 588–9
- Freedman, D.A. 758–9
- Freeman, John R. 18, 147, 150, 609, 612, 905
- frenetic failed legislation 235–6
- Fruchterman-Reingold layout 559
- Fujii, Anne Lee 1212, 1219–21, 1224
- Fukumoto, Kentaro 585, 678
- full-information maximum-likelihood (FIML) 583, 689
- funding (money) 1182–3
- Gailmard, S. 132–3
- games, timing 272–3
- game theory, applied 192–4
  - breadth and limits of an intuition 196–7
  - conclusion 203
  - consistency of explanations 197–9
  - estimation 201–2
  - language 194–5
  - modeling 199–200
  - quantifying models 200–1
  - theory development 195–6
- Gamson's Law 246, 249, 254
- Gamson, William 1198–9, 1204
- gaps, filling of 31–2
- Geddes, Barbara 16, 818
- Gelfand, Alan E. 950, 981
- Gelman, Andrew 438, 975, 1085–86
  - Bayesian methods 896, 902, 904
  - Bayesian models 947, 949
  - measuring attitudes 372, 380
- General Inquirer* 471
- George, Alexander 105, 117, 1135, 1137, 1143
- Gerber, Angela S. 778, 1001
- Gerring, J. 1120, 1123, 1134, 1135, 1139, 1144
- Gerrish, Sean 930–1
- Ghironi, Fabio 147, 149
- Ghitza, Yair 372, 380, 1085–6
- Gibbons, S. 736, 737
- Gilardi, Fabrizio 315, 318
- Gilligan, M.J. 706, 994
- Gill, Jeff 666, 901, 963, 965, 967–8, 970, 978–80
- Girvan-Newman algorithm 561
- Gneezy, U. 168–9
- goals, common 184–5
- Goertz, Gary 116, 1111, 1120
- Golder, Matt 253, 257
- Golder, Scott A. 406–7
- gradient tree-boosting 1084–5
- Graduate Institute of International and Development Studies 63
- Granger causality tests 609
- graphical user interfaces (GUIs) 73–4
- graphs, use of 438–41
  - parallel coordinate plots 446–7
  - small multiple designs 447–9
  - table lens plots 441–6
- Green, D.P. 722, 752, 778
- Grimmer, Justin 301–2, 476, 488, 523, 529, 906–7, 930–1

- Groseclose, Tim 225, 228, 229, 231, 237  
*Guide to Professional Ethics in Political Science* 1179  
 Guttman, Louis 357–8
- Hafner-Burton, Emilie 708–9, 859  
 Hainmueller, Jens 761, 1040  
 Hamilton Monte Carlo (HMC) 972–5  
*Hamilton's Paradox: The Promise and Peril of Fiscal Federalism* 183  
 Hansen, Ben B. 794, 795, 810  
 Harrod-Balassa-Samuelson (HBS) 147  
 Hausman, Jerry A. 625, 627  
 Hays, Jude C. 323, 589, 673  
 hazard model 663–4  
 HDF5 (Hierarchical Data Format) 87  
 HDFS (Hadoop File System) 87  
 Heckman (1976) Selection Model 702–3, 712  
 Heckman, James J. 707, 817, 829, 1013  
 Hennink, Monique M. 1200, 1203  
 Herzog, Alexander 511, 930  
 heteroskedasticity 618–19  
 hierarchical linear model (HLM) 684–6  
   extending levels 686–7  
   longitudinal data 687–8  
   precursors 682–4  
 hierarchical models 605–6  
 Highton, Benjamin 377–8  
*Hill, The* 232  
 Ho, Daniel E. 297, 709, 817  
 Hoff, Peter D. 904, 927  
 Hopkins, Daniel J. 526, 530, 531, 538, 542, 806  
 Hotelling, Harold 206, 498  
 Houser, Daniel 147, 150  
 human rights levels, measuring 903  
 Humphreys, Macarten 1009, 1128  
 Huntington, Samuel 9, 17  
 hypothesis testing 691, 781  
   Fisherian 782–8  
   Neymanian 788–91  
 Hyttinen, A. 843, 845
- ideal points, measuring 902–3  
 IDE (integrated Development Environment) 81  
 identification 359–61, 688–9  
 Imai, Kosuke 813, 815, 817, 906, 926, 1004–5  
 Imbens, Guido W. 813, 847  
*iMostly Harmless Econometrics* 616  
 impulse response functions (IRFs) 607–8  
 independence 186–7  
   conditional 517  
   parametric 673  
   stochastic 670–1  
 Independent Component Analysis (ICA) 95  
 independent durations 668–70  
   serial 676–8  
 indicators 337–9  
   coding 342–3  
   data collection 339–43  
   evaluation 343–4  
   measurement scales 339
- indices  
   combining data 344–6  
   evaluation 346–7
- inference 606–11, 813, 1013–14, 1028–9  
   design-based *see* causality  
   RD 841–3  
   visual 454–7  
   within known window 845
- inferential strategies 1013  
 Ingelgom, Virginie Van 1205–6  
 Ingram, Matthew C. 1128  
 Institutional Analysis and Development (IAD) 178  
 institutionalism 187–8  
   *see also* institutions
- institutions 173–4  
   centrality in political science 174–6  
   collective action 186  
   common goals and internal diversity 184–5  
   contradictions and paradoxes 183–4  
   discursive approach 181–2  
   empirical approach 181  
   historical approach 180–181  
   independence and shared spaces 186–7  
   individuals 185  
   institutionalization 187  
   major approaches 176–7  
   methodological implications 182–3  
   multiple approaches 187–8  
   normative approach 177  
   as organizations 186  
   rational choice approach 178–80  
   stability and change 184  
   structures and individual agency 185–6
- instrumental variables (IV)  
   analysis in design-based causal inference 752–6  
   conclusion 763  
   illustration, demand for coffee 759–62  
   introduction 748–9  
   SEM 756–9  
   in structural equation models 749–52
- Integrated Conflict Early Warning System (ICEWS) 415  
 Integrative Complexity 1151–2  
*International Organization* 425  
*International Studies Quarterly* 425  
 interpretation 606–11, 692, 1203–5  
 interpretation, valid *see* survey experiments  
 interpretive approaches  
   conclusion 1225–6  
   interpretivist explanation 1215–19  
   introduction 1211–12  
   research action 1219–25  
   revisiting the binaries 1212–15  
 interpretive policy analysis 1222–4  
 interstate conflict 261–2  
   blotto games 273  
   complete information 262–4

- conclusion 273–4
- games of timing and wars of attrition 272–3
- incomplete information 266–71
- models linked to domestic politics 271–2
- repeated games 264–5
- stochastic games 265–6
- interval truncation 675–6
- interview research
  - challenges 1176–83
  - conclusion 1186
  - ethical considerations 1183–6
  - introduction 1167–8
  - use in political science 1172–6
  - when and how 1168–2
- introductions 9–10
- Introduction to Models in the Social Sciences, An* 14
- intuitions 196–7
- Item-Response Theory (IRT) 343, 359, 506–7, 969
- iterated prisoner's dilemma (IPD) 294–5
  
- Jackman, Simon 690, 899, 901, 904, 917–8, 922
- Jacobs, Alan M. 1009, 1128
- jangle fallacy 88
- Jeffreys, Harold 942, 976
- Jeliazkov, Ivan 976–7
- Jervis, Robert 199, 1150
- jingle fallacy 88
- John Hopkins University 63
- Jo, Jinhee 921, 927
- Journal of Public Administration Research and Theory* 425
- Journal of Experimental Political Science, The* 1180
- Journal of Mixed Methods Research (JMMR)* 1121
- Journal of Policy Analysis and Management* 425
- Journal of Politics, The* 425
- journals 899
- JSON 85
- Judd, Kenneth 291–2
- judiciary
  - American courts and agendas 281–2
  - case space 279–80
  - current and future research in case space 280–1
  - decisions 285–6
  - introduction 277–9
  - law and stare decisis 279
  - other opportunities 281
  - shaping dockets 282–5
- justification 38
  
- Kaarbo, Juliet 1149–50, 1154
- Kahnman, Daniel 164–6
- Kamada-Kawai model 560
- Kastellec, J. 283–4, 381, 440
- Katz, Elihu 1199, 1201
- Katz, Jonathan N. 618, 722
- Keech, William R. 150–1
- Keele, Luke J 607, 622, 825, 829–31
  
- Kendall, Patricia L. 1194, 1201
- Kenkel, Brenton 201–2
- Kenwick, Michael R. 363
- Kim, In Song 813, 815, 928–30
- King, Gary 410, 812
  - classification and clustering 526, 531
  - dyadic data analysis 720, 722
  - sentiments and social media 538, 542
- Klašnja, M. 849, 851
- Klein, Benjamin S. 279, 1213
- k-Means 528
- knowledge, common 271
- Koch, Julianna 379–80
- Kornhauser, Lewis A. 278, 279, 284
- Kosinski, Michael 409, 1071
- Kousser, Thad 372, 380
- Krehbiel, Keith 183, 281
- Krueger, Alan B. 760, 832
- Kuhn, Thomas 7, 20–1
- Kurowska, Xymena 1219, 1221
- Kwan, Mei-Po 425, 429–30, 1128
- Kydland, Finn, E. 146–7
- Kyung, Minjung 978–9
  
- laboratory experiments 986–9
  - types in political science 989–90
- ladder of abstraction 29–30
- Lake, David 114, 1179
- languages 512–13, 544–5
- large graph layout (LGL) 560
- large-scale approach 847
- Lasswell, Harold 470
- Latent Dirichlet Allocation (LDA) 475–6, 522, 528
- latent networks 321–2
- latent semantic analysis (LSA) 477
- latent space model (LSM) 877, 878–9, 883–8
- Lauderdale, Benjamin 511, 926, 930
- Laver, Michael 247–8, 250–3, 255–6, 257, 292–3, 499, 504
- Lax, Jeffrey R. 283, 285
- Lazarsfeld, Paul F. 345, 347
- leaders 1154–5
  - automated content analysis 1155
  - Leadership Trait Analysis (LTA) 1155–9
- learning 56–7
- learning and diffusion
  - diffusion networks 318–21
  - dyadic event history analysis 317–18
  - event history analysis (EHA) 314–16
  - latent networks 321–2
  - moving forward 323–4
  - policy diffusion 311–12
  - pooled event history analysis (PEHA) 316–17
  - spatial economic models 322–3
  - theoretical background 312–14
- learning, supervised 537–42
  - see also* machine learning
- Least Squares Dummy Variable (LSDV) 623

- leave-one-out cross-validation (LOO-CV) 950–1
- Leavitt, Thomas 986
- Lechner, Michael 823, 829
- Leeman, Lucas 372, 379, 380, 714
- Levendusky, M.S. 989–90
- Liebes, Tamar 1199, 1201
- Lijphart, Arend 181, 1137
- limited-information maximum-likelihood (LIML) 583
- Linz, Juan J. 181, 335
- literature 10–14
- literature reviews 38
- Liu, Jun S. 978–9
- Lo, Albert 978–9
- local average treatment effects (LATE) 757
- local independence 361–3
- Local Interpretable Model-agnostic Explanations (LIME) 299
- Longford, Nicholas T. 689, 690
- long run multiplier (LRM) 607
- Lord's Resistance Army (LRA) 65
- Loughran, Tim 472, 536
- Louvaine method 561–2
- Lowe, William 477, 490, 504, 505–6, 509, 511, 513
- Lucas, Christopher 399, 906
- Lucas, Robert 13, 146
- Luengo-Oroz, M. 1070
- Lupia, A. 990, 992
- Lyall, Jason 826
- McCarty, Nolan 225, 227–8, 229, 231, 237, 381, 922
- McDonald, Bill 472, 536
- MacEachern, Steven N. 978–9
- machine learning 295–6
  - document positions 503–4
  - supervised 473–5, 504–6
  - unsupervised 475–7, 506–12
- machine learning, supervised learning models
  - concluding remarks 1091–2
  - introduction 1079–80
  - kernel methods 1086–91
  - tree-based approaches 1082–6
  - types and concepts 1080–2
- MacQueen, James B. 978–9
- Macy, Michael W. 406–7
- Mahalanobis Distance Matching (MDM) 811
- Mahoney, James 105, 116, 180
- malnutrition 63–4
- Manifesto Project 470
- Man, the State, and War* 679
- March, James G. 16, 176, 177
- margins, vote 229–30
- Markov chain Monte Carlo (MCMC) 967–70
- Martin, Andrew D. 902–3, 905, 919–21, 922
- Martin, Gregory 506
- Martin, Lanny W. 245, 254–5, 257
- Marx, Karl 13–14
- matching, DID 829–30
- matching model 706–7
- matching, statistical 805–7
  - algorithms 810–11
  - categorical and continuous treatments 813
  - Coarsened Exact Matching (CEM) 812
  - comparison to regression 813–14
  - constructing and analyzing samples 808–10
  - inference 813
  - Mahalanobis Distance Matching (MDM) 811
  - matching frontier 812
  - match or mismatch? 817–19
  - philosophy 807–8
  - Propensity Score Matching (PSM) 811–12
  - SUTVA 816–17
  - TSCS 814–16
- Matthews, Steven A. 228, 1128
- maximum likelihood estimation (MLE) 606
- Mayne, Stephanie L. 825, 826
- measurement
  - combining data 344–5
  - conclusion 347–8
  - indicators 337–44
  - indices 344–7
  - see also* economic modeling
- measurement models 353–4, 902
  - assumptions 356–63
  - best practice for applied research 365–6
  - conclusion 366–7
  - extensions and future research 363–6
  - process 354–6
- measurement techniques 1042–6
- mechanism design 268–70
- Median Voter Theorem 208
- Meirowitz, Adam 924–5
- Melitz, Marc J. 147, 149
- Merrill, Samuel, III 219
- Merton, Robert K. 30, 1190, 1192, 1194, 1201
- message legislation 236–9
- message votes 228–9
- Metropolis, Nicholas 967, 973
- Mikhaylov, Slava Jankin 470, 505
- military threat model (MTM) 50–1
- Million or So Correlation Coefficients: Three Experiments on the Modifiable Areal Unit Problem, A* 425
- Mill, J.S. 174, 1113
- ministerial autonomy model 250–3
- Minozzi, William 829, 992
- mixed methods designs
  - challenges 1126–8
  - conclusion 1129
  - introduction 1117–8
  - MMR 1118–9
  - rise of MMR 1119–23
  - types 1123–6
- Modelling Early Risk Indicators to Anticipate Malnutrition (MERIAM) 62, 71–2, 74



- models, choice 696–7
- models, count 697
- models, formal 14–15
- as arguments 44–8
  - exposition 56–7
  - game theory 199–200
  - inspiration for 48–54
  - puzzles 54–6
- model specification
- conclusion 743–4
  - introduction 730–2
  - Monte Carlo analysis 739–43
  - specifying models 732–9
- moderation 1200–1
- modifiable area unit problem (MAUP) 425–6, 428–9, 430–2
- Mokken Scaling Analysis (MSA) 358
- Møller, Jørgen 340
- money 1182–3
- MongoDB 87
- Monroe, Burt, L. 95, 397, 461, 483, 509
- Monte Carlo analysis 739–43
- Montgomery, Jacob M. 1085–6
- Moore, Ryan T. 431–2
- Morelli, Massimo 248–50, 257
- Morgan, David L. 1191, 1194
- Morrow, James 265, 273
- Morton, Rebecca B. 990
- Mosley, Layna 1170
- Mosteller, Frederick 303, 524
- Mostly Harmless Econometrics* 622–3
- Müller, Peter 978–9
- multicausality 580, 584
- multilevel analysis 679
- conclusion 697–8
  - cross-classifications and MMMCS 692–3
  - general linear models 693–7
  - hierarchical linear model (HLM) 681–8
  - inference 688–2
  - interpretation 692
  - multilevel data 679–81
  - statistical properties 681
- multilevel regression with post-stratification (MrP) 371–3
- example 373–4
  - methodological limitations and extensions 377–9
  - subnational preference estimates 375–6
  - technical remarks 376–7
  - uses in literature 379–81
- Multiple-Indicator Multiple-Causes (MIMIC) 364–5
- multiple membership and multiple classification (MMMC) 680
- multi-rater/Aldrich-McKelvey scaling 364
- multi-state model 678
- Munzert, Simon 378, 380, 401
- Naïve Bayes model 540, 1067
- classification and clustering 522, 524, 526–7, 529–30
  - text as data 474, 488
- narratives, analytical 132–3
- National Science Foundation (NSF) 143–4
- natural language processing (NLP) 537
- see also* deep learning; dictionaries
- Neal, Radford M. 973, 978–9
- network analysis 991–2
- added value in political science 858–61
  - causal inference 862–3
  - conclusion 872
  - fundamental components 861–2
  - graph theory 863–5
  - introduction 858
  - measuring, sampling, and missingness 866–8
  - practical guide to matrices 865–6
  - useful properties 869–72
  - visualization 868–9
- network modeling
- additional methods 881–3
  - application of ERGM, LSM, and SBM 883–8
  - discussion 888–90
  - ERGM 879–81
  - introduction 876–7
  - LSM 878–9
  - model description 877–8
  - SBM 881
- networks, diffusion 318–21
- networks, large 553–60
- networks, latent 321–2
- Neumayer, Eric 722, 723
- New European Regime for Venture Capital, A* 509
- New Institutionalism 176
- Newman, Benjamin 426–7
- Neyman–Holland–Rubin (NHR) 584–5, 587–9, 591, 753
- Neyman, J 753, 778, 780
- Nickel-bias 620
- Nielsen, Richard A. 706, 1128
- noncompliance 791–4
- Nonnegative Matrix Factorization (NMF) 95
- non-parametric model 660–2
- observational studies 795
- observations
- adding 94–5
  - combining 93–4
- observations, multiple 816
- O'Donnell, Guillermo 114–15
- Old Institutionalism 175
- Olivella, Santiago 1085–6
- Olsen, Johan P. 176, 177
- omnicausality 580
- opacity 294–5, 299–301, 303–4
- Openshaw, Stan 425, 428
- Operational Code Analysis (OCA) 1151
- opinion mining *see* sentiment analysis
- Ordinary Least Squares (OLS) 295–6, 750, 841–2
- organizations 186
- originality 50

- OSEMN (Obtain, Scrub, Explore, Model, iNterpret) 79
- Ostrom, Elinor 178, 186
- outcomes 693–6, 1011  
RD 851–4
- out-of-sample prediction error 584
- Overman, H.G. 736, 737
- Page, Scott E. 273, 291–2
- Palfrey, Thomas 144, 989
- parallel coordinate plots 446–7
- parallelization 568
- Parameswaman, Giri 282, 285
- parameters 294, 298–9, 302–3
- parametric model 662–7
- parametric structural models  
benefits 1027–32  
setting up and estimating 1022–7
- Park, David 375, 378
- Park, Jong Hee 905
- participant recruitment 1198–0
- path weighted local regression (PWR) 566–7
- Patterns of Democracy* 181
- PDF (Portable Document Format) 86
- Pearl, Judea 1009, 1010, 1019, 1021–2
- Pedersen, Rasmus 105, 1138
- Pemstein, Daniel 364, 903
- Pickup, Mark 620–2, 624, 627, 648
- Pischke, Joern-Steffen 622, 758, 823, 828, 829
- pivots 229–30
- plausible general model (PGM) 600, 601–3  
simplified 603–6
- Plümper, Thomas 622, 625–7, 707, 710–12, 722, 723
- Poast, P. 717, 726
- poisson scaling *see* wordfish scaling model
- policy concessions 229–30
- Political Analysis* 440, 899, 1180
- Political Order in Changing Societies* 9
- political positions  
inference about latent positions 499–503  
machine learning 503–12  
what is the road ahead? 516–20  
when assumptions are broken 512–16  
why scale texts? 499
- political psychology  
accepted procedures 161–3  
disciplinary cultural differences 167–70  
research issues and questions 165–7  
starting assumptions 163–5
- political science, experimental 985–9
- Political Science, or The State Theoretically and Practically Considered* 175
- Political Science Research and Methods* 1180
- pooled event history analysis (PEHA) 316–17
- pooled time series cross section (PTSCS) 616
- Poole, Keith T. 902, 910, 917, 921
- Pope, Jeremy C. 924–5
- Population-Based Survey Experiments* 1036
- potential-outcomes framework (POF) 578–81
- Powell, Robert 114, 198, 266–7, 270, 273
- practices, good 20–3
- precision 778–91
- prediction 297–301, 577–8, 584, 591, 593
- preferences, true 500
- preferential trade agreements (PTAs) *see* selection bias
- Prestcott, Edward C. 146–7
- principal-agent (PA) 178–9
- Principal Component Analysis (PCA) 95
- privacy, data 99
- problems, real world 32–3
- programming 80–1
- Proksch, Sven-Oliver 477, 509, 511
- Propensity Score Matching (PSM) 811–12
- proportionality 665–6
- Prospect Theory 164
- Protestant Ethic and the Spirit of Capitalism, The* 177
- protest movements 412–17
- Public Administration Review* 425
- puzzles 31
- qualitative comparative analysis (QCA) 183
- qualitative research projects 104–5  
causal pathways 105–7  
conclusion 116–18  
implications for case selection 114–16  
implications for data collection 112–13  
implications for identifying alternative explanations 113–14  
models as fields 107–8  
specifying the field 108–12
- Quantal Response Equilibrium (QRE) 144–5
- questions 1027–28  
*see also* puzzles
- questions, asking of 7–8  
conclusion 23–4  
good practices 20–3  
question 1: introduction 9–10  
question 2: literature review 10–14  
question 3: theory 14–15  
question 4: research design 15–17  
question 5: findings 17–19  
summary 20
- Quinn, Kevin M. 476, 531, 902–3, 905, 919–21, 922
- Ragin, Charles C. 1097, 1101–3, 1106, 1113–14
- Ramisa, A. 1068–9
- Ramsay, Kristopher W. 46, 197–8, 201–2, 269–70, 763
- random effects analysis of variance (ANOVA) 682–3
- randomized controlled trial (RCT) 578–83
- randomized responses 1043–4
- random utility model (RUM) 911–13
- random walk 562
- Rationalist Explanations for War* 35
- Real Business Cycle (RBC) 146–51
- Reeves, Andrew 431–2
- regression 813–14
- Regression Discontinuity (RD) design 582

- continuity-based framework 838–43
- empirical illustration 849–54
- falsification methods 848–9
- final remarks 854–5
- general setup 836–8
- introduction 835–6
- local randomization framework 843–8
- Reinforcing Sanctioning Regimes in the Financial Services Sector* 509
- Reiter, Dan 271
- reliability 1176–80
- representativeness 579–80
- reproducibility 97–9
- research designs 15–17
  - causality 770–2
  - EITM 128–35
  - illustrative example 772–4
- research projects
  - conclusions 37–40
  - developing research questions 34–5
  - examples 35–7
  - feasibility 33–4
  - research question pillars 30–3
  - research questions 27–30
- research questions 27–30, 34–7
- see also* research projects
- restricted maximum likelihood (REML) 689
- Rethinking Social Inquiry* 1119
- Reuning, Kevin 363
- Review of the Investor Compensation Scheme Directive* 509
- Richman, Jesse 938–0, 945
- Ritter, Nolan 705–6, 707, 713
- Roberts, Margaret E. 440, 476, 489, 531, 906
- Rodden, Jonathan 183, 375, 378, 380
- Rohlfing, Ingo 1112, 1119
- Romer, Thomas 200, 219, 226
- root mean squared error (RMSE) 513
- Rosenbaum, Paul R. 781–2, 786, 794, 796, 811, 1016
- Rosenthal, Howard 200, 226, 902, 910, 917
- Rubin, Donald B. 753, 757, 811, 847
- Rubinstein, Ariel 292, 1009
- Ruffa, Chiara 1140
- sample effects 1018
- sample sizes 688–9
- Sanders, James 475, 1055
- Sandler, Todd 606, 905
- Sartori, Anne E. 704–5
- Sartori, Giovanni 29, 332
- Sartori Selection Model 704–5
- Savage, Leonard J. 895, 899
- scale dependence 824–825
- scaling, multidimensional 560
- Schelling, Thomas C. 32, 47, 192
- Schmidt, Vivien A. 181, 182
- Schneider, Carsten 1111–12
- Schneider, Christina 52–3
- Schumpeter, Joseph A. 333, 336
- scientific discovery 7–8
- Seawright, J. 1120, 1134, 1139
- seemingly unrelated regression (SUR) 632, 671–2
- Sekhon, Jasjeet S. 709, 795, 844
- Selb, Peter 378, 380
- selection 1012–13
- selection bias
  - conclusion 714
  - empirical illustrations 707–14
  - empirical methods 702–7
  - introduction 701–2
- Selznick, Philip 175, 177, 186, 187
- sentiment analysis 534–5
  - approaches 535–42
  - social media 542–5
  - worked example 545–9
- separation-of-powers (SOP) 224–5
- set theoretical methods
  - formulating hypotheses 1100–1
  - introduction 1097–98
  - necessity and sufficiency 1108–12
  - polarity and asymmetry 1105–8
  - qualitative comparative analysis (QCA) 1112–14
  - set calibration 1101–5
  - set membership scores vs. probabilities 1105
  - set operations 1099–1100
  - short background 1098–9
- settings 1011
- Shehata, S. 1216, 1219
- Shepsle, K.A. 176, 250–3, 257
- Shifting Legal Visions: Judicial Change and Human Rights Trials in Latin America* 108
- Shor, Boris 921–2
- Signorino, Curtis S. 145, 629
- Silveira, B.S. 136–7
- Simulating Acute Malnutrition Toolkit (SAMT) 73
- simulation 943–5
- simulations, counterfactual scenarios 136–7
- Single Value Decomposition (SVD) 95
- Skaaning, Svend-Erik 340
- Skocpol, Theda 16, 175, 176, 180, 1112, 1141
- Slantchev, Branislav L. 198, 267
- Slapin, Jonathan 477, 509, 511
- Smith, Alastair 198, 271
- Snow, Jon 824, 826
- Snyder, J.M.J. 134–5
- social media 542–5
  - social media data 404–5
  - advantages 405–7
  - affects on political behavior 411–12
  - conclusion 417
  - limitations 407–409
  - measuring political behavior 409–11
  - protest movements 412–17
- social psychology
  - accepted procedures 161–3

- disciplinary cultural differences 167–70
- research issues and questions 165–7
- starting assumptions 163–5
- software 692, 899–902
- Southern Politics in State and Nation* 426
- spaces, political 500–2
- spaces, shared 186–7
- spatial autoregressive (SAR) 323
- spatial data 424–6
  - addressing MAUP and UGCoP 430–2
  - conclusion 432
  - context in political research 426–8
  - modifiable area unit problem (MAUP) 428–9
  - uncertain geographic context problem (UGCoP) 429–30
- spatial error model (SEM) 734–5, 738, 756–9
  - spatial interdependence
    - conclusion 743–4
    - introduction 730–2
    - Monte Carlo analysis 739–43
    - specifying models 732–9
- spatial lag model (SAR) 734–6, 738
- spatially lagged X model (SLX) 734–6, 739
- spatial temporal autoregressive (STAR) 323
- specificity 38
- Spilker, Gabriele 707, 709
- Spirling, Arthur 302, 466, 484, 487, 905, 1063
- split population 669–70
- SQL (Structured Query Language) 81
- stability 184
- Stable Unit Treatment Value Assumption (SUTVA) 579, 757, 770, 799, 808, 816–17, 844
- stacking 955
- Stam, Allan C. 198, 271
- States and Social Revolutions: A Comparative Analysis of Social Revolutions in Russia, France, and China* 16, 175
- static models 361–2
- statics, comparative 133–4
- Steenbergen, Marco 585
- Steinert-Threlkeld, Zachary C. 406, 408
- Stewart, Brandon M. 301–2, 488, 523, 529, 930–1
- stochastic block model (SBM) 877, 881, 883–8
- stochastic writing processes 502
- Stokes, Donald E. 175, 207
- Stone, D. 1214, 1223
- Stromberg, D. 134–5
- structural estimation 135–7
- structural topic model (STM) 476, 531–2
- structural vector autoregression (SVAR) model 632, 639–41
- Study of War, A* 32
- Support Vector Machines (SVMs) 526–7, 1086–91
- survey experiments
  - causal inference 1038–42
  - conclusion 1047
  - introduction 1036–38
  - as measurement 1042–6
  - summary 1046–7
- table lens plots 441–6
- Take-it-or-Leave-it (TILI) 226–7
- Talking of the Royal Family* 1205
- Taylor, P.J. 425, 428
- testing 577–8, 580, 583–4, 587, 589–90, 592–3, 1016
- text analysis
  - CNNs 1063–6
  - LSTMs 1066–8
- text as data 461–2
  - analysis using appropriate procedure 486–9
  - conclusions and future directions 491–2
  - conversion to common electronic format 479–80
  - converting features into quantitative matrix 484–6
  - defining documents and units of analysis 480–1
  - defining features 481–4
  - defining the corpus 478–9
  - dictionary analysis (texts) 470–2
  - distributional semantic models (DSM) 477–8
  - interpreting and summarising results 489–91
  - latent vs. manifest characteristics 465–7
  - literary analysis (texts) 469
  - machine learning 473–7
  - qualitative text analysis 469–70
  - statistical summaries (texts) 472–3
  - text as text vs. text as data 462–5
  - what it is not 467–8
  - see also* political positions
- theories 14–15
- Thistlethwaite, D.L. 835–6
- three-stage least-squares (3SLS) 583
- time 1182–3
- time frameworks
  - continuous 661–2, 669, 674–5
  - discrete 660–1, 668–9, 673–4
- time series analysis 599–600
  - additional considerations 611–13
  - conclusion 613–14
  - estimation 606
  - inference and interpretation 606–11
  - principled approach to dynamic specification 600–6
- time-series-cross-section (TSCS) analysis
  - conclusion 629
  - dynamics 619–22
  - heterogeneity 622–7
  - heteroskedasticity and contemporaneous error 618–19
  - introduction 616–17
  - limited dependent variable 627–9
  - PTSCS advantages and disadvantages 617–18
- time-series cross-section data *see* matching, statistical
- time-series models 904–5
- Tingley, Dustin 906, 992
- Titunik, Rocío 844, 849, 851
- Tobler, Waldo 428, 552, 721, 733
- tokenization (tokens) 482–3
- tone analysis *see* sentiment analysis
- training, interview 1180–2

- Traummüller, Richard 381  
 treatments 1011  
 treaty design 707–10  
 Treier, Shawn 903, 922, 924–5  
 Troeger, Vera E. 621–2, 624, 625–7, 684  
 Tversky, Amos 164–6  
 Two-Part Model 705–6  
 two-stage least-squares (2SLS) 583
- unbiasedness 775–6  
 uncertain geographic context problem (UGCoP)  
   425–6, 429–32  
 units 1011  
 University of Maryland 63  
 usefulness 584, 593
- validation 949–50  
 validity 1176–80  
 validity, external 579, 580, 582–4, 591, 593  
 validity, internal 579, 584, 591, 593  
 Vanberg, Georg 254–5, 257, 506  
 Vance, Colin 705–6, 707, 712  
 Van Ingelgom, Virginie *see* Ingelgom, Virgnie Van  
 variables 39  
 variation 1019–20  
 Vazquez-Bare, Gonzalo 839  
 vector autoregression (VAR) 602–3, 632, 635–9  
 vector error-correction (VEC) 632, 641–6  
 vector moving average (VMA) 609, 636  
 vector quantization 95  
 Vehtari, Aki 950–1  
 Vennesson, P. 1140, 1143, 1144  
 verification 1176–0  
 veto bargaining 224–5  
   before an audience 239–40  
   classical games 225–9  
   empirical anomalies 229–36  
   model of message legislation 236–9  
 virtue signaling 236–9  
 visualizing data 436–8  
   exploratory data analysis (EDA) 449–54  
   graph use 438–49  
   inference 454–7  
   recent advances 449–57  
 Voeten, Erik 919, 921  
 Volden, Craig 314, 315, 318, 323  
 voting, spatial 205–6  
   conclusion 220  
   consequences 216–19  
   election models 208–11  
   empirical puzzle 211–12  
   non-positional dimension 207–8  
   positional dimensions 206–7  
   possible solution 212–16
- Walker, Jack L. 311, 314  
 walk, random 562  
 Wallace, David L. 303, 524  
 Waltz, Kenneth N. 679, 1148, 1150  
 Wang, Y. 1071  
 Ware, Colin 438  
 Warsaw, Christopher 372, 375, 378, 380  
 Wasserfallen, Fabio 372, 379, 380  
 Watanabe-Akaike information criteria 365, 949  
 Waterman, Richard 965, 968, 970, 978, 980  
 Webb, Clayton 612  
 web data collection 387–8  
   applied example 399–400  
   current challenges 396–9  
   fundamentals 388–96  
   in practice 400–1  
 Weber, Max 174, 177, 1101  
 web scaping 393–6  
 Western, Bruce 683, 904, 905  
 West, Mike 921, 978–9  
 Wickham, Hadley 93, 438  
 Wilkerson, John 473, 530, 531  
 window selection 847, 849  
 Wojcik, Stefan 877, 883, 885  
 Wolford, Scott 267, 271  
 Wong, Cara 431–2  
 word embedding 1063  
 wordfish scaling model 476, 506–12  
 wordscores 504–6  
*World Politics* 425  
 Wright, Philip G. 749–50, 760, 763
- Yamamoto, T. 1004–5  
 Yao, Yuling 945, 955
- Zadeh, Lotfi 1100, 1106  
 Zellner, Arnold 954, 964  
 zero-intelligence model (GGS) 253–5  
 Zirakzadeh, C. 1215, 1217