A mind map is a diagram used to visually outline information. A mind map is often created around a single word or text, placed in the center, to which associated ideas, words and concepts are added. Major categories radiate from a central node, and lesser categories are sub-branches of larger branches.[1] Categories can represent words, ideas, tasks, or other items related to a central key word or idea. Mind maps can be drawn by hand, either as "rough notes" during a lecture or meeting, for example, or as higher quality pictures when more time is available. An example of a rough mind map is illustrated. Mind maps are considered to be a type of spider diagram.[2] A similar concept in the 1970s was "idea sun bursting".

**Origins**

Diagrams that visually map information using branching and radial maps trace back centuries. These pictorial methods record knowledge and model systems, and a long history in learning, brainstorming, memory, visual thinking, and problem solving by educators, engineers, psychologists, and others. Some of the earliest examples of such graphical records were developed by Porphyry of Tyros, a noted thinker of the 3rd century, as he graphically visualized the concept categories of Aristotle. Philosopher Ramon Llull (1235–1315) also used such techniques. The semantic network was developed in the late 1950s as a theory to understand human learning and developed further by Allan M. Collins and M. Ross Quillian during the early 1960s.

**Popularization of the term "mind map"**

The term "mind map" was first popularized by British popular psychology author and television personality Tony Buzan when BBC TV ran a series hosted by Buzan called Use Your Head.[4][5] In this show, and companion book series, Buzan enthusiastically promoted his conception of radial tree, diagramming key words in a colorful, radiant, tree-like structure. Buzan says the idea was inspired by Alfred Korzybski's general semantics as popularized in science fiction novels, such as those of Robert A. Heinlein and A.E. van Vogt. Buzan argues that while "traditional" outlines force readers to scan left to right and top to bottom, readers actually tend to scan the entire page in a non-linear fashion. Buzan also uses popular assumptions about the cerebral hemispheres in order to promote the exclusive use of mind mapping over other forms of note making. When compared with the concept map (which was developed by learning experts in the 1970s) the structure of a mind map is a similar radial, but is simplified by having one central key word.

**Mind map guidelines**

Buzan suggests the following guidelines for creating mind maps:

Start in the center with an image of the topic, using at least 3 colors.

Use images, symbols, codes, and dimensions throughout your mind map.

Select key words and print using upper or lower case letters.

Each word/image is best alone and sitting on its own line.

The lines should be connected, starting from the central image. The central lines are thicker, organic and thinner as they radiate out from the centre.

Make the lines the same length as the word/image they support.

Use multiple colors throughout the mind map, for visual stimulation and also to encode or group.

Develop your own personal style of mind mapping.

Use emphasis and show associations in your mind map.

Keep the mind map clear by using radial hierarchy, numerical order or outlines to embrace your branches.

This list is itself more concise than a prose version of the same information and the mind map of these guidelines is itself intended to be more memorable and quicker to scan than either the prose or the list. This is the latest technique used by today's psychologists.

**Uses**

Rough mindmap notes taken during a course session

As with other diagramming tools, mind maps can be used to generate, visualize, structure, and classify ideas, and as an aid to studying[7] and organizing information, solving problems, making decisions, and writing.

Mind maps have many applications in personal, family, educational, and business situations, including notetaking, brainstorming (wherein ideas are inserted into the map radially around the center node, without the implicit prioritization that comes from hierarchy or sequential arrangements, and wherein grouping and organizing is reserved for later stages), summarizing, as a mnemonic technique, or to sort out a complicated idea. Mind maps are also promoted as a way to collaborate in color pen creativity sessions.

Mind maps can be used for:

problem solving

outline/framework design

structure/relationship representations

anonymous collaboration

marriage of words and visuals

individual expression of creativity

condensing material into a concise and memorable format

team building or synergy creating activity

enhancing work morale

In addition to these direct use cases, data retrieved from mind maps can be used to enhance several other applications, for instance expert search systems, search engines and search and tag query recommender.[8] To do so, mind maps can be analysed with classic methods of information retrieval to classify a mind map's author or documents that are linked from within the mind map.[8]

**Differences from other visualizations**

Concept maps - Mind maps differ from concept maps in that mind maps focus on only one word or idea, whereas concept maps connect multiple words or ideas. Also, concept maps typically have text labels on their connecting lines/arms. Mind maps are based on radial hierarchies and tree structures denoting relationships with a central governing concept, whereas concept maps are based on connections between concepts in more diverse patterns. However, either can be part of a larger personal knowledge base system.Modelling graphs - There is no rigorous right or wrong with mind maps, relying on the arbitrariness of mnemonic systems. A UML diagram or a semantic network has structured elements modelling relationships, with lines connecting objects to indicate relationship. This is generally done in black and white with a clear and agreed iconography. Mind maps serve a different purpose: they help with memory and organization. Mind maps are collections of words structured by the mental context of the author with visual mnemonics, and, through the use of colour, icons and visual links, are informal and necessary to the proper functioning of the mind map.

**Research**

Farrand, Hussain, and Hennessy (2002) found that spider diagrams (similar to concept maps) had limited, but significant, impact on memory recall in undergraduate students (a 10% increase over baseline for a 600-word text only) as compared to preferred study methods (a 6% increase over baseline). This improvement was only robust after a week for those in the diagram group and there was a significant decrease in motivation compared to the subjects' preferred methods of note taking. Farrand et al. suggested that learners preferred to use other methods because using a mind map was an unfamiliar technique, and its status as a "memory enhancing" technique engendered reluctance to apply it. Nevertheless the conclusion of the study was "Mind maps provide an effective study technique when applied to written material. However before mind maps are generally adopted as a study technique, consideration has to be given towards ways of improving motivation amongst users."

**The Semantic Differential and Attitude Research**

The Semantic Differential (SD) measures people's reactions to stimulus words and concepts in terms of ratings on bipolar scales defined with contrasting adjectives at each end. An example of an SD scale is:



Usually, the position marked 0 is labeled "neutral," the 1 positions are labeled "slightly," the 2 positions "quite," and the 3 positions "extremely." A scale like this one measures directionality of a reaction (e.g., good versus bad) and also intensity (slight through extreme). Typically, a person is presented with some concept of interest, e.g., Red China, and asked to rate it on a number of such scales. Ratings are combined in various ways to describe and analyze the person's feelings.

A number of basic considerations are involved in SD methodology:

(1) Bipolar adjective scales are a simple, economical means for obtaining data on people's reactions. With adaptations, such scales can be used with adults or children, persons from all walks of life, and persons from any culture.

(2) Ratings on bipolar adjective scales tend to be correlated, and three basic dimensions of response account for most of the co-variation in ratings. The three dimensions, which have been labeled Evaluation, Potency, and Activity (EPA), have been verified and replicated in an impressive variety of studies.

(3) Some adjective scales are almost pure measures of the EPA dimensions; for example, good-bad for Evaluation, powerful-powerless for Potency, and fast-slow for Activity. Using a few pure scales of this sort, one can obtain, with considerable economy, reliable measures of a person's overall response to something. Typically, a concept is rated on several pure scales associated with a single dimension, and the results are averaged to provide a single factor score for each dimension. Measurements of a concept on the EPA dimensions are referred to as the concept's profile.

(4) EPA measurements are appropriate when one is interested in affective responses. The EPA system is notable for being a multi-variate approach to affect measurement. It is also a generalized approach, applicable to any concept or stimulus, and thus it permits comparisons of affective reactions on widely disparate things. EPA ratings have been obtained for hundreds of word concepts, for stories and poems, for social roles and stereotypes, for colors, sounds, shapes, and for individual persons.

(5) The SD has been used as a measure of attitude in a wide variety of projects. Osgood, et al., (1957) report exploratory studies in which the SD was used to assess attitude change as a result of mass media programs (pp. 305-311) and as a result of messages structured in different ways (pp. 240-241). Their chapter on attitude balance or congruity theory (pp. 189-210) [excerpted in Chapter 13 of this volume] also presents significant applications of the SD to attitude measurement. The SD has been used by other investigators to study attitude formation (e.g., Barclay arid Thumin, 1963), attitudes toward organizations (e.g., Rodefeld, 1967), attitudes toward jobs and occupations (e.g., Triandis, 1959; Beardslee and O'Dowd, 1961; Gusfield and Schwartz, 1963), and attitudes toward minorities (e.g., Prothro and Keehn, 1957; Williams, 1964; 1966). The results in these, and many other studies, support the validity of the SD as a technique for attitude measurement. The question of validity, and other issues in assessing attitudes with the SD, will be treated in more detail after a general discussion of SD theory and technique.

The Semantic Differential and Attitude Research

DAVID R. HEISE

This review was facilitated greatly by the "Contemporary Bibliography of Research Related to the Semantic Differential Technique," (Urbana, Ill.: January, 1967; mimeographed), made available by Charles E. Osgood. The work was carried out while the author was a staff member in the Methodology in Sociology program at the University of Wisconsin, a project funded by the Institute of General Medical Sciences of NIH. This chapter was prepared especially for this volume.

The Semantic Differential (SD) measures people's reactions to stimulus words and concepts in terms of ratings on bipolar scales defined with contrasting adjectives at each end. An example of an SD scale is:

Usually, the position marked 0 is labeled "neutral," the 1 positions are labeled "slightly," the 2 positions "quite," and the 3 positions "extremely." A scale like this one measures directionality of a reaction (e.g., good versus bad) and also intensity (slight through extreme). Typically, a person is presented with some concept of interest, e.g., Red China, and asked to rate it on a number of such scales. Ratings are combined in various ways to describe and analyze the person's feelings.

A number of basic considerations are involved in SD methodology:

(1) Bipolar adjective scales are a simple, economical means for obtaining data on people's reactions. With adaptations, such scales can be used with adults or children, persons from all walks of life, and persons from any culture.

(2) Ratings on bipolar adjective scales tend to be correlated, and three basic dimensions of response account for most of the co-variation in ratings. The three dimensions, which have been labeled Evaluation, Potency, and Activity (EPA), have been verified and replicated in an impressive variety of studies.

(3) Some adjective scales are almost pure measures of the EPA dimensions; for example, good-bad for Evaluation, powerful-powerless for Potency, and fast-slow for Activity. Using a few pure scales of this sort, one can obtain, with considerable economy, reliable measures of a person's overall response to something. Typically, a concept is rated on several pure scales associated with a single dimension, and the results are averaged to provide a single factor score for each dimension. Measurements of a concept on the EPA dimensions are referred to as the concept's profile.

(4) EPA measurements are appropriate when one is interested in affective responses. The EPA system is notable for being a multi-

—end p. 235—

variate approach to affect measurement. It is also a generalized approach, applicable to any concept or stimulus, and thus it permits comparisons of affective reactions on widely disparate things. EPA ratings have been obtained for hundreds of word concepts, for stories and poems, for social roles and stereotypes, for colors, sounds, shapes, and for individual persons.

(5) The SD has been used as a measure of attitude in a wide variety of projects. Osgood, et al., (1957) report exploratory studies in which the SD was used to assess attitude change as a result of mass media programs (pp. 305-311) and as a result of messages structured in different ways (pp. 240-241). Their chapter on attitude balance or congruity theory (pp. 189-210) [excerpted in Chapter 13 of this volume] also presents significant applications of the SD to attitude measurement. The SD has been used by other investigators to study attitude formation (e.g., Barclay arid Thumin, 1963), attitudes toward organizations (e.g., Rodefeld, 1967), attitudes toward jobs and occupations (e.g., Triandis, 1959; Beardslee and O'Dowd, 1961; Gusfield and Schwartz, 1963), and attitudes toward minorities (e.g., Prothro and Keehn, 1957; Williams, 1964; 1966). The results in these, and many other studies, support the validity of the SD as a technique for attitude measurement. The question of validity, and other issues in assessing attitudes with the SD, will be treated in more detail after a general discussion of SD theory and technique.

THE EPA STRUCTURE

One of the distinctive features of the SD is its reduction of ratings to three basic dimensions of variation. A number of early studies were conducted to determine the dimensions of bipolar adjective ratings (Osgood, et al., pp. 47-66). Of special importance was the thesaurus study in which 76 adjective pairs were chosen from Roget's Thesaurus to represent a great variety of semantic contrasts and the corresponding bipolar scales were used by one hundred college students to rate 20 different concepts. Correlations between the ratings on different scales were calculated and factored. The EPA structure was clearly evident in the results of this and other early analyses; in the thesaurus study the EPA dimensions accounted for more than two-thirds of the common variance. Some additional dimensions were found in the early studies, and several scales that made distinctions too narrowly descriptive or too highly abstract were found to be unrelated to any of the major dimensions. Yet, for the most part, early work with the SD revealed that ratings on most scales are highly predictable in the three EPA dimensions alone.

The EPA structure holds up with a wide variety of subjects, concepts, and scales. Bopp (reported in Osgood, et al., pp. 223-226) had 40 schizophrenics rate 32 words on a 13 scale form; the usual EPA structure was recognizable. Wright (1958) had 40 concepts rated on a 30 scale SD by a survey sample of 2,000 men and women distributed over the spectrum of socioeconomic status. In this study each concept was rated by a different sample of 50 persons so the mean ratings for different concepts were entirely independent. Wright found four factors in his data, the first three of which clearly were EPA. Heise (1965) had 1,000 concepts rated on eight scales by Navy enlistees; factor analyses of the data based on mean ratings for the 1,000 different words yielded the usual EPA structure. DiVesta (1966) had 100 concepts rated on 27 scales by subjects in grades two through seven (20 subjects for each concept). The usual EPA structure emerged, though there was some tendency for Potency and Activity to merge into a single Dynamism dimension up until the fifth grade. DiVesta also reports another study in which grade school children used 21 scales to rate 100 different concepts (this time with 100 subjects rating each concept) and, combining the data for all grades, the usual EPA structure was found.

Osgood (1962) reports several early studies designed to determine whether the EPA structure is idiosyncratic to English or

—end p. 236—

whether it holds up within other languages and other cultures. G. J. Suci (1960) had illiterate Navajo, Hopi, and Zuni respondents make ratings by pointing; the data obtained revealed Evaluation and Potency factors; Activity did not appear separately, possibly because of the roughness of the data, or perhaps because not enough Activity scales were included. H. Akuto had 100 Japanese subjects rate 90 concepts on 50 scales in Japanese and found that the EPA structure was clearly evident in the factor structure.

More recently, a program of research has been set up to validate the SD in 24 different languages (Osgood, 1964; Jakobovits, 1966). Analyses now have been completed for 15 languages: American, English, Arabic, Cantonese, Dutch, Finnish, Flemish, French, Greek, Hindi, Italian, Japanese, Kannada, Serbo-Croatian, Swedish, and Spanish. In each culture a set of 50 bipolar scales is developed in the native language (rather than by translation) and these are used to rate 100 basic concepts (the concepts are the same for all cultures, having been drawn to be meaningful everywhere and easily translatable). Ratings are made by adolescent males using 20 subjects per concept, and correlations and factor analyses are calculated for the mean ratings on 50 scales over 100 concepts. In these analyses an EPA structure emerges by blind machine analysis in all but two cases, and in these (Hindi and Arabic) the EPA structure can be obtained by appropriate rotation of the factor axes. Of course, the impression of an EPA structure emerging throughout is based on translation of scales back into English, and it could be that translation introduces a cultural bias. To test this possibility, a pan-cultural factor analysis was conducted (Jakobovits) in which the 50 scales from the 15 cultures were entered as variables in one giant factor analysis and correlations were calculated over concepts. In this analysis the first three factors were clearly recognizable as EPA and every culture clearly contributed to the definition of the EPA dimensions. Jakobovits commented, "The fact that each pan-cultural factor is defined by scale loadings of comparable size across all languages proves the true pan-cultural nature of the semantic space as measured by these procedures" (p. 26).

Characterization of the EPA dimensions—Considering the generality of the EPA dimensions and their importance in research using the SD, it is worth considering in more detail the distinctions that are involved. In the following paragraphs the EPA dimensions are characterized in two ways. First, some of the typical adjective contrasts that define each dimension are presented (taken from Jakobovits). Second, a number of concepts which typically are rated near the extremes of each dimension are given (taken from Jenkins, 1960, and from Heise, 1965).

Evaluation is associated with the adjective contrasts: nice-awful, good-bad, sweet-sour, and helpful-unhelpful. Some concepts which lie on the positive (good) side of this dimension are: DOCTOR, FAMILY, GOD, CHURCH, HAPPY, PEACE, SUCCESS, TRUTH, BEAUTY, and MUSIC. Some concepts which lie toward the negative (bad) pole are: ABORTION, DEVIL, DISCORDANT, DIVORCE, FRAUD, HATE, DISEASE, SIN, WAR, ENEMY, and FAILURE.

Some scales which define the Potency dimension are big-little, powerful-powerless, strong-weak, and deep-shallow. Concepts which lie toward the positive (powerful) pole are: WAR, ARMY, BRAVE, COP, MOUNTAIN, ENGINE, BUILDING, DUTY, LAW, STEEL, POWER, and SCIENCE. Concepts which lie toward the negative (powerless) pole are: GIRL, BABY, WIFE, FEATHER, KITTEN, KISS, LOVE, and ART.

Activity scales are fast-slow, alive-dead, noisy-quiet, and young-old. Some concepts high in Activity are: DANGER, ANGER, ATTACK, CITY, ENGINE, FIRE, SWORD, TORNADO, WAR, WIN, CHILD, and PARTY. Among concepts which lie toward the negative pole on the Activity dimension are: CALM, SNAIL, DEATH, EGG, REST, STONE, and SLEEP.

SD space—Sometimes it is convenient to think of the EPA dimensions as forming a three-dimensional space. The SD, or affective, space is illustrated in Figure 1; the origin or center of this space represents neutrality

—end p. 237—

Figure 1. The SD Space.

on all three dimensions. Treating EPA measurements of a stimulus as coordinates allows the stimulus to be positioned as a point in the space, and this point graphically represents the affective response to the stimulus. Some interesting geometric indices can be devised to measure a stimulus' total affectivity and its affective similarity to other stimuli; these indices are presented later.

CONSTRUCTION AND USE OF SD

The following sections discuss how one makes and uses an SD for research purposes, and what kinds of information are provided for analyses. This discussion serves to introduce vocabulary which will be helpful. later on.

The primary question in constructing an SD is what scales should be used. Two basic criteria enter into scale selection; relevance and factorial composition.

Scale Relevance

Subjects find it easier to use scales which relate meaningfully to the concepts being judged and which make distinctions that are familiar (Triandis, 1959). For example, in rating persons, sweet-sour is less relevant, and thus harder to use, than helpful-unhelpful; among laymen, talkative-quiet would be a better scale than manic-depressive. Furthermore (and more important), relevant scales provide more sensitive measurements. More variance is obtained in using relevant scales and the variance of ratings involves less random error (Koltuv, 1962; Mitsos, 1961).

There are two approaches to identifying scales which are relevant for a given class of concepts and a given sample of persons. On the one hand, subjects can be presented with a set of scales and asked to rank them in terms of their meaningfulness in thinking about x, where x is a class of concepts to be rated like People, Newspapers, Organizations, etc. (Mitsos). One then would use the scales ranking highest in meaningfulness for a given population of raters.

A second, more meticulous approach would be to present pairs or triads of concepts from the stimulus concept domain and ask subjects how these concepts differ. One would make up bipolar scales from the distinctions respondents make, omitting any purely denotative distinctions {e.g., blond versus brunette). For example, if subjects frequently drew the distinction of crudeness, an appropriate scale might be crude-gracious. This approach, developed for the study of individuals by Kelly (1955), has been applied successfully in SD studies (e.g., Triandis).

Factorial Composition

The basic goal in an SD study is to get measurements on the EPA dimensions, and since factor analyses show these dimensions to be independent, one seeks measurements that are independent. This means that appropriate scales will measure the dimensions (i.e., scales that have high factor loadings on the EPA dimensions) and will give relatively pure measures of the dimensions (i.e., each scale has a high loading on just one dimension). The only objective way to select factorially pure scales is on the basis of actual factor analyses. Researchers experienced with the SD are aware that intuition is an unreliable guide in selecting factorially pure

—end p. 238—

scales. One can conduct ad hoc factor analyses to learn the factorial composition of new scales, but this is an expensive procedure since studies based on less than 30 concepts and hundreds of subjects are likely to be misleading. The most common procedure is to select scales on the basis of published factor analyses and following are some available reports which indicate the factorial composition of SD scales. The thesaurus study (Osgood et al., pp. 53-61) has been a standard source of factor analytic information on SD scales. Because of the large number of scales considered (76), this is an important source, but the factor loadings should be treated only as rough indicators because of the unusual method of factoring and because only 20 concepts were rated in this study. Wright presents the factorial structure for 30 scales based on data from a survey sample of 2,000 adults rating 40 concepts. DiVesta gives the factor loadings of 27 scales used to rate 100 concepts by a large sample of children. Jakobovits gives the highest loading EPA scales for 15 languages (including English) as derived from the pan-cultural factor analyses.

The published factor analytic studies provide a large fund of scales to draw on and usually one can obtain a subset of scales which are relevant to the concept domain of interest. It should be noted, however, that another problem arises in selecting scales from previous studies—the matter of semantic stability. When applied to a special class of concepts, the words in a scale may take on special meanings and thus the scale is literally a different one than previously studied. For example, the words HOT and COLD are used connotatively in rating many concepts (like PEOPLE) but may be used denotatively in rating physical objects. Since the scale takes on different meanings with different concepts, its factorial composition may be different for the special class of objects. The problem of semantic stability is (along with the problem of relevance) the primary impetus for carrying out special factor analyses for each new content area.

Number of Scales

Assuming that one has a set of relevant scales, each of which loads on one and only one of the EPA factors, the next question is how many scales should be included in the final instrument. More than one scale for each dimension is desirable since this improves the reliability of factor scores. On the other hand, reliability characteristics of SD scales are such that it would rarely be useful to include more than ten scales to measure a dimension, and generally speaking, four scales per dimension can give adequate sensitivity for most purposes.

Contrary to the practice in many published studies, the number of Evaluation scales should not be more than the number of Potency and Activity scales. Evaluation scales always are found to be more reliable than Potency or Activity scales and thus fewer, not more, are needed for a given level of precision.

Equivalent Forms

In research it is often necessary or desirable to do repeated measurement. This introduces the question of equivalent forms. There is evidence that subjects may recall the SD rating they have made previously when the time periods between repeated measurements are short (Miron, 1961). Consequently, such repeated measurements using the same form may not be independent. An example of how this could confound research is given by Coyne and Holzman (1966) who had subjects give SD ratings for their voice at points before and after listening to themselves on a tape recorder. No significant differences were found when the same SD form was used in all ratings, but highly significant changes appeared when subjects used alternate forms of the SD for the different points of time. This experiment suggests that equivalent forms of the SD are necessary in experiments dealing with short range changes in attitudinal reaction.

The primary problem in the development

—end p. 239—

and use of equivalent forms is the large fund of factor analyzed scales that is required; making up two equivalent forms calls for twice as many scales. Given a fund of scales to draw on, one should try to match factor loadings of scales in different forms. Then an experimental design should be used such that some subjects should use Form A at time 1 and Form B at time 2 while other subjects use Form B at time 1 and Form A at time 2.

Format of SD Test Booklets

There are three possible ways of graphically setting up scales and the concepts to be rated:

(1) Concepts can be presented one at a time, with each concept followed by all of the scales on which it is to be rated; typically, the concept is printed at the top of a page and the scales are arrayed below, one after another, and centered on the page.

(2) A concept and one of the scales on which it is to be rated can be presented as a single item with the various concept-scale combinations arrayed randomly one after another. For example, item 1 might be NEGRO followed by the good-bad scale, item 2 RUSSIAN followed by the passive-active scale, item 3 JEW followed by helpful-unhelpful, etc. (It is immaterial whether the stimulus word is placed to the left or right of the scale [Osipow and Grooms, 1962].)

(3) A single scale can be presented along with all of the concepts which are to be rated on it; for example, the good-bad scale could be presented at the top of the page and concepts listed down along the side, each followed by scale marking positions.

Studies show that measurements differ very little in going from one format to another (Osgood, et al., pp. 81-82; Wells and Smith, 1960), although format 3 is least desirable since there is some slight tendency for ratings of one concept to affect ratings on another concept. From the standpoint of data handling, format 1 is preferable since it groups the data for a single concept, facilitating keypunching and statistical analyses.

When format 1 is used, the order of concepts in the test booklet is immaterial since anchoring or order effects are not evident using this format. Sommer (1965) made a determined effort to produce anchor effects and found none: for example, POLITICIAN was rated the same whether preceded by JANITOR, GARBAGE COLLECTOR, FARMER or whether preceded by STATESMAN, SCHOLAR and SCIENTIST. (See also Osgood, et al., pp. 84-85.)

To disguise the nature of an SD test and to prevent subjects from developing response sets which could reduce sensitivity of measurements, it is customary to mix the scales as much as possible. This means alternating Evaluation, Potency and Activity scales rather than presenting them in blocks and alternating directionality so that the scales' good poles, strong poles, or active poles are not always on the same side.

Adverbial quantifiers—To facilitate the rating of intensity, SD scale positions usually are labeled with adverbs like "extremely," "quite," and "slightly." The study by Wells and Smith inquired into whether the adverbs serve any useful function. SD scales with and without adverbial quantifiers were employed with a survey sample of 400 housewives. It was found that the amount of differentiation in SD ratings was substantially greater when adverbial labels were used: no labels led to many more ratings at the end-points of the scales. Furthermore, interviewers reported that the labeled scales were better understood by the respondents and led to greater cooperation in the rating task. Hence, use of adverbial quantifiers is justified.

The metric characteristics of adverbial quantifiers have been investigated in a number of studies (Cliff, 1959; Howe, 1962, 1966a, 1966b). The results indicate that adverbs "extremely," "quite," and "slightly" do define rating positions which are about equidistantly spaced. The results from these studies also suggest some other adverbs which might be used in some SD studies. For example, the adverbs of frequency—"seldom," "often," "always"—might be meaningful in SD studies of roles (i.e., is a LAWYER sometimes power-

—end p. 240—

ful, usually powerful, always powerful; is a MOTHER sometimes nice, usually nice, always nice). However, the relationship between such frequency ratings and intensity ratings using the customary adverbs is not known.

Administration of an SD

SDs are easily administered to groups and, when possible, this is certainly the most efficient way to obtain SD data. However, an SD also can be administered successfully on an individual basis by survey interviewers.

Instructions should routinely contain a statement that the purpose of the SD is to find out how people feel about things and so the respondent should rate the way he feels. He should use his first impressions and not try to figure out the "right answer" or the answer that makes most sense. Instructions also should contain an example in which the concept presented would elicit a unanimous response from the subjects, for example, TORNADO. The concept is rated by the test administrator, who explains while making the ratings what the scale positions mean. It has been suggested (Osgood, et al., pp. 82-84) that subjects should be urged to work quickly; however, Miron found that subjects could be urged to work slowly and thoughtfully and the same results were obtained, mainly because after the first few ratings, subjects worked quickly, regardless of what they were instructed to do.

For many subject populations, one can turn to the literature to check the experiences and procedures of others who have worked with similar groups, for example: college students—Osgood, et al.; children—DiVesta (1966), and Kagan, Hosken and Watson (1961); survey respondents—Wright, and Wells and Smith; factory workers—Triandis; juvenile delinquents—Gordon, et al. (1963); illiterates—Suci (1960).

Test length—Osgood, et al. (p. 8O) suggested that a subject should be allowed about one hour to make 400 SD judgments (for example, to rate 40 concepts on ten scales). Most college students work faster than this and the allowance is generous for even the stragglers in a college student population. On the other hand, this timing estimate is a convenient round figure, and it is perhaps minimal for subjects not accustomed to taking tests. In any case, the patience and endurance of unpaid subjects can rarely be strained beyond 400 judgments, and for non-college subjects (such as survey respondents), the maximum number of ratings undoubtedly is far less—probably more like 50 judgments.

DESCRIPTIVE MEASURES AND PROCEDURES

A typical SD study dealing with a number of concepts, using several scales for each EPA dimension, and employing a sample of respondents, results in thousands of ratings. Various statistics and procedures are available to compress this data to a comprehensible set of measurements.

Factor Scores

The first step in data reduction is to combine ratings on the separate scales into factor scores. This involves first assigning numerical values to the scale positions; for example, -3, -2, -1,0,1,2,3, going from one end of the scale to the other. (To simplify calculations, numerical values should be adjusted for the directionality of scales; for example, the positions numbered 3 through -3 for the scale nice-awful, and -3 through 3 for the scale bad-good.) The responses that were obtained are then coded and a subject's ratings on a concept averaged over all the scales representing a single factor. The product is a single number representing one subject's reaction to one concept on one of the SD dimensions.

Scale weights—Assuming that the factor loadings of the scales for a given dimension are all high and comparable in size, that all the scales load mainly on the one dimension, and that all the scales are of approximate equal relevance so that the rating variances

—end p. 241—

are approximately equal, then it is reasonable to weight the scales equally in calculating the factor scores (i.e., find the simple mean of the ratings). Only if these assumptions are seriously violated, is a more complicated procedure of differential weighting desirable; this could involve weighting each scale by the squared factor loading or the use of multiple regression formulas. Textbooks on factor analysis provide information on the more complicated procedures.

Group means—A frequent second step in data reduction is finding the group means for the factor scores corresponding to different concepts. This simply involves averaging the factor scores over the subjects in the sample. The group means can be viewed as estimates of true factor scores for the particular concept in the particular group or culture—they are the points around which individuals vary. Group means computed from the SD tend to be extremely stable.

Polarization

The factor scores for a concept constitute a complete description of an affective reaction in terms of the EPA dimensions. For some purposes one might not want such detailed information but simply a general measure of the intensity of the affective response independent of its character. This kind of measure, the emotionality of the concept, is given by the polarization measure—the distance between the neutral point or origin of the SD three-dimensional space and the particular concept under consideration. If the neutral point of the scale was assigned a value of zero in the coding process, the factor scores also have their neutral point at zero and polarization is calculated simply by squaring the factor scores, adding, and taking the square root of the sum. That is:

where e, p, and a, are factor-score measurements of a given concept on the three dimensions.

Profile Analyses

The majority of SD studies involve some hypothesis about differences in affective reaction. For example, one might be interested in reactions to NEGROES versus JEWS; the difference in reaction to NEGROES before and after seeing The Birth of a Nation, or the difference in reactions to NEGROES among Southerners and Northerners. Various approaches for analyzing differences in affective response have been developed.

Dimensions treated separately—One approach examines the differences on each EPA dimension separately. That is, one would compare the means for concept a versus concept b, for time 1 versus time 2, or for group x versus group y on each of the three dimensions separately. This approach provides the most detailed results, the statistical procedures for comparing means are well-studied and relatively non-problematic, and it is definitely the preferred procedure in most SD studies. In any case, it should accompany other types of profile analysis.

D scores—There are instances in which one would like to have a measure of the combined differences on all three EPA dimensions—a summary measure of the total difference in affective reactions. D scores have come to serve this purpose in SD research. These represent the distance between two sets of SD measurements when both are plotted as points in the three-dimensional SD space. The formula for calculating D scores is as follows: let el, pl, al, be the factor score for concept 1 (or time 1 or group 1) ; e2, p2, a2, the measurements for concept 2 (or time 2 or group 2). Then

The meaning of D scores can be illustrated by an example. The average EPA factor scores for the concepts HOME, OFFICE, and WORK were drawn from Heise (1965) and entered into the formula for D. It was found that the distance between HOME and WORK is about 3.8 units while the distance between OFFICE and WORK is .8 units. Thus, the affec-

—end p. 342—

tive reaction to WORK is more similar to that for OFFICE than to that for HOME.

Considerations in using D—The reliability of D scores based on group means (where N = 30) is adequate; the correlations between test and retest or between alternate groups are above .90 (Norman, 1959). The random distribution of D under various conditions of rating errors has been studied by Cozens and Jacobs (1961).

Despite the simplicity and the reliability of the measure, D scores should be employed conservatively. D scores completely hide the character of a difference, and a large D could be due to a big difference on one dimension or small differences on all three dimensions. When only the D scores are presented, a reader has no way of determining which is the case.

Beyond that, however, they can be misinterpreted and lead to artifactual findings. For example, at one time a popular project was to show that the difference (D) between evaluative ratings of Ideal Self and Actual Self is greater for neurotics than for normals. To simplify matters, suppose that all persons see their ideal selves as quite good (an assumption which is realistic). Now suppose that neurotics have low evaluation of their actual selves, rating their actual selves as slightly bad, whereas normals rate their actual selves as slightly good. Since both groups have the same rating of the ideal self, it inevitably follows that the neurotics are further from their ideal selves. It could be a serious error to say that "what's wrong" with neurotics is the discrepancy between their actual and ideal selves, since perhaps what is really wrong with them is merely their low evaluation of actual self, which produces the discrepancy as an artifact or inevitable side effect. In fact, Bass and Fiedler (1959) did find that D scores added very little to the basic factor scores in predicting maladjustment. Pitfalls involved in D scores are discussed in greater detail in a series of articles by Cronbach (1955, 1958; Cronbach and Gleser, 1953).

EPA Compounds

There are instances in which the EPA dimensions separately are not nearly as interesting as some compound or combination of them. This is true when there is a particular type of affective reaction that is of special interest. Several techniques are available for analyzing compounds of E, P, and A.

Multiple regression—Suppose one wanted to estimate voting behavior from reactions to the concept VOTING, and it was believed that Evaluation, Potency and Activity reactions all are important. An appropriate procedure to find the total relationship between the EPA structure and the behavioral criterion would be to run a multiple regression of voting behavior simultaneously on all three EPA measurements. That is, one would take the observed voting behavior of a subject and relate it, simultaneously, through multiple regression procedures to his observed reactions to the concept VOTING in terms of Evaluation, Potency and Activity. Since the three dimensions are essentially independent, their estimation power will cumulate, and generally speaking, multiple regressions will give significantly better predictions than simple regressions based on each dimension separately.

Characteristic attributes—Different compounds of E, P, and A seem to have special meanings and psychological significances. That is, the shape of an EPA profile can be a fruitful subject for analysis and may itself provide special information for prediction and explanation. Osgood, et al. (p. 116-124) developed the idea of characteristic attribute. This is a compounded or complex dimension of reaction involving Evaluation, Potency, and Activity all together in some constant proportions. These proportions are derived by rotating the EPA axis and then recalculating the scores of concepts in terms of the new, compounded dimensions. For example, one could define a dimension of reaction which consists of one part Evaluation, one-half part Potency, and one part Activity—this is the characteristic attribute of SUC-

—end p. 243—

CESSFUL. Any other concept whose EPA measurements are in the same ratio could be compared directly with SUCCESSFUL in the sense that the other concept evokes either more or less of the same complex reaction.

Further, they found (pp. 120-124) that the major reaction to political personages and concepts involved such a compound of feelings—about equal amounts of EPA defining roughly a dimension of progress versus decay. A methodological implication is that in political studies it may be more economical and meaningful to measure candidates and issues directly in terms of the characteristic attribute for politics, using scales like healthy-sick or progressive-degenerative (these are hypothetical scales whose actual properties are not known). Theoretically, the special dimensions suggest that psychological processes involved in politics are of a different nature and character than those involved in family life, for instance; there are habitual modes of response for political objects which do not apply in other domains.

Position indices—Sometimes it is of interest to identify all concepts whose profiles are similar both in terms of shape and elevation; that is, concepts that are at approximately the same place in the SD space. One very rough measure of this type of position is specification of the octant of the SD space in which a concept falls. Jakobovits specified the octant for 100 common concepts as rated in different cultures and showed that there are major differences among cultures in the proportions of concepts lying in each octant.

A more refined procedure is to define a point in the SD space in terms of EPA measurements and then identify all concepts that lie within some distance from this point, calculating the distance by the D score formula. For example, Heise (1965; 1966a) defined the point in the space that corresponds to the affective condition for n affiliation by having a variety of affiliative behaviors rated on the SD. The ratings for all behaviors were averaged on each EPA dimension, giving a single motive profile, which in turn, defined a single motive point in the SD space. The distances between this point and a large number of SD responses to words were calculated, and it was shown that persons aroused in n affiliation tend to use words with affectivity similar to that associated with the motive; that is, they use words that lie close to the motive point.

Jenkins, Russell, and Suci (1959) have prepared tables giving the distances between all 360 concepts in their Semantic Atlas Study (1958).

Visual Presentations

It has been suggested that the EPA relations (Osgood, et al., pp. 93-97) among a set of concepts can be represented visually by making up three dimensional models using wires and small rubber balls. To make a model one calculates the distances between concepts using the D score formula, cuts wires proportional to the distances, and builds the model by inserting the wires in soft rubber balls, each of which represents one concept. One then would have a three-dimensional structure in which the relationships between concepts are visually evident and can be viewed from different perspectives.

Arthur (1965) advocated three-dimensional graphing using perspective charts as a means of reducing EPA measurements to a visual form and claimed that this procedure was substantially easier than model building while giving much the same kind of insight.

Reliability of SD Measurements

A study of the absolute deviations between ratings of a concept in test and retest (with retest up to three months later) was reported in Osgood, et al. (p. 127). For evaluation scales it was found that the average difference between ratings on the test and retest was somewhat more than one-half scale units. For Potency and Activity scales the average difference between the test and retest ranged from .7 to 1.0 scale units. The authors concluded from their data that a difference of 3 scale units or more between two ratings on

—end p. 244—

the same scale could be considered statistically significant at the .05 level in a two-tailed test.

DiVesta and Dick (1966) studied the test-retest reliabilities of SD ratings made by grade school children. In their study each subject rated a different concept on a series of scales, and reliabilities were determined by correlating the ratings made on a first test with ratings made one month later on a second test. The correlations for different scales ranged from .27 to .56. DiVesta and Dick found that reliabilities are somewhat higher in the higher grades and also that Evaluation scales tend to be somewhat more reliable at all grade levels.

A reliability study by Norman (1959) gives information on how much shift occurs in ratings, relative to what might be expected if the ratings were purely random. Norman had 30 subjects rate 20 concepts on 20 scales in a test and retest spaced four weeks apart. On the average he found that the amount of shift in ratings was about 50 per cent of what would be expected if the ratings were completely random. More specifically, his results showed that 40 per cent of the scale ratings do not shift at all from test to retest, 35 per cent of the ratings shift by one scale unit, and 25 per cent of the ratings shift two or more scale units. Norman found that ratings are more stable for some concepts than for others, and this seems to be related to the number of meanings for a concept. This may also be a function of how extremely the word is rated. Other studies suggest that concepts whose true values are neutral are rated with less reliability (Peabody, 1962; Luria, 1959). Norman also found that some subjects were more stable than others in making their ratings; in particular, there is a tendency for those who use the end-points of scales more often to have lower test-retest stability. Finally, he found that certain scales are associated with greater stability; in particular, Evaluation scales evoke fewer shifts.

The general impression produced by these test-retest reliability studies is that a person's rating of a single concept on a single scale constitutes a measurement, albeit not an extremely delicate one. Such results may be somewhat misleading, however, because test-retest statistics measure stability as well as reliability. Consequently, low correlations may be due to actual changes in subjects' reactions as well as random errors. In any case, single ratings rarely are used in SD research; instead, factor scores, which should be more reliable because they are the averages of ratings on several scales, are more commonly employed.

Factor score reliability—A study is reported in Osgood, et al. (p. 192) in which several controversial topics were rated on six evaluation scales, and factor scores, representing each subject's evaluative reaction to a given topic, were obtained by summing the ratings on the six scales. The correlations between test and retest factor scores ranged from .87 to .97 with a mean of .91. DiVesta and Dick in their study of SD reliability among children made up factor scores by averaging ratings on two scales for a given dimension and correlating the measurements from the first test with those from the second test given one month later. For children in the fourth grade or higher the correlations ranged between .5 and .8 and were highest for Evaluation factor scores; for students in the third grade or lower, test-retest correlations ranged between .4 and .5. (Reduced reliability of factor scores among younger children also was found by Maltz, 1963.) DiVesta and Dick found that test-retest correlations were somewhat higher when the retest followed the first test immediately. In this case the r's ranged between .6 and .8. Norman examined the effect of making up factor scores from various numbers of scales. His results indicate that factor scores are more reliable than single ratings and that most of the gain in precision is accomplished by averaging just three or four scales; going up to an eight-scale factor score seems to add very little additional stability when looking at data from a test and retest spaced one month apart.

The various studies indicate that there is indeed a significant gain in test-retest corre-

—end p. 245—

lations when factor scores are used rather than individual scale ratings. Furthermore, it appears that most of the possible improvement can be obtained using relatively few scales in making up the factor scores.

Group means—Many SD studies do not focus on an individual's rating of a concept but on a group mean. That is, interest is in the average score in a certain group rather than the score for anyone person. In such case, there is averaging both across scales (factor scores) and across persons, and reliabilities should be even higher.

DiVesta and Dick calculated factor score means for groups of three to five children. The immediate test-retest correlations ranged from .73 to .94, figures that are significantly higher than the correlations based on individual subjects. Norman calculated scale means for 20 concepts using groups of 30 raters. The test-retest correlations between means was .96, and the correlations between means produced by two different samples of student respondents was .94. Miron averaged the factor scores for 20 concepts across 112 subjects and obtained test-retest correlations of .98 or more.

These studies reveal that group means on the EPA dimensions are highly reliable and stable even when the samples of subjects involved in calculating the means are as small as 30.

THE SD IN ATTITUDE RESEARCH

Validity

The general validity of the SD for measuring attitudes is supported by the fact that it yields predicted results when it is used for this purpose and also supported by studies which compare SD measurements with attitude measurements on traditional scales. A series of substantive studies was mentioned in the introduction. Now we turn to studies comparing SD measurements with measurements on traditional scales.

Evaluation factor scores for the concepts Negro, Church, and Capital Punishment were found to parallel Thurstone attitude measurements, and correlations between the two types of measurement were in the range of .74 to .82 (Osgood, et al., p. 230 in this book). Evaluation factor scores for Crop Rotation were found highly related to scores on a Guttman scale for the same concept in a sample of farmers; the rank order correlation was .78 between the two sets of measurements (Osgood, et al., p. 231 in this book).

A study by Nickols and Shaw (1964) supported the validity of SD ratings as attitude measurements, but also found that the relationship between SD measurements and Thurstone measurements varies under certain conditions. Attitudes toward College Professors and toward the Church were assessed using Evaluative factor scores and two Thurstone scales. When the attitude object was non-salient for subjects, the relationship between the measurements was high: r = .71 for non-college subjects rating Professor and r = .76 for college students rating the Church. However, when the attitude object was salient for the subjects, the relationship between the two measures dropped to low values: r = .29 for college students rating Professors and r = .39 for church attenders rating the Church. Nickols and Shaw noted that while the variations in attitude were less in the high saliency groups, this difference was not enough to explain the drops in correlation. They also presented evidence that the Thurstone scales retained their reliability as measurements of individual differences in the high saliency groups. An ad hoc analysis by this author (using the data summarized by Heise, 1965) indicates that in the case of SD ratings, saliency of attitude objects does not affect their reliability either.

Nickols and Shaw hypothesize that subjects are more sensitive to the social repercussions of their ratings when dealing with salient objects, and that the SD is more transparent as a measure of attitude. Thus, social desirability1 may enter as a factor in SD ratings of salient objects. This interpretation re-

1 Social desirability is the tendency of subjects to give what they believe are socially acceptable responses.

—end p. 246—

ceives indirect support from a study by Ford and Meisels (1965) which showed that the social desirability of SD scales corresponded directly to their loading on the Evaluation dimension (the Potency and Activity dimensions are unrelated to social desirability). That being the case, direct SD ratings of objects may not be an efficient approach to attitude measurement when salient or delicate topics are involved. However, before concluding this firmly, one would like to see replications which involve more than two attitude objects. The replications should include: (1) subjects' need for approval as an actual control variable, (2) a third criterion used to show that the SD measurements are the less valid, and (3) SD measurements on all three dimensions.

Tittle and Hill (1967) conducted a study in which various scales designed to measure attitudes toward student political participation were compared with each other and evaluated in terms of their ability to estimate voting behavior [Chapter 30 in this volume]. The SD measurements were obtained by averaging Evaluation factor scores for five concepts (voting, discussing student politics, holding office, assisting in campaigns, and keeping informed). The SD attitude measurement was found to be related to measurements of attitude using other scales. The correlation between the SD measurement and a measurement on a Likert-type scale was .62 (presumably product moment correlation; the type of association measure was not identified by the authors). The SD measure also had some capability for estimating voting behavior. However, the SD measure was less strongly related to behavior than the measurements on other scales, especially the Likert-type scale. The study is an important contribution to the literature on attitude and behavior, but it could have been improved from the standpoint of SD methodology. First, measurements were made only on the Evaluation dimension; past work (Osgood, et al., pp. 120-124) indicates that all three dimensions are important elements in political attitude. Second, the SD scales apparently were all oriented in the same direction (i.e., the positive poles of the scales were all presented on the same side) so that response sets could interfere with measurement precision. Finally, it was merely assumed that attitudes toward a variety of concepts would all be equally relevant for predicting voting behavior; actually, this is an empirical question and attitudes toward the five different concepts should not merely have been summed, but entered in a regression analysis.

In summary, most studies provide confirmation that the SD can be used to measure attitudes. Too little methodological research is available to decide whether SD ratings always provide as sensitive a measure of attitude as is given by traditional scales. However, for the present, one perhaps should be cautious in using the SD with highly salient topics since there is some evidence that measures may be confounded by social desirability effects in such instances.

Dimensionality

Measurement on SD scales has been found to have a three-dimensional structure whereas attitude measurements typically are made on a single dimension. What is the relationship between the EPA structure and an attitude measurement?

Osgood, et al. (p. 228 in this book) proposed that in measuring attitudes, just the Evaluation dimension of the SD need be considered. The justification for this suggestion was simply that it seemed reasonable in light of previous writings on attitudes. At one point (p.233 in this book) they did suggest that considerably more information could be obtained by also measuring Potency and Activity, but they treated Potency and Activity as distinct from attitude. In retrospect, it appears that their identification of the Evaluation dimension with attitude measurement was erroneous, for it seems that the single dimension involved in attitude measurement is only sometimes pure Evaluation, and other times is a compound dimension involving Evaluation, Potency, and Activity.

Taking the extreme empiricist position that attitude is what attitude scales measure,

—end p. 247—

one question that can be asked is, Do traditional attitude scales measure pure Evaluation or do Potency and Activity also get involved? If traditional attitude scales do measure Potency and Activity as well as Evaluation, this should be evident in the content of at least some of the items. The following are a few selected items from scales presented by Shaw and Wright (1967) which suggest that traditional scales in fact do tap Potency and Activity content. The Anti-Semitism Scale. (pp. 384-386) "In order to handle the Jewish problem, Gentiles must meet fire with fire and use the same ruthless tactics with the Jews that the Jews use with the Gentiles" (Potency). "Jewish power and control in money matters is far out of proportion to the number of Jews in the total population" (Potency). Attitude Toward the German People. (pp.396-398) "The German soldiers were, almost without exception, cruel and brutal" (Potency). "Germans are slow and unimaginative" (Activity). Acceptance of Self. (pp. 432-436) "I am quite shy and self-conscious in social situations" (Activity). "I seem to have a real inner strength in handling things. I'm on a pretty solid foundation and it makes me pretty sure of myself" (Potency).

The Bogardus Social Distance Scale can be taken as a measurement of attitudes toward different nationality groups. Rather low and unstable correlations have been reported between Evaluation ratings of nationality groups and measurements on the Bogardus scale—.22 for Germans, .62 for Chinese, and .59 for Hindus (Osgood, et al., pp. 233-234 in this volume). However, multiple regressions of Bogardus scores on all three EPA dimensions resulted in much higher correlations, and the degree of relationship was about the same for all three nationalities—.78, .80 and .72 for Germans, Chinese, and Hindus respectively. Particularly in the case of the Germans, Evaluation accounted only for about 5 per cent of the variance in Bogardus scores while Potency and Activity accounted for 55 per cent.

In a study by Heise (1966b) EPA ratings were obtained from subjects in three status groups for a number of concepts specially selected to represent key areas of attitudinal differences. Most of the significant differences between groups appeared on the Potency and Activity dimensions rather than on Evaluation.

Ratings of occupational prestige can be interpreted as measurements of attitudes toward different status positions. It is noteworthy that prestige ratings correlate highest with ratings on the scale successful-unsuccessful (Gusfield and Schwartz, 1963) which loads jointly on Evaluation, Potency, and Activity.

From the above results it appears that what we mean by attitude, both in theory and in measurement, is simply the affective reaction to an object, and this reaction frequently is along a dimension which is a compound of Evaluation, Potency, and Activity. In the vocabulary of SD research, the single dimension represented in an attitude scale corresponds to the characteristic attribute for the attitude object, and this is only sometimes pure Evaluation. Accordingly, studies employing the SD for attitude measurements should make use of all three dimensions to get measurements paralleling those on traditional attitude scales.

Special Features of SD Attitude Measurement

A generalized method—An SD can be used as a generalized technique in the sense that a subject's attitude toward any object might be assessed by having the subject give ratings on the same set of SD scales. The SD offers the usual advantages of generalized attitude scales (Osgood, et al., pp. 231-233 in this volume).

(1) Economy. The same bipolar scales can be used to measure attitudes toward any object, so the costs of preparing a different scale for every object are eliminated.

(2) Instant Readiness. An SD for measuring attitudes can be made up immediately for crash programs or for topical projects in

—end p. 248—

social research like studies of disasters, riots or the appearance of new political figures.

(3) Cross-concept Comparability. Since attitudes toward various objects are all measured on the same scales, there is the potential for comparing different attitudes.

The major problem in using a single set of SD scales as a generalized attitude scale is the matter of scale relevancy or, more generally, of concept-scale interactions. A single set of scales used for all objects would provide relatively insensitive measurements for some. This may not be objectionable (for example, in exploratory work or in research involving a large number of attitude objects) where a set of scales like those from the pan-cultural factor analyses can be used as a rough and ready instrument for general attitude measurement. Where, however, sensitivity is necessary, it probably is desirable to use the SDs developed for the particular content areas of interest.

Standard metric—One of the unique features of the SD is that attitudes toward a vast array of objects can be measured in terms of basically the same metric on the three EPA dimensions. Thus, all of the objects can be positioned in a single attitudinal space. This feature of the SD has yielded developments and insights that would have been difficult or impossible to obtain using attitude scales in which the metric changes for each object considered.

Congruity theory—Attitude balance theories postulate that when two concepts are associated the attitudes toward the concepts tend to converge, and when two concepts are dissociated (contrasted), the attitudes tend to diverge. While this basic theory has been expressed in a number of ways (see Brown, 1962) a quantitative statement was possible only when attitudes toward different objects could be measured on comparable scales, for only then could attitude convergence or divergence take on a rigorous meaning. The SD provided the required metric and permitted the development of congruity theory, an approach to attitude balance in which predictions about attitude change are made in terms of formulas (Osgood, et al., 1957, chapter 5). Congruity theory is a topic in itself and is not of central interest here. The interested reader should consult the chapter in Osgood, et al. and articles extending and modifying the theory by Fishbein and Hunter (1964), Manis, Gleason, and Davies (1966), and Tannenbaum (1966).2 What is of immediate interest here is that a mathematical approach to attitude balance required a standardized metric like that provided by the SD.

Homogeneity in attitudes—Differences between the average attitudes in various groups are found frequently using both traditional attitude scales and the SD. However, the measurement of many different attitudes on the uniform metric of the SD provides an insight which previously was obscured; attitude differences between groups tend to be smaller than one might expect. This finding develops out of comparing differences between groups with differences between concepts. For example, taking a case where group differences should be extreme—the evaluation of NEGROES by Southern whites and by Negroes—we find a difference of 1.2 evaluation units while the average difference in evaluation between the concepts of FRIEND and ENEMY within the same groups is 3.4 units (Williams, 1966). Similarly, we find that while gang delinquents evaluate FIGHTING more favorably than do middle-class adolescents, the difference between groups on this concept is far less than the differences in attitudes toward the two different concepts of FIGHTING and SCHOLARSHIP within either group. The delinquents do not feel that fighting is really good, it is just somewhat less bad than the middle-class boys view it (Gordon, et al., 1963). When a series of concepts are chosen to tap attitude differences between males and females or between working-class and middle-class subjects, it is found that the differences

2 Recent research by Gollob (1968) and Heise (1969) reveals that the above formulation must be modified, but an attitude-balance formula can be derived that is highly efficient in predicting attitude change.

—end p. 249—

between groups on the EPA dimensions, while statistically significant, do not approach the maximum possible values (Heise, 1966b).

These results indicate that attitudes toward specific concepts are not as divergent in different sectors of society as one might imagine; thus they imply (assuming an attitude-behavior relationship) that small differences in the distribution of attitudes have a large impact on behavior rates, or that persons in different sectors of society operate in terms of different concepts (Heise, 1966b).

Cross-cultural Comparisons—With cross-cultural validation and extension of SD measurements, the way is opened for cross-cultural comparisons of attitudes. This work has barely begun (Osgood, 1965; Jakobovits, 1966), but already results suggest that while major cultural variations do exist in attitudes toward various objects, there are also some striking uniformities. GIRLS, LOVE, and MARRIAGE, for example, seem to be positively evaluated in several cultures (Osgood, 1965). Much of the variation in attitudes is ecologically determined by the nature of the objects and, even cross-culturally, there is less variation in attitudes than one might expect.

The total image created by this work (admittedly mostly speculation at this point) is that cultural and group differences may amount to shifts and distortions of a basic attitudinal structure rather than complete reformations; while within any culture or group, at the individual level, there is a considerable amount of ongoing attitudinal variation and flux which probably contributes to variations in behavior.

SUMMARY

The SD is a general procedure for assessing affective responses. The technique has three features that distinguish it as an instrument for social psychological research. First, SDs are easy to set up, administer, and code. This, in conjunction with the demonstrated reliability and validity of the procedure, gives it favorable cost-effectiveness. Second, the EPA structure, which has an unprecedented amount of cross-cultural validation, is interesting theoretically, and measurements on all three dimensions yield a wealth of information about affective responses to a stimulus. The information that the three independent scores give about the character of responses inevitably is lost with alternative measures depending on unidimensionality. Third, since the form of an SD is basically the same whatever the stimulus, research using the SD (and methodological research about the SD) can cumulate.

The SD has been applied frequently as a technique for attitude measurement. Its usefulness in this respect is indicated by the wide variety of meaningful results that have been obtained. Further, SD measurements have been found to correlate highly with measurements on traditional attitude scales. There are, however, a number of questions in the use of SDs for attitude measurement.

When subjects are highly invested in a topic and want to give socially desirable answers, it may be advisable to use an instrument that is less direct than the SD. Social desirability ratings of SD scales correlate very highly with the Evaluation factor loadings of the scales. Thus, if subjects choose to distort their responses toward social desirability, Evaluation scores would be biased upward. If one does use the SD with especially sensitive topics (or respondents) it is worth taking some precaution to guard against social desirability effects (e.g., giving anonymity to respondents). Note, however, that Potency and Activity measurements should be free of this problem since the typical scales for measuring these dimensions are essentially free of social desirability contamination.

Thus far almost all applications of the SD to attitude measurement have relied only on Evaluation measurements. This appears to be an unfortunate tradition. A subjective examination of items in traditional attitude scales suggests that Potency and Activity do get involved in traditional attitude measurements. Furthermore, the multiple correla-

—end p. 250—

tions of EPA ratings with traditional scales often are much higher than the correlations of Evaluation ratings only with the scales. In the future it would be advisable to obtain ratings on all three dimensions when one is interested in attitudes. Almost certainly the full EPA information will increase the power of analyses.

Perhaps the most important general contribution of the SD is the provision of a single attitude space for all stimuli. This permits analyses, comparisons, and insights that. were virtually impossible with traditional instruments.