**Measurement of Constructs-Lecture 8**

Theoretical propositions consist of relationships between abstract constructs. Testing theories (i.e., theoretical propositions) require measuring these constructs accurately, correctly, and in a scientific manner, before the strength of their relationships can be tested. Measurement refers to careful, deliberate observations of the real world and is the essence of empirical research. While some constructs in social science research, such as a person’s age, weight, or a firm’s size, may be easy to measure, other constructs, such as creativity, prejudice, or alienation, may be considerably harder to measure. In this chapter, we will examine the related processes of conceptualization and operationalization for creating measures of such constructs.

**Conceptualization**

Conceptualization is the mental process by which fuzzy and imprecise constructs (concepts) and their constituent components are defined in concrete and precise terms. For instance, we often use the word “prejudice” and the word conjures a certain image in our mind; however, we may struggle if we were asked to define exactly what the term meant. If someone says bad things about other racial groups, is that racial prejudice? If women earn less than men for the same job, is that gender prejudice? If churchgoers believe that non -believers will burn in hell, is that religious prejudice? Are there different kinds of prejudice, and if so, what are they? Are there different levels of prejudice, such as high or low? Answering all of these questions is the key to measuring the prejudice construct correctly. The process of understanding what is included and what is excluded in the concept of prejudice is the conceptualization process.

The conceptualization process is all the more important because of the imprecision, vagueness, and ambiguity of many social science constructs. For instance, is “compassion” the same thing as “empathy” or “sentimentality”? If you have a proposition stating that “compassion is positively related to empathy”, you cannot test that proposition unless you can conceptually separate empathy from compassion and then empirically measure these two very similar constructs correctly. If deeply religious people believe that some members of their society, such as nonbelievers, gays, and abortion doctors, will burn in hell for their sins, and forcefully try to change the “sinners” behaviors to prevent them from going to hell, are they acting in a prejudicial manner or a compassionate manner? Our definition of such constructs is not based on any objective criterion, but rather on a shared (“inter-subjective”) agreement between our mental images (conceptions) of these constructs.

While defining constructs such as prejudice or compassion, we must understand that sometimes, these constructs are not real or can exist independently, but are simply imaginary creations in our mind. For instance, there may be certain tribes in the world who lack prejudice and who cannot even imagine what this concept entails. But in real life, we tend to treat this concept as real. The process of regarding mental constructs as real is called reification , which is central to defining constructs and identifying measurable variables for measuring them.

One important decision in conceptualizing constructs is specifying whether they are unidimensional and multidimensional. Unidimensional constructs are those that are expected to have a single underlying dimension. These constructs can be measured using a single measure or test. Examples include simple constructs such as a person’s weight, wind speed, and probably even complex constructs like self-esteem (if we conceptualize self-esteem as consisting of a single dimension, which of course, may be a unrealistic assumption). Multidimensional constructs consist of two or more underlying dimensions. For instance, if we conceptualize a person’s academic aptitude as consisting of two dimensions – mathematical and verbal ability – then academic aptitude is a multidimensional construct. Each of the underlying dimensions in this case must be measured separately, say, using different tests for mathematical and verbal ability, and the two scores can be combined, possibly in a weighted manner, to create an overall value for the academic aptitude construct.

**Operationalization**

Once a theoretical construct is defined, exactly how do we measure it? Operationalization refers to the process of developing indicators or items for measuring these constructs. For instance, if an unobservable theoretical construct such as socioeconomic status is defined as the level of family income, it can be operationalized using an indicator that asks respondents the question: what is your annual family income? Given the high level of subjectivity and imprecision inherent in social science constructs, we tend to measure most of those constructs (except a few demographic constructs such as age, gender, education, and income) using multiple indicators. This process allows us to examine the closeness amongst these indicators as an assessment of their accuracy (reliability).

Indicators operate at the empirical level, in contrast to constructs, which are conceptualized at the theoretical level. The combination of indicators at the empirical level representing a given construct is called a variable . As noted in a previous chapter, variables may be independent, dependent, mediating, or moderating, depending on how they are employed in a research study. Also each indicator may have several attributes (or levels) and each attribute represent a value . For instance, a “gender” variable may have two attributes: male or female. Likewise, a customer satisfaction scale may be constructed to represent five attributes: “strongly dissatisfied”, “somewhat dissatisfied”, “neutral”, “somewhat satisfied” and “strongly satisfied”. Values of attributes may be quantitative (numeric) or qualitative (non-numeric). Quantitative data can be analyzed using quantitative data analysis techniques, such as regression or structural equation modeling, while qualitative data require qualitative data analysis techniques, such as coding. Note that many variables in social science research are qualitative, even when represented in a quantitative manner. For instance, we can create a customer satisfaction indicator with five attributes: strongly dissatisfied, somewhat dissatisfied, neutral, somewhat satisfied, and strongly satisfied, and assign numbers 1 through 5 respectively for these five attributes, so that we can use sophisticated statistical tools for quantitative data analysis. However, note that the numbers are only labels associated with respondents’ personal evaluation of their own satisfaction, and the underlying variable (satisfaction) is still qualitative even though we represented it in a quantitative manner.

Indicators may be reflective or formative. A reflective indicator is a measure that “reflects” an underlying construct. For example, if religiosity is defined as a construct that measures how religious a person is, then attending religious services may be a reflective indicator of religiosity. A formative indicator is a measure that “forms” or contributes to an underlying construct. Such indicators may represent different dimensions of the construct of interest. For instance, if religiosity is defined as composing of a belief dimension, a devotional dimension, and a ritual dimension, then indicators chosen to measure each of these different dimensions will be considered formative indicators. Unidimensional constructs are measured using reflective indicators (even though multiple reflective indicators may be used for measuring abstruse constructs such as self-esteem), while multidimensional constructs are measured as a formative combination of the multiple dimensions, even though each of the underlying dimensions may be measured using one or more reflective indicators.

**Levels of Measurement**

The first decision to be made in operationalizing a construct is to decide on what is the intended level of measurement. Levels of measurement , also called rating scales , refer to the values that an indicator can take (but says nothing about the indicator itself). For example, male and female (or M and F, or 1 and 2) are two levels of the indicator “gender.” In his seminal article titled “On the theory of scales of measurement” published in Science in 1946, psychologist Stanley Smith Stevens (1946) defined four generic types of rating scales for scientific measurements: nominal, ordinal, interval, and ratio scales. The statistical properties of these scales are shown in Table 6.1.

|  |
| --- |
| **Table 6.1. Statistical properties of rating scales** |
| **Scale** | **Central Tendency** | **Statistics** | **Transformations** |
| Nominal | Mode | Chi-square | One-to-one (equality) |
| Ordinal | Median | Percentile, nonparametric statistics | Monotonic increasing (order) |
| Interval | Arithmetic mean, range, standard deviation | Correlation, regression, analysis of variance | Positive linear (affine) |
| Ratio | Geometric mean, harmonic mean | Coefficient of variation | Positive similarities(multiplicative, logarithmic) |

Note: All higher-order scales can use any of the statistics for lower order scales.

Nominal scales , also called categorical scales, measure categorical data. These scales are used for variables or indicators that have mutually exclusive attributes. Examples include gender (two values: male or female), industry type (manufacturing, financial, agriculture, etc.), and religious affiliation (Christian, Muslim, Jew, etc.). Even if we assign unique numbers to each value, for instance 1 for male and 2 for female, the numbers don’t really mean anything (i.e., 1 is not less than or half of 2) and could have been easily been represented non-numerically, such as M for male and F for female. Nominal scales merely offer names or labels for different attribute values. The appropriate measure of central tendency of a nominal scale is mode, and neither the mean nor the median can be defined. Permissible statistics are chi-square and frequency distribution, and only a one-to-one (equality) transformation is allowed (e.g., 1=Male, 2=Female). 

Ordinal scales are those that measure rank-ordered data, such as the ranking of students in a class as first, second, third, and so forth, based on their grade point average or test scores. However, the actual or relative values of attributes or difference in attribute values cannot be assessed. For instance, ranking of students in class says nothing about the actual GPA or test scores of the students, or how they well performed relative to one another. A classic example in the natural sciences is Moh’s scale of mineral hardness, which characterizes the hardness of various minerals by their ability to scratch other minerals. For instance, diamonds can scratch all other naturally occurring minerals on earth, and hence diamond is the “hardest” mineral. However, the scale does not indicate the actual hardness of these minerals or even provides a relative assessment of their hardness. Ordinal scales can also use attribute labels (anchors) such as “bad”, “medium”, and “good”, or “strongly dissatisfied”, “somewhat dissatisfied”, “neutral”, or “somewhat satisfied”, and “strongly satisfied”. In the latter case, we can say that respondents who are “somewhat satisfied” are less satisfied than those who are “strongly satisfied”, but we cannot quantify their satisfaction levels. The central tendency measure of an ordinal scale can be its median or mode, and means are uninterpretable. Hence, statistical analyses may involve percentiles and non-parametric analysis, but more sophisticated techniques such as correlation, regression, and analysis of variance, are not appropriate. Monotonically increasing transformation (which retains the ranking) is allowed.

Interval scales are those where the values measured are not only rank-ordered, but are also equidistant from adjacent attributes. For example, the temperature scale (in Fahrenheit or Celsius), where the difference between 30 and 40 degree Fahrenheit is the same as that between 80 and 90 degree Fahrenheit. Likewise, if you have a scale that asks respondents’ annual income using the following attributes (ranges): $0 to 10,000, $10,000 to 20,000, $20,000 to 30,000, and so forth, this is also an interval scale, because the mid-point of each range (i.e., $5,000, $15,000, $25,000, etc.) are equidistant from each other. The intelligence quotient (IQ) scale is also an interval scale, because the scale is designed such that the difference between IQ scores 100 and 110 is supposed to be the same as between 110 and 120 (although we do not really know whether that is truly the case). Interval scale allows us to examine “how much more” is one attribute when compared to another, which is not possible with nominal or ordinal scales. Allowed central tendency measures include mean, median, or mode, as are measures of dispersion, such as range and standard deviation. Permissible statistical analyses include all of those allowed for nominal and ordinal scales, plus correlation, regression, analysis of variance, and so on. Allowed scale transformation are positive linear. Note that the satisfaction scale discussed earlier is not strictly an interval scale, because we cannot say whether the difference between “strongly satisfied” and “somewhat satisfied” is the same as that between “neutral” and “somewhat satisfied” or between “somewhat dissatisfied” and “strongly dissatisfied”. However, social science researchers often “pretend” (incorrectly) that these differences are equal so that we can use statistical techniques for analyzing ordinal scaled data.

Ratio scales are those that have all the qualities of nominal, ordinal, and interval scales, and in addition, also have a “true zero” point (where the value zero implies lack or non-availability of the underlying construct). Most measurement in the natural sciences and engineering, such as mass, incline of a plane, and electric charge, employ ratio scales, as are some social science variables such as age, tenure in an organization, and firm size (measured as employee count or gross revenues). For example, a firm of size zero means that it has no employees or revenues. The Kelvin temperature scale is also a ratio scale, in contrast to the Fahrenheit or Celsius scales, because the zero point on this scale (equaling -273.15 degree Celsius) is not an arbitrary value but represents a state where the particles of matter at this temperature have zero kinetic energy. These scales are called “ratio” scales because the ratios of two points on these measures are meaningful and interpretable. For example, a firm of size 10 employees is double that of a firm of size 5, and the same can be said for a firm of 10,000 employees relative to a different firm of 5,000 employees. All measures of central tendencies, including geometric and harmonic means, are allowed for ratio scales, as are ratio measures, such as studentized range or coefficient of variation. All statistical methods are allowed. Sophisticated transformation such as positive similar (e.g., multiplicative or logarithmic) are also allowed.

Based on the four generic types of scales discussed above, we can create specific rating scales for social science research. Common rating scales include binary, Likert, semantic differential, or Guttman scales. Other less common scales are not discussed here.

Binary scales. Binary scales are nominal scales consisting of binary items that assume one of two possible values, such as yes or no, true or false, and so on. For example, a typical binary scale for the “political activism” construct may consist of the six binary items shown in Table 6.2. Each item in this scale is a binary item, and the total number of “yes” indicated by a respondent (a value from 0 to 6) can be used as an overall measure of that person’s political activism. To understand how these items were derived, refer to the “Scaling” section later on in this chapter. Binary scales can also employ other values, such as male or female for gender, full-time or part-time for employment status, and so forth. If an employment status item is modified to allow for more than two possible values (e.g., unemployed, full-time, part-time, and retired), it is no longer binary, but still remains a nominal scaled item.

|  |
| --- |
|  Table 6.2. A six-item binary scale for measuring political activism |
| Have you ever written a letter to a public official |  | Yes |  | No |  |
| Have you ever signed a political petition |  | Yes |  | No |  |
| Have you ever donated money to a political cause |  | Yes |  | No |  |
| Have you ever donated money to a candidate running for public office |  | Yes |  | No |  |
| Have you ever written a political letter to the editor of a newspaper or magazine |  | Yes |  | No |  |
| Have you ever persuaded someone to change his/her voting plans |  | Yes |  | No |  |

Likert scale. Designed by Rensis Likert, this is a very popular rating scale for measuring ordinal data in social science research. This scale includes Likert items that are simply-worded statements to which respondents can indicate their extent of agreement or disagreement on a five or seven-point scale ranging from “strongly disagree” to “strongly agree”. A typical example of a six-item Likert scale for the “employment self-esteem” construct is shown in Table 6.3. Likert scales are summated scales, that is, the overall scale score may be a summation of the attribute values of each item as selected by a respondent.

| Table 6.3. A six-item Likert scale for measuring employment self-esteem |
| --- |
|  | Strongly Disagree | Somewhat Disagree | Neutral | Somewhat Agree | Strongly Agree |
| I feel good about my job | 1 | 2 | 3 | 4 | 5 |
| I get along well with others at work | 1 | 2 | 3 | 4 | 5 |
| I’m proud of my relationship with my supervisor at work | 1 | 2 | 3 | 4 | 5 |
| I can tell that other people at work are glad to have me there | 1 | 2 | 3 | 4 | 5 |
| I can tell that my coworkers respect me | 1 | 2 | 3 | 4 | 5 |
| I feel like I make a useful contribution at work | 1 | 2 | 3 | 4 | 5https://s3-us-west-2.amazonaws.com/courses-images/wp-content/uploads/sites/538/2016/08/25203114/image13.jpg |

Likert items allow for more granularity (more finely tuned response) than binary items, including whether respondents are neutral to the statement. Three or nine values (often called “anchors”) may also be used, but it is important to use an odd number of values to allow for a “neutral” (or “neither agree nor disagree”) anchor. Some studies have used a “forced choice approach” to force respondents to agree or disagree with the LIkert statement by dropping the neutral mid-point and using even number of values and, but this is not a good strategy because some people may indeed be neutral to a given statement and the forced choice approach does not provide them the opportunity to record their neutral stance. A key characteristic of a Likert scale is that even though the statements vary in different items or indicators, the anchors (“strongly disagree” to “strongly agree”) remain the same. Likert scales are ordinal scales because the anchors are not necessarily equidistant, even though sometimes we treat them like interval scales.

|  |
| --- |
| **Table 6.4. A semantic differential scale for measuring attitude toward national health insurance** |
| How would you rate your opinions on national health insurance? |
|  | Very much | Somewhat | Neither | Somewhat | Very much |  |
| Good |  |  |  |  |  | Bad |
| Useful |  |  |  |  |  | Useless |
| Caring |  |  |  |  |  | Uncaring |
| Interesting |  |  |  |  |  | Boring |

Semantic differential scale. This is a composite (multi-item) scale where respondents are asked to indicate their opinions or feelings toward a single statement using different pairs of adjectives framed as polar opposites. For instance, the construct “attitude toward national health insurance” can be measured using four items shown in Table 6.4. As in the Likert scale, the overall scale score may be a summation of individual item scores. Notice that in Likert scales, the statement changes but the anchors remain the same across items. However, in semantic differential scales, the statement remains constant, while the anchors (adjective pairs) change across items. Semantic differential is believed to be an excellent technique for measuring people’s attitude or feelings toward objects, events, or behaviors.

Guttman scale. Designed by Louis Guttman, this composite scale uses a series of items arranged in increasing order of intensity of the construct of interest, from least intense to most intense. As an example, the construct “attitude toward immigrants” can be measured using five items shown in Table 6.5. Each item in the above Guttman scale has a weight (not indicated above) which varies with the intensity of that item, and the weighted combination of each response is used as aggregate measure of an observation.

|  |
| --- |
| **Table 6.5. A five-item Guttman scale for measuring attitude toward immigrants** |
|  How will you rate your opinions on the following statements about immigrants? |
| Do you mind immigrants being citizens of your country | Yes | No |
| Do you mind immigrants living in your own neighborhood | Yes | No |
| Would you mind living next door to an immigrant | Yes | No |
| Would you mind having an immigrant as your close friend | Yes | No |
| Would you mind if someone in your family married an immigrant | Yes | No |

**Scaling**

The previous section discussed how to measure respondents’ responses to predesigned items or indicators belonging to an underlying construct. But how do we create the indicators themselves? The process of creating the indicators is called scaling. More formally, scaling is a branch of measurement that involves the construction of measures by associating qualitative judgments about unobservable constructs with quantitative, measurable metric units. Stevens (1946) said, “Scaling is the assignment of objects to numbers according to a rule.” This process of measuring abstract concepts in concrete terms remains one of the most difficult tasks in empirical social science research.

The outcome of a scaling process is a scale , which is an empirical structure for measuring items or indicators of a given construct. Understand that “scales”, as discussed in this section, are a little different from “rating scales” discussed in the previous section. A rating scale is used to capture the respondents’ reactions to a given item, for instance, such as a nominal scaled item captures a yes/no reaction and an interval scaled item captures a value between “strongly disagree” to “strongly agree.” Attaching a rating scale to a statement or instrument is not scaling. Rather, scaling is the formal process of developing scale items, before rating scales can be attached to those items.

Scales can be unidimensional or multidimensional, based on whether the underlying construct is unidimensional (e.g., weight, wind speed, firm size) or multidimensional (e.g., academic aptitude, intelligence). Unidimensional scale measures constructs along a single scale, ranging from high to low. Note that some of these scales may include multiple items, but all of these items attempt to measure the same underlying dimension. This is particularly the case with many social science constructs such as self-esteem, which are assumed to have a single dimension going from low to high. Multi-dimensional scales, on the other hand, employ different items or tests to measure each dimension of the construct separately, and then combine the scores on each dimension to create an overall measure of the multidimensional construct. For instance, academic aptitude can be measured using two separate tests of students’ mathematical and verbal ability, and then combining these scores to create an overall measure for academic aptitude. Since most scales employed in social science research are unidimensional, we will next three examine approaches for creating unidimensional scales.

Unidimensional scaling methods were developed during the first half of the twentieth century and were named after their creators. The three most popular unidimensional scaling methods are: (1) Thurstone’s equal-appearing scaling, (2) Likert’s summative scaling, and (3) Guttman’s cumulative scaling. The three approaches are similar in many respects, with the key differences being the rating of the scale items by judges and the statistical methods used to select the final items. Each of these methods are discussed next.

Thurstone’s equal-appearing scaling method. Louis Thurstone. one of the earliest and most famous scaling theorists, published a method of equal-appearing intervals in 1925. This method starts with a clear conceptual definition of the construct of interest. Based on this definition, potential scale items are generated to measure this construct. These items are generated by experts who know something about the construct being measured. The initial pool of candidate items (ideally 80 to 100 items) should be worded in a similar manner, for instance, by framing them as statements to which respondents may agree or disagree (and not as questions or other things). Next, a panel of judges is recruited to select specific items from this candidate pool to represent the construct of interest. Judges may include academics trained in the process of instrument construction or a random sample of respondents of interest (i.e., people who are familiar with the phenomenon). The selection process is done by having each judge independently rate each item on a scale from 1 to 11 based on how closely, in their opinion, that item reflects the intended construct (1 represents extremely unfavorable and 11 represents extremely favorable). For each item, compute the median and inter-quartile range (the difference between the 75 th and the 25 th percentile – a measure of dispersion), which are plotted on a histogram, as shown in Figure 6.1. The final scale items are selected as statements that are at equal intervals across a range of medians. This can be done by grouping items with a common median, and then selecting the item with the smallest inter-quartile range within each median group. However, instead of relying entirely on statistical analysis for item selection, a better strategy may be to examine the candidate items at each level and selecting the statement that is the most clear and makes the most sense. The median value of each scale item represents the weight to be used for aggregating the items into a composite scale score representing the construct of interest. We now have a scale which looks like a ruler, with one item or statement at each of the 11 points on the ruler (and weighted as such). Because items appear equally throughout the entire 11-pointrange of the scale, this technique is called an equal-appearing scale.

Thurstone also created two additional methods of building unidimensional scales – the method of successive intervals and the method of paired comparisons – which are both very similar to the method of equal-appearing intervals, except for how judges are asked to rate the data. For instance, the method of paired comparison requires each judge to make a judgment between each pair of statements (rather than rate each statement independently on a 1 to 11 scale). Hence, the name paired comparison method. With a lot of statements, this approach can be enormously time consuming and unwieldy compared to the method of equal-appearing intervals.



Figure 6.1. Histogram for Thurstone scale items

Likert’s summative scaling method . The Likert method, a unidimensional scaling method developed by Murphy and Likert (1938), is quite possibly the most popular of the three scaling approaches described in this chapter. As with Thurstone’s method, the Likert method also starts with a clear definition of the construct of interest, and using a set of experts to generate about 80 to 100 potential scale items. These items are then rated by judges on a 1 to 5 (or 1 to 7) rating scale as follows: 1 for strongly disagree with the concept, 2 for somewhat disagree with the concept, 3 for undecided, 4 for somewhat agree with the concept, and 5 for strongly agree with the concept. Following this rating, specific items can be selected for the final scale can be selected in one of several ways: (1) by computing bivariate correlations between judges rating of each item and the total item (created by summing all individual items for each respondent), and throwing out items with low (e.g., less than 0.60) item-to-total correlations, or (2) by averaging the rating for each item for the top quartile and the bottom quartile of judges, doing a t-test for the difference in means, and selecting items that have high t-values (i.e., those that discriminates best between the top and bottom quartile responses). In the end, researcher’s judgment may be used to obtain a relatively small (say 10 to 15) set of items that have high item-to-total correlations and high discrimination (i.e., high t-values). The

Likert method assumes equal weights for all items, and hence, respondent’s responses to each item can be summed to create a composite score for that respondent. Hence, this method is called a summated scale. Note that any item with reversed meaning from the original direction of the construct must be reverse coded (i.e., 1 becomes a 5, 2 becomes a 4, and so forth) before summating.

Guttman’s cumulative scaling method . Designed by Guttman (1950), the cumulative scaling method is based on Emory Bogardus’ social distance technique, which assumes that people’s willingness to participate in social relations with other people vary in degrees of intensity, and measures that intensity using a list of items arranged from “least intense” to “most intense”. The idea is that people who agree with one item on this list also agree with all previous items. In practice, we seldom find a set of items that matches this cumulative pattern perfectly. A scalogram analysis is used to examine how closely a set of items corresponds to the idea of cumulativeness.

Like previous scaling methods, the Guttman method also starts with a clear definition of the construct of interest, and then using experts to develop a large set of candidate items. A group of judges then rate each candidate item as “yes” if they view the item as being favorable to the construct and “no” if they see the item as unfavorable. Next, a matrix or table is created showing the judges’ responses to all candidate items. This matrix is sorted in decreasing order from judges with more “yes” at the top to those with fewer “yes” at the bottom. Judges with the same number of “yes”, the statements can be sorted from left to right based on most number of agreements to least. The resulting matrix will resemble Table 6.6. Notice that the scale is now almost cumulative when read from left to right (across the items). However, there may be a few exceptions, as shown in Table 6.6, and hence the scale is not entirely cumulative. To determine a set of items that best approximates the cumulativeness property, a data analysis technique called scalogram analysis can be used (or this can be done visually if the number of items is small). The statistical technique also estimates a score for each item that can be used to compute a respondent’s overall score on the entire set of items. Y indicates exceptions that prevents this matrix from being perfectly cumulative.

|  |
| --- |
| **Table 6.6. Sorted rating matrix for a Guttman scale** |
|  Respondent |  Item 12 | Item 5 | Item 3 | Item 22 | Item 8 | Item 7 |
| 29 | Y | Y | Y | Y | Y | Y |
| 7 | Y | Y | Y |  | Y |  |
| 15 | Y | Y | Y | Y |  |  |
| 3 | Y | Y | Y | Y |  |  |
| 32 | Y | Y | Y |  |  |  |
| 4 | Y | Y |  | Y |  |  |
| 5 | Y | Y |  |  |  |  |
| 23 | Y | Y |  |  |  |  |
| 11 | Y |  |  | Y |  |  |

**Indexes**

An index is a composite score derived from aggregating measures of multiple constructs (called components) using a set of rules and formulas. It is different from scales in that scales also aggregate measures, but these measures measure different dimensions or the same dimension of a single construct . A well-known example of an index is the consumer price index (CPI), which is computed every month by the Bureau of Labor Statistics of the U.S. Department of Labor. The CPI is a measure of how much consumers have to pay for goods and services in general, and is divided into eight major categories (food and beverages, housing, apparel, transportation, healthcare, recreation, education and communication, and “other goods and services”), which are further subdivided into more than 200 smaller items. Each month, government employees call all over the country to get the current prices of more than 80,000 items. Using a complicated weighting scheme that takes into account the location and probability of purchase of each item, these prices are combined by analysts, which are then combined into an overall index score using a series of formulas and rules.

Another example of index is socio-economic status (SES), also called the Duncan socioeconomic index (SEI). This index is a combination of three constructs: income, education, and occupation. Income is measured in dollars, education in years or degrees achieved, and occupation is classified into categories or levels by status. These very different measures are combined to create an overall SES index score, using a weighted combination of “occupational education” (percentage of people in that occupation who had one or more year of college education) and “occupational income” (percentage of people in that occupation who earned more than a specific annual income). However, SES index measurement has generated a lot of controversy and disagreement among researchers.

The process of creating an index is similar to that of a scale. First, conceptualize (define) the index and its constituent components. Though this appears simple, there may be a lot of disagreement among judges on what components (constructs) should be included or excluded from an index. For instance, in the SES index, isn’t income correlated with education and occupation, and if so, should we include one component only or all three components? Reviewing the literature, using theories, and/or interviewing experts or key stakeholders may help resolve this issue. Second, operationalize and measure each component. For instance, how will you categorize occupations, particularly since some occupations may have changed with time (e.g., there were no Web developers before the Internet). Third, create a rule or formula for calculating the index score. Again, this process may involve a lot of subjectivity. Lastly, validate the index score using existing or new data.

Though indexes and scales yield a single numerical score or value representing a construct of interest, they are different in many ways. First, indexes often comprise of components that are very different from each other (e.g., income, education, and occupation in the SES index) and are measured in different ways. However, scales typically involve a set of similar items that use the same rating scale (such as a five-point Likert scale). Second, indexes often combine objectively measurable values such as prices or income, while scales are designed to assess subjective or judgmental constructs such as attitude, prejudice, or self-esteem. Some argue that the sophistication of the scaling methodology makes scales different from indexes, while others suggest that indexing methodology can be equally sophisticated. Nevertheless, indexes and scales are both essential tools in social science research.

**Typologies**

Scales and indexes generate ordinal measures of unidimensional constructs. However, researchers sometimes wish to summarize measures of two or more constructs to create a set of categories or types called a typology . Unlike scales or indexes, typologies are multi-dimensional but include only nominal variables. For instance, one can create a political typology of newspapers based on their orientation toward domestic and foreign policy, as expressed in their editorial columns, as shown in Figure 6.2. This typology can be used to categorize newspapers into one of four “ideal types” (A through D), identify the distribution of newspapers across these ideal types, and perhaps even create a classificatory model to classifying newspapers into one of these four ideal types depending on other attributes.



Figure 6.2. A multi-dimensional typology of newspapers

**Summary**

In closing, scale (or index) construction in social science research is a complex process involving several key decisions. Some of these decisions are:

* Should you use a scale, index, or typology?
* How do you plan to analyze the data?
* What is your desired level of measurement (nominal, ordinal, interval, or ratio) or rating scale?
* How many scale attributes should you use (e.g., 1 to 10; 1 to 7; −3 to +3)?
* Should you use an odd or even number of attributes (i.e., do you wish to have neutral or mid-point value)?
* How do you wish to label the scale attributes (especially for semantic differential scales)?
* Finally, what procedure would you use to generate the scale items (e.g., Thurstone, Likert, or Guttman method) or index components?

This chapter examined the process and outcomes of scale development. The next chapter will examine how to evaluate the reliability and validity of the scales developed using the above approaches.