A TECHNIQUE FOR DETERMINING THE EVALUATIVE DISCRIMINATION CAPACITY AND POLARITY OF SEMANTIC DIFFERENTIAL SCALES FOR SPECIFIC CONCEPTS

Thesis for the Degree of Ph. D.

MICHIGAN STATE UNIVERSITY

Donald Keith Darnell

1964

A TECHNIQUE FOR DETERMINING THE EVALUATIVE DISCRIMINATION CAPACITY AND POLARITY OF SEMANTIC DIFFERENTIAL SCALES FOR SPECIFIC CONCEPTS

Thesis for the Degree of Ph. D.

MICHIGAN STATE UNIVERSITY

Donald Keith Darnell

1964

This is to certify that the

thesis entitled
A Technique for Determining the Evaluative Discrimination Capacity and Polarity of Semantic Differential Scales for Specific Concepts

presented by

Donald Keith Darnell

has been accepted towards fulfillment
of the requirements for

Ph.D. degree in Communication

David K. Berlo

Date November 14, 1963

O-169

A TECHNIQUE FOR DETERMINING THE EVALUATIVE DISCRIMINATION CAPACITY AND POLARITY OF SEMANTIC DIFFERENTIAL SCALES FOR SPECIFIC CONCEPTS

By

DONALD KEITH DARNELL

AN ABSTRACT OF A THESIS

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Communication

1964

ABSTRACT

A TECHNIQUE FOR DETERMINING
THE EVALUATIVE DISCRIMINATION CAPACITY AND POLARITY
OF SEMANTIC DIFFERENTIAL SCALES
FOR SPECIFIC CONCEPTS

by Donald Keith Darnell

The purpose of this study was to develop a technique of measurement which can be used to investigate the objective criteria that people use in making evaluative judgments about events or objects in their environment. The technique developed employs a scaling technique similar to that employed in the semantic differential but uses a different system of analysis.

The first chapter is devoted to a reassessment of the semantic differential, based on ten years research with that instrument. Empirical and theoretical arguments are presented that the results of SD research do not support the general conclusion that evaluations of events are "independent" of objective judgments for particular events or categories of events.

The major hypothesis of this study was: Bipolar adjectival scales such as those used in the semantic differential,

and including those identified in factor analysis as non-evaluative, can be shown to have an evaluative discrimination capacity for some concepts.

Subjects were asked to respond to the "best imaginable" and the "worst imaginable" examples of the categories of events named by concepts. Twenty concepts and 75 scales, borrowed from earlier research with the SD, were used.

The sign test was used to determine if there was a significant agreement among subjects on the polarity of each scale for each concept. It was assumed that the "best-worst" stimulus would permit each subject to indicate a preferred direction for each scale-concept item and that significant agreement among subjects on the relation between "best" and "worst" responses would indicate an evaluative discrimination capacity of the scale for the concept.

Affirmative results were obtained. Of the 46 scales identified in earlier factor analyses as non-evaluative, 44 showed a significant evaluative discrimination capacity. In all, 72 of 75 scales demonstrated this capacity.

A second hypothesis was also tested: There is a positive relation between discrimination capacity of a scale for a concept and the importance of that scale as an evaluative criterion for a particular concept.

Subjects ranked the 75 scales in order of importance to an evaluative decision about each of six concepts. A rank order correlation between importance and discrimination capacity provided support for the second hypothesis.

The conclusions of this study were:

1. The evaluative judgments that people make about events are related to their "objective" judgments of those events.

2. The objective criteria on which people base their evaluations of particular events are discoverable, using the best-worst technique.

3. The greater the statistical confidence in the evaluative discrimination capacity of a given scale the more likely it is to be an <u>important</u> criterion of evaluation.

4. The fact that a particular scale discriminates evaluatively (or does not) for a particular concept is not generalizable to other, unrelated, concepts.

The implications of this study for the semantic differential as a measurement technique, for meaning, and new directions in research are discussed.

A TECHNIQUE FOR DETERMINING THE EVALUATIVE DISCRIMINATION CAPACITY AND POLARITY OF SEMANTIC DIFFERENTIAL SCALES FOR SPECIFIC CONCEPTS

By

DONALD KEITH DARNELL

A THESIS

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Communication

1964

Acknowledgments

I wish to express my appreciation to my principal adviser, Dr. David K. Kerlo, and to Dr. Erwin P. Bettinghaus, Dr. Hideya Kumata, and Dr. Malcolm Maclean, Jr., who have provided counsel and encouragement throughout this program of research.

A special word of appreciation goes to Dr. Charles Osgood and his associates who contributed greatly to this dissertation and to my general intellectual development.

I must also thank Dr. Norma Bunton of Kansas State University who tolerated my preoccupation.

To Dan Costley and Mike Miller, for acting as interpreters with the computers, goes my sincere gratitude.

The greatest debt is owed to the family and friends who never doubted my sanity.

Table of Contents

List of Tables

List of Illustrations

Introduction

The semantic differential (SD) was first introduced in 1952 by Osgood and Suci, and it stimulated a flood of research which was summarized in 1957 by Osgood, Suci, and Tannenbaum.

Barely ten years after the debut of the SD, the writer is able to report more than seventy-five publications that make some mention of the technique. The name has gained sufficient importance to researchers of human behavior that it warrants a place in the index of Psychological Abstracts. The SD is currently a standard measurement technique at many universities and other research organizations. The variety of ways in which it has been applied is evident from a glance at the appended bibliography.

All of this argues the importance of the technique to a field short on measuring instruments. It does not, of course, imply that the SD is a technique without imperfections. Instead, it implies urgency in discovering the flaws that may exist in the technique and repairing these flaws.

A detailed description of the technique of semantic differentiation is available in several places (e.g., Osgood & Suci, 1952; Osgood, Suci, & Tannenbaum, 1957), and there

seems to be no need to reproduce that detail here. Perhaps it will add continuity, however, to give a brief review of that technique.

The Semantic Differential

"The semantic differential is essentially a combination of controlled association and scaling procedures" (Osgood et al., 1957, p. 20). It is a means of eliciting subjects' responses that indicate which member of a pair of adjectives is more closely associated with a particular concept, and the intensity of that association. In its most common form, the SD form looks like this:

TREE

good ____:____:____:____:____:____:____: bad

happy ____:____:____:____:____:____:____: sad

large ____:____:____:____:____:____:____: small

The subject (S) is instructed to mark in the middle of the scale if the adjectives at either end are equally associated with the concept at the top of the page. If one is more closely associated than the other, S can indicate "extremely" (by marking the box next to the stronger associate), "quite" (by marking the second box from the stronger

associate), or "slightly" (by marking the third box from the stronger associate--next to the center).

It is assumed that an adequate sample of such scales would provide a fairly specific profile of S's meaning for a concept. Given that each scale represents a choice among seven alternatives, and that with N independent scales S could differentiate the universe into 7^N categories (five scales would then allow for 16,807 categories of response), this does not seem an unreasonable assumption.

There is evidence that Ss can make these responses reliably (Osgood et al., 1957, Chapter 4; Norman, 1959). Norman tested reliability on the instrument for individuals in his sample and reports a median test-retest (four weeks intervening) reliability coefficient of .66. Osgood reports an over-all test-retest (immediate) reliability coefficient of .85. Both of these are reasonable when one considers that the indicated error includes mistakes in marking, changes in meaning, and differences in the situational context of the marking behavior.

The scaling procedure also has high face validity. In fact, the S's marks may be said to constitute a complex, graphic definition of a concept for an individual. Osgood et al. (1957, p. 142) also report significant relations

between SD markings and independent measures of attitude as well as significant predictions of voting behavior by "undecided" voters, which indicates a kind of predictive validity for the instrument. But, given that the meaning of a concept to an individual is his response to it, the validity of the SD rests on the assumption that Ss follow instructions and do, in fact, respond to the concepts by means of the scales.

One of the strong advantages that the SD has over other comparable instruments is the speed with which it generates data. It can be administered to groups of subjects limited in number only by convenience. It has been estimated (Osgood et al., 1957, p. 80) that an S can respond to 10 to 20 items per minute and can sustain a rate of 5 to 10 items per minute for periods up to an hour. (An item is defined as a scale-concept pairing.) The writer's experience in administering the SD indicates that this is probably a conservative estimate, or a fairly good estimate of the rate maintained by the slowest member in the average experimental group of college students. By comparison, S can respond to four SD scales in the time required for one typical multiple choice question.

The graphic scale system also presents an advantage to the experimenter if he wishes to compare two subjects or two concepts. Given an S's responses to concepts A and B on scales a, b, c, d, and e, a direct profile comparison can be made by preparing the following kind of chart:

a			A		B			-a
b		A				B		-b
c	A						B	-c
d			A		B			-d
e				AB				-e

It is easy to see at a glance that the two profiles have certain similarities and certain differences, and it does not seem unreasonable to assume that there are corresponding similarities and differences in the meanings of the two concepts. However, it is difficult to express what the eye can see, even in this simple example. When groups of subjects or concepts are compared it is necessary to translate to some other system, such as the language of mathematics, to express the complex relationships.

To make mathematical description possible, Osgood and his associates assume that the intervals of a scale are equal intervals and assign successive integers to the scale

categories. To further simplify the problem of communicating the comparative data collected from Ss about concepts, they set up a mathematical model described in The Measurement of Meaning (p. 25) as follows:

> We begin by postulating a semantic space, a region of some unknown dimensionality and Euclidian in character. Each semantic scale, defined by a pair of polar (opposite in meaning) adjectives, is assumed to represent a straight line function that passes through the origin of this space, and a sample of such scales then represents a multidimensional space. The larger or more representative the sample, the better defined is the space as a whole. Now, . . . , many of the "directions" established by particular scales are essentially the same (. . .) and hence their replication adds little to the definition of the space. To define the semantic space with maximum efficiency, we would need to determine that minimum number of orthogonal dimensions or axes (again assuming the space to be Euclidian) which exhausts the dimensionality of the space--in practice, we shall be satisfied with as many such independent dimensions as we can identify and measure reliably. The logical tool to uncover these dimensions is factor analysis, ...

Semantic differentiation is explained in terms of this model (p. 26) as

> the successive allocation of a concept to a point in the multidimensional semantic space by selection from among a set of given scaled alternatives. Difference in the meaning of two concepts is then merely a function of the differences in their respective allocations within the same space.

Osgood's development of the SD stems, at least in part, from his interest in a theory of meaning. In Method and Theory of Experimental Psychology (Osgood, 1953) and again

in The Measurement of Meaning (Osgood et al., 1957) he discusses his learning theory conceptualization of meaning. He defines meaning there as "representational mediating processes." These processes are said to be both responses and sources of self stimulation. These processes are learned. "They may well be purely neural events rather than actual muscular contractions or glandular secretions . . ." (1957, p. 7). And, it follows from the statement that they may be "purely neural events" that these meaning processes are, at present, not directly observable in an intact organism. On the other hand, with the *SD* there is a definition of meaning, a point in semantic space, that can be mathematically traced to the responses of subjects. Osgood says (1957, pp. 26-27):

> We now have two definitions of meaning. In learning-theory terms, the meaning of a sign in a particular context and to a particular person has been defined as the representational mediation process which it elicits; in terms of our measurement operations the meaning of a sign has been defined as that point in semantic space specified by a series of differentiating judgments. We can draw a rough correspondence between these two levels as follows: The point in space which serves us as an operational definition of meaning has two essential properties--direction from the origin, and distance from the origin. We may identify these properties with the quality and intensity of meaning respectively. The direction from the origin depends on the alternative polar terms selected, and the distance depends on the extremeness of the scale positions checked.

> What properties of learned associations--here, associations of signs with mediating reactions--correspond to these two attributes of direction and intensity? At this point we must make a rather tenuous assumption, but a necessary one. Let us assume that there is some finite number of representational mediation reactions available to the organism and let us further assume that the number of these alternative reactions (excitatory or inhibitory) corresponds to the number of dimensions or factors in the semantic space. Direction of a point in the semantic space will then correspond to what reactions are elicited by the sign, and distance from the origin will correspond to the intensity of the reactions.

After several paragraphs designed to "clarify this assumed isomorphism somewhat," the writers (Osgood et al., 1957, p. 30) conclude with three statements that are significant for the present study.

1. "This, then, is one rationale by which the semantic differential, as a technique of measurement, can be considered as an index of meaning."

2. "It is true that many of the practical uses of the semantic differential, indeed its own empirical validity, depend little, if at all, on such a tie-in with learning theory."

3. "If we are to use the semantic differential as an hypothesis testing instrument, and if the hypotheses are to be drawn from learning-theory analysis, some such rationale as has been developed here is highly desirable."

It seems to be clearly implied by these three statements that one need not concern himself with the mediation hypothesis to be concerned with the semantic differential, and since it is the purpose of this study to revaluate the

measuring technique, it does not seem desirable to employ any unnecessary theoretical assumptions.

Purpose of This Dissertation

This dissertation is intended to focus attention on the problems that have arisen in the application and interpretation of the semantic differential. Most of the problems that will be discussed have been pointed out by the authors of The Measurement of Meaning; however, these problems have not been attacked systematically to determine what revisions in the technique they might suggest.

In the course of the discussion, data will be presented to show that (a) certain assumptions made in the standard SD analysis are untenable, (b) there is a more parsimonious explanation than the one given for the results obtained so far, (c) the factor structures obtained, though they may describe the universe of concepts within sampling limitations, may not describe any particular member of that universe, and (d) there is a need not met by the SD that could be met by a slight modification of that technique.

An alternate technique will be described which (a) shares the advantages of ease and speed of data collection with the SD, (b) makes fewer assumptions than does the SD, (c)

permits a new kind of semantic differentiation, and (d) has direct and immediate implications for efficiency in persuasion.

The two techniques will be compared--in terms of their relative appropriateness to various kinds of problems--through analysis of two sets of data collected from the same subjects and using the same scales and concepts.

Chapter 1

This chapter is a discussion of methodological problems that are associated with the use of the semantic differential and a suggestion for the solution of those problems.

Methodological Problems with the Semantic Differential

In his review of The Measurement of Meaning, Gulliksen (1958, p. 116) summarized a number of the problems that arise in the interpretation of SD data.

> From the point of view of the general stability of the data, it is encouraging to find that several different factor studies give similar results. With regard to factor analysis, however, the authors mention (Chap. 4) a number of disturbing characteristics of the data, such as "concept-scale interaction" (p. 187), variation in scales contributing to a given factor (p. 180), variation in inter-scale correlation for different concepts (p. 177), ... In the same vein, Osgood states that "the vast majority of scales show significant variation in their correlations with other scales across concepts" (p. 177), that there is variation in the "relevance" of particular scales to particular concepts (p. 78), and that some scales shift in meaning with the concept being judged (p. 179).
>
> The foregoing comments may be summarized by saying that there is a marked "concept-scale interaction." For data which exhibit this characteristic, a general factor analysis of a number of concepts may give quite misleading results. Such interaction means that the emphasis throughout the book on correlational analysis is to be regretted. Other methods of analysis should be considered.

There are at least nine references in The Measurement of Meaning to concept-scale interaction (pp. 39, 93, 108, 176,

177, 178, 187, 200, and 326). The import of this interaction to the interpretation of SD data is suggested by the following examples from these references.

> To the extent that there are differences in factor structure as between concepts, and to the extent that our sampling of only 20 concepts was nonrepresentative, the factorial results of the first analysis could be biased.
>
> To the extent that the relations among scales (and factors) vary with the classes of concepts being judged (see section in Chapter 4 on comparability across concepts), some error in the interpretation of D is being introduced for certain concepts.
>
> For purposes of generalized semantic measurement we would like to have a set of scales which consistently load heavily on a certain factor and are independent of other factors, despite variations in the concepts being judged. We have had difficulty trying to isolate a set of scales having these properties.
>
> What do these findings have to say about the practical problems of semantic measurement? For one thing, it now seems less likely that we will be able to discover a single set of scales which represent an adequate set of factors and which are stable across whatever concepts may be judged. On the other hand, it may be possible to identify classes of concepts for which general instruments may be used, and perhaps, in course, the principles which operate in determining a common frame of reference can be discovered.

These statements indicate that Osgood and his associates were not entirely satisfied with the results that had been obtained by 1957. But, enthusiasm is apparently easier to communicate than caution. More than half of the empirical studies reported in the journals since that time have

borrowed scales on the basis of the general factor loadings and applied them without qualification to many different kinds of concepts. However, three research projects, carried out since 1957, show clearly the need for caution (Osgood, Ware, & Morris, 1961; Smith, 1959, 1961, 1962; Triandis, 1959, 1960).

Smith's three studies employed scales chosen from Osgood's lists on the basis of factor loadings on the evaluation, potency, and activity factors and "literal application to speech concepts." His concepts were "speech related," "theater," and "speech correction" concepts. With each new set of concepts and each new factor analysis, differences in the factor structures were noted--in spite of the fact that he started with a select set of scales, and all of the concepts were drawn from the "same" academic discipline. Smith (1961) remarked,

> . . . the dimensions of any special subject matter area must be individually determined even with areas as closely related as those of general speech and the theater arts since there are both factor and scale variations in significant amounts. This necessitates for any special area of investigation in which the semantic differential is to be used, a specific factor analysis to determine the important factors and the scales which measure them.

Smith (1959) also noted two other problems with the SD. From the fact that his subjects seemed to treat "worthless"

and "meaningless" as positive values, he inferred, "It is impossible to determine an absolute scale polarity apart from the conceptual structure within which it is to operate." This bit of evidence, added to earlier findings (Osgood et al., 1957, p. 68) that scales may reverse polarity with a change of concepts, raises a serious question about the meaning of correlations between scales across concepts using a constant polarity assigned by the experimenter.

The third problem noted by Smith (1961) was that the hot-cold scale seemed to be the best available measure of Factor I, although subjects could neither apply nor interpret it. This seems to be related to the fact noted by Osgood et al. (1957, p. 323) that they were unable to work back from the profile or the point in semantic space to identify the concept. At any rate, it emphasizes the fact that scales vary in relevance from concept to concept (Osgood et al., 1957, pp. 78-79).

Triandis (1959, 1960) used restricted samples of job-related concepts and a quite different set of scales. He found factors that differed from those found by Osgood, and "there were also certain differences between the factors obtained from managers and those obtained from the workers." He describes (1960, p. 300) the changes in factor structure as follows:

> Instead of evaluation we have objective and subjective job evaluation factors. Instead of potency and activity we have a fusion of the two in a relatively insignificant dynamism factor. New factors, such as the white collar, variety, and job level factors, that are specific to the job domain of meaning, have taken the place of the potency and activity factors and account for a portion of the variance that was previously accounted for by these factors.

It is difficult to tell how much of this change in factor structure must be attributed to the changes in scales, how much to the changes in concepts, and how much to the author's interpretation. It probably makes little difference, for it is clear that the reproducibility of the original factor structure is less than satisfactory--for one reason or another.

Osgood, Ware, and Morris (1961), in their study of values, used a comparable set of subjects and some of the same scales as used in an earlier study, but they found an entirely different factor structure in this restricted sample of concepts. The following quotations from their report round out the empirical case for concept-scale interaction:

> The scales selected for this study provide ample opportunity for at least the three general factors usually obtained, "evaluation," "potency," and "activity," to appear (p. 67).

> The semantic space of connotative meanings generated when these value statements (the Ways) are judged is

> clearly not the same as that obtained when more varied samples of concepts are used (p. 68).
>
> In the case of value statements (the Ways) being judged as concepts, by American students, "evaluation," "potency," "activity," and "receptivity" fuse together as a single "successfulness" factor (p. 69).
>
> Comparing these results with those of other factor analyses (cf. Osgood, Suci, & Tannenbaum, 1957, Ch. 2), then, we have clear evidence for concept-scale interaction (p. 69).
>
> The results . . . make it clear that factors derived from more "representative" samples of concepts are not necessarily independent, and hence visible, when some specific subset of concepts is judged (p. 72).

What does all this mean? How is it that factors derived from representative samples of concepts are not visible when some specific concept, or set of concepts, is judged? What difference does it make? Perhaps all three questions can be answered with an example.

Take the three scales that name the three general factors mentioned above, good-bad, strong-weak, and active-passive. To simplify the example, assume that these are dichotomous like heads-tails. On one "flip" then, it would be possible to obtain any one of eight combinations. And, if it is possible to think of a word in English appropriate to each combination, it can be said that the three dichotomies are logically independent--all combinations can occur. The eight combinations are enumerated below with some likely

candidates for the set of concepts that would prove logical independence.

g s a - ATHLETE
g s p - ARMY RESERVE
g w a - KITTEN
g w p - ANTIQUE CHAIR
b w p - DERELICT
b w a - HOUSEFLY
b s p - QUICKSAND
b s a - ROGUE ELEPHANT

If it is agreed that these concepts are not only possibly described by the three adjectives indicated but that subjects would be highly likely to pick just those combinations to go with those concepts, then it is also highly likely, in standard SD procedure, that the three scales would correlate zero with each other if the data were summed over these concepts. However, it is also evident that the relation of independence (factor structure) indicated by summing over concepts does not hold for any particular concept in the sample--that the concepts fit better with some combinations of adjectives than with others. Thus, if one were to apply these scales to a set of concepts which name subclasses of one of these concepts (e.g., FOOTBALL PLAYER, BASEBALL PLAYER, MILER, and BOXER), he would likely obtain significant correlations between the scales (or indeterminate ones if the variance among subjects were very low).

Adding a set of DERELICTS to the set of ATHLETES would assure sufficient variability in the data to permit significant correlations between the three scales in question (see Osgood, et al., 1957, p. 35), and, in either case, a quite different picture of semantic space is obtained than that produced by the broader set of concepts.

Osgood et al. (1957, p. 180) and Bettinghaus (1961 a) have made comments to the effect that one would not expect the evaluation of some concepts (ATHLETE, POLITICIAN, SECRETARY) to be independent of activity, potency, or stability. Experience tells us that there is a relation between strength and stability for rigid structures and between strength and activity for living organisms. Yet, these seem to be independent factors in diverse samples of concepts. All of this supports the proposition that the preceding example is not an isolated one.

There is an explanation for this seeming inconsistency. Factor structures depend on correlations, and correlations depend on variance--they are indices of covariation. In the normal SD analysis, there are two kinds of variance contributing to the outcome. Variance_1 is the dispersion of scores around the concept means. Variance_2 is the dispersion of concept means around the grand mean.

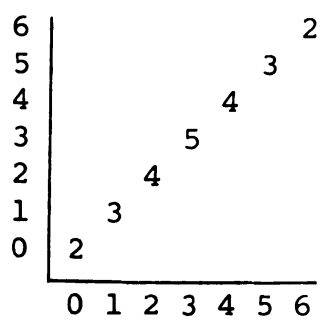
Variance$_1$ and variance$_2$ are independent in that neither is predictable from the other on a priori grounds. Thus, there is no reason to expect that a factor analysis based on either kind of variance alone would be the same as one based on both. It is possible to construct plausible cases in which the addition of two concepts--the inclusion of variance$_2$--would increase, decrease, and have no effect on the correlation between two scales based on the covariation within either concept.

Osgood, Ware, and Morris (1961) found that when V_2 was severely reduced, by the use of closely related concepts, "the factor structure was clearly not the same as that obtained when more varied samples of concepts are used." They also found that V_1 could be eliminated (by using the concept means in the correlation instead of individual scores) with no effect on the pattern of loadings, although the proportion of the total variance accounted for was greatly increased. These findings strongly suggest that the factor structure is heavily dependent on V_2--the among-concept variance.

The most familiar SD factor structure was based on concepts selected on the criterion "that they be as diversified in meaning as possible so as to augment the total variability

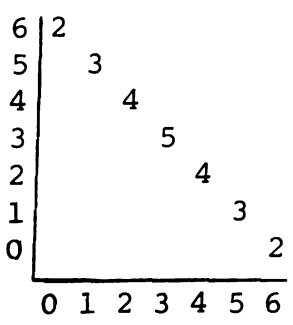
in judgments" (Osgood et al., 1957, p. 34). Again these authors say (p. 85), "Ordinarily in making up a sample of concepts for a differential we try to balance off good concepts with bad, strong with weak, and so forth. . . ." Given these pieces of information, it seems reasonable to expect that when concepts are selected on other criteria (such as a specific problem of meaning) that different results would be obtained.

Given that scales may reverse polarity with a change of concepts (Smith, 1959; Osgood et al., 1957, p. 68), there is still another way that adding concepts together can affect the correlations between scales. Figures 1, 2, and 3 illustrate a possible relation between the scales <u>good</u>-<u>bad</u> and <u>large</u>-<u>small</u> for the concepts PROFIT and LOSS. If 23 subjects were to respond to the concept PROFIT as in Fig. 1 and to the concept LOSS as in Fig. 2, the results



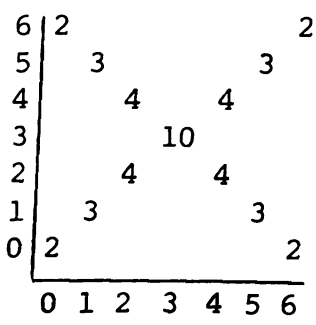
$\underline{r}_{xy} = 1.00$

Fig. 1



$\underline{r}_{xy} = -1.00$

Fig. 2



$\underline{r}_{xy} = 0.00$

Fig. 3

would be completely obscured by adding the two concepts together, as indicated in Fig. 3. The numbers in the figures are frequencies, and their positions in the matrices indicate covariation on the two scale axes. Although no data has been reported on the frequency with which this polarity reversal might occur, it has been observed in SD data, and the possibility must be taken into account in the interpretation of SD results.

If the line of reasoning in the preceding discussion is valid, and concept-scale interaction is inevitable in the SD technique, then it follows that there are as many factor structures or "semantic spaces" as there are samples of concepts. If that is true, then the SD must be treated as a purely descriptive technique and a relatively inefficient one at that. That is, for anyone not highly sophisticated in mathematics, it probably takes more effort--more words--to describe the meanings of a concept in terms of a factor pattern than would be required without the factor pattern. And, without the inferential power that derives from an assumption of a fundamental pattern there is little to be gained for the effort.

"On the other hand, it may be possible to identify classes of concepts for which general instruments may be

used, and perhaps in course, the principles which operate in determining a common semantic frame of reference can be discovered" (Osgood et al., 1957, pp. 326-327). There are, however, reasons why the suggested reorientation may not be entirely effective without some changes in the technique.

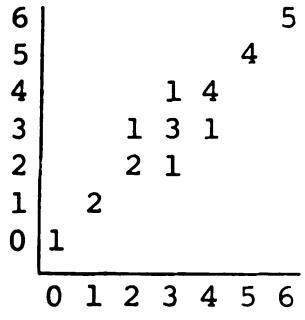
The first of these reasons to be considered involves the assumption of equal intervals in the semantic differential scales. This assumption is essential to the usual method of analysis, but the evidence that it is tenable may be described as inconclusive. Osgood et al. (1957, pp. 146-152) report evidence that the intervals of the nine most frequently used scales are approximately equal. The argument, of course, is not whether the intervals are actually equal but whether a significant distortion is introduced by assuming equal intervals. Osgood argued that little distortion would be introduced for these nine scales. However, no evidence is available on this question with regard to less frequently used scales, the ones that have behaved less predictably in analysis. Since the cost of such evidence would be quite high, it seems advisable, for the moment at least, to employ some means of analysis with SD data that does not require this assumption.

Gulliksen (1958) suggests another problem in the SD technique. His argument concerns the advisability of using the linear correlation coefficient instead of the curvilinear coefficient. He points out that the linear correlation may lead one to draw unjustified conclusions about the presence or absence of functional relations, because a high linear correlation does not imply that the curvilinear relation is negligible. Of course, it is also true that a curvilinear relation may exist when the linear coefficient is zero (see Ferguson, 1959, p. 109). The greater sensitivity of the curvilinear coefficient, eta (E), relative to the linear coefficient, r, can be made clear by comparing basic assumptions. When one computes a correlation coefficient between two variables (X and Y) he predicts that there is a line that can be fitted to the data matrix which will enable him to predict X from Y or Y from X with greater success than he could from the mean of the predicted variable. E puts no restriction on the nature of the line, but r restricts the possibilities to straight lines. In other words, r assumes that the relation is linear or nonexistent. When this assumption is tenable, r has its merits: A single coefficient can describe the reciprocal relation between X and Y while two E coefficients are required. An

r may be calculated from raw data on a desk calculator, but *E*, which is essentially an analysis of variance, requires that the data be grouped and plotted, or more complex equipment. Finally, *r* shows direction of the relationship (ranges from -1.00 to +1.00) which *E* does not do. Thus, if its straight line assumption is met, *r* is more efficient than *E*, but, according to Senders (1958, p. 271),

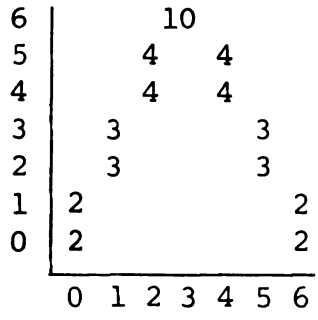
> When there is any doubt about whether or not the relationship between two variables is linear, both *E* and *r* should be computed. *E* will always be equal to or greater than *r* in absolute value, but if the relationship is linear the difference will be small. A large difference indicates a non-linear relationship, in which case *E* rather than *r* should be used.

The following examples should make the import of Sender's statement clear. The numbers in the matrices are hypothetical frequencies. Computation is after Walker and Lev (1953, pp. 238 & 279). Figure 4 illustrates the case in which *r* and *E* are equally good predictors, because the relation is linear. Figure 5 illustrates the case in which *r* is zero, and the predictability of X, given Y, is zero, but the predictability of Y, given X, is nearly perfect. Figure 6 illustrates a case in which either variable is perfectly predictable from the other, but *r* shows only about 13% of the variance accounted for.



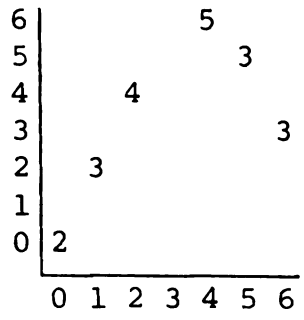
r_{xy} = .97
E_{xy} = .97
E_{yx} = .97

Fig. 4



r_{xy} = .00
E_{xy} = .00
E_{yx} = .97

Fig. 5



r_{xy} = .36
E_{xy} = 1.00
E_{yx} = 1.00

Fig. 6

The only question that remains is whether there is any doubt about the relations between SD scales being linear. McNelly (1961), in searching for an index of "interest" in news stories about "countries," found a strong positive relation between <u>intensity</u> of evaluation and <u>absolute</u> strength and activity. He argued that it made sense to think that news about a country would become more interesting as the perceived strength and activity of that country increased. It also seemed reasonable that extremely good or bad countries should be more interesting than neutral countries, and his findings support these contentions. McNelly's findings also suggest the possibility of curvilinear relations between the evaluative scale and the other two for this particular set of concepts. The relation could be that in

Fig. 5 with passive-active or weak-strong as the ordinate axis and good-bad as the abscissa. In the normal SD analysis, making the linear-or-nothing assumption, this relation would be overlooked. The reverse curvilinear relation with good-bad might reasonably be predicted for strong-weak on the concept COFFEE, for fast-slow on the concept CLOCK, or hot-cold in regard to the weather. Although proof that E would be more appropriate than r in SD analysis is limited, given concept-scale interaction, there seems to be a reasonable doubt that the linear assumption is universally tenable.

Osgood et al. (1957, p. 91), in their justification of the use of D, point out that "the product-moment correlation not only distorts the information, but may be completely inapplicable in some cases." They refer to the fact that correlation disregards differences in the means of the two variables correlated, but the truth of the statement seems to be generalizable to some other instances in which the technique has been employed.

Given that the scales do not behave the same way from concept to concept--that there is concept-scale interaction--the SD must be restricted to the task of describing the relations among a specific set of concepts. Further,

it has been argued, there are problems inherent in the technique that make its value as a descriptive instrument something less than certain. On the other hand, it must be noted that the weaknesses in the SD seem to be in the technique of analysis while the strengths (speed and ease of data collection, reliability, and validity) rest on the scaling technique itself. That alternative methods of analysis should be explored seems to be the most reasonable conclusion.

An Alternative Approach

The original problem that prompted this survey of the SD seems a reasonable starting point for laying the foundation for the alternate approach.

It has been observed that people rather consistently make evaluative judgments about events in their environment. They show preferences for one event over another. They can quite often give reasons (or rationalizations) for their preferences, such as, it's sweeter, more dependable, more durable, smells better, cheaper, larger, smaller, faster, slower, hotter, colder, or it matches my shoes. Observations like these suggest the hypothesis that there are discriminable characteristics of events that are linked by

some psycho-logical value system to evaluative judgments about those events.

A question was formulated: Is there some efficient way of finding out what these criteria of evaluation are, for particular events? The semantic differential was immediately suggested, but, upon examination, it seemed that the SD was not only unsuitable for the task, but the results seemed incompatible with the hypothesis. That is, the SD seemed to say that evaluative judgments are independent of all kinds of sense related discriminations, across concepts.

This conclusion, based on repeated SD results, is easily countered by the technique of argument called reductio ad absurdum: If it is true that evaluative judgments are independent of sense related discriminations, then it must be true that evaluative judgments are independent of the events themselves. And, if it is true that evaluative judgments are independent of the events being judged, then it is as likely that an individual would feel favorable toward a rattlesnake bite as that he would feel favorable toward a dish of ice cream.

The conclusion to this line of reasoning is obviously false, but it does not explain the contrary result of the SD research. It is believed, however, that a sufficient

explanation for the apparent independence of evaluative and objective judgments has been given in the first part of this chapter, so that it is now reasonable to attempt to support the counter proposition--that evaluative judgments are related to objective characteristics of events.

Any kind of reliable discrimination between two events, evaluative or otherwise, would seem to require that the events be objectively different in at least one respect. In discrimination learning experiments, for example, if an experimental subject develops a reliable preference for one of a pair of stimuli, under conditions of controlled reinforcement, it is taken for granted that S has discovered the relevant distinctive cue, which in various experiments may be weight, brightness, size, configuration, etc. (Osgood, 1953, pp. 446-453).

It does not follow from this proposition, however, that the heavier, brighter, or larger of the pair of stimuli will be universally preferred. The "direction" of the preference depends, instead, on the "direction" of the reinforcement. In the reference cited immediately above, experiments are mentioned in which the "significance of these cues" was intentionally reversed by the manipulation of the reinforcing conditions, but in all cases it was assumed that the

formation of a reliable preference depended on the association of the reinforcement differential with discriminable differences in the stimuli.

From this assumption it follows that there are at least two necessary conditions for the development of a reliable preference; discriminable differences in the stimulus events, and discriminable differences in the reinforcement conditions associated with each of the stimuli. If this is true, then, a reliable preference implies both of these conditions but is not, in turn, implied by either.

In the laboratory situation, the relation between reinforcement and the observable qualities of the stimulus is usually, by design, quite arbitrary and in the control of an experimenter. The arbitrariness of the relation permits the experimenter to exercise his control in the manipulation of the subject's preferences.

In the natural (non-laboratory) situation, however, this level of control does not obtain, and the predictability of preferential responses is nil under the assumption that the reinforcement differential and the observable qualities of the stimulus are independent. On the other hand, if the relation is not one of independence, an observer with a knowledge of the relation and of the stimulus should enjoy

approximately the same predictive power as the experimenter who controls both the stimulus and the reinforcement. And, it seems likely that the reinforcement associated with a given stimulus is frequently, in the real world, directly (causally) related to the objective properties of the stimulus, limited of course by the wants of the individual and his ability to employ the stimulus in the satisfaction of those wants.

If this is true, one might ask, why have these relations not been discovered, in SD research for instance? The answer has already been given. It has been assumed that the relations are linear and constant across categories or nonexistent. But, the attribute of an event that is associated with positive reinforcement in one category of events, or in one situation, may be associated with negative or non-reinforcement in other categories or in other situations. For example, heaviness is usually quite desirable in football players and quite undesirable in jockies. Again, certain attributes of coffee which make it very desirable to some people at 7 a.m. make it very undesirable to the same people at 10 p.m. In other words, there is reason to believe that the "significance of cues" quite naturally changes from one category of events to

another, from one combination of cues to another, so that a given objective cue can be strongly related to evaluations of particular categories of events, and yet appear to be independent of those evaluations when examined across categories under the linear-or-nothing assumption. Examples were given earlier showing how this change of the relation might affect the computation of a correlation.

Using the SD factor analysis as a base, evaluative judgments appear to be independent of objective criteria insofar as evaluative and objective variables are measured by the semantic differential. Yet, logically, evaluative judgments must be related to objective variables if they are related to events at all. Or, if events and evaluations are independent, no predictive propositions about evaluative behavior can be made from knowledge of physical events. The assumption that there is a "natural" change of cue significance from category to category provides a very simple and parsimonious explanation for this apparent inconsistency. The evidence of concept-scale interaction, reported earlier, supports this assumption. It seems reasonable, therefore, to employ this parsimonious assumption and to hypothesize that evaluative judgments are related to objective variables for particular concepts or categories.

In retrospect, Osgood et al. (1957, pp. 62, 180, 188, & 78) offer some support for the position taken here.

> The evaluative factor is itself further analyzable into a set of secondary factors--various "modes" of evaluation which are appropriate to different frames of reference or objects of judgment.
>
> What is good depends heavily upon the concept being judged--strong may be good in judging athletes and politicians, but not in judging paintings and symphonies; harmonious may be good in judging organized processes like family life, symphony, and hospital, but not so much so in judging people or objects.
>
> Evaluation thus appears as a highly generalizable attribute which may align itself with almost any other dimension of meaning, depending on the concept being judged--and it is most often the dominant attribute of judgment.
>
> Another criterion in scale selection is relevance to the concept being judged. For example, in judging a concept like ADLAI STEPHENSON, one evaluative scale like beautiful-ugly may be comparatively irrelevant while another like fair-unfair may be highly relevant; on the other hand, just the reverse would be true for judging paintings.

These are only a few of the more than thirty instances in The Measurement of Meaning that this writer interprets as supporting this proposal. But, the most important one to this argument is, "What is good depends heavily upon the concept being judged. . . ," for from this assertion it follows that evaluative judgments are not independent of the concepts being judged and are not independent of the objective attributes of events named by those concepts. If this

is true, and the nature of the relationship for particular events can be ascertained, then it should be possible to predict evaluations and changes in evaluation with greater success than we now enjoy and to change or stabilize an evaluation more effectively by the efficient use of influence. None of these things is likely to be accomplished under the assumption that evaluations are independent of the objective observations that a person makes of events.

There are several ways that the validity of this argument could be tested empirically. But, the most direct way also happens to be the one that offers the greatest promise of generality, because it has a methodological emphasis.

In the following chapter, a design will be presented for a study to test the hypothesis that bipolar adjectival scales such as those used in the semantic differential, and including those previously identified as non-evaluative, have an evaluative discrimination capacity for some concepts. Since extensive correlation analysis is not available to many people who might be interested in the question posed here, and since there are serious doubts about the appropriateness of that type of analysis to data produced by the SD scaling technique, a simpler alternative method will be introduced to test this hypothesis.

Chapter 2

This chapter includes the designs for two experiments. The second experiment is contingent upon the outcome of the first one, so its design will appear as a separate section following the complete design of the major experiment.

Experiment 1

Hypothesis 1

Bipolar adjectival scales, such as those used in the semantic differential, and including those identified by factor analysis as "non-evaluative," have an evaluative discrimination capacity for some concepts.

Rationale

The elements of a rationale for this theoretic hypothesis have been given in Chapter 1, but they may be summarized as follows:

It is assumed that a necessary condition for a reliable discrimination between two events is that the discriminator reliably perceives an objective difference between the two events. It is further assumed that an individual cannot reliably perceive differences between events when there are, in fact, no differences between the events. These

assumptions imply that if an individual makes an evaluative judgment about an event, it either does not differentiate that event from any other, or it is related to some objective characteristic of that event, and that characteristic is a variable among events. Therefore, evaluative judgments must be related to objective variables or they are independent of events.

If the latter is assumed, then there is no ready explanation for the variability and apparent reliability of evaluative judgments and no basis for predicting the behavior of an individual from a knowledge of his environment. On the other hand, acceptance of the proposition that evaluative judgments are related to objective variables presumes an explanation for previous research findings obtained with the semantic differential.

In the previous chapter, two kinds of explanations were suggested for the apparent independence of evaluative and objective judgments: (a) It may be partly accounted for by the fact that the linear correlation is insensitive to certain kinds of relations. Several examples were given to illustrate the importance of the fact that the curvilinear relation is always equal to or greater than the linear relation. (b) It may be the direct result of the practice

of summing across concepts which assumes that any relations that do obtain between evaluative and non-evaluative judgments are themselves independent of the categories of events (concepts) being judged.

This study, then, assumes that evaluative judgments are related to objective variables; that bipolar adjectival scales, such as those used in the semantic differential, do measure both evaluative and objective variables; that events within a category may be evaluated differently if and only if they are objectively different; and that concepts name categories of events. It does not assume that the evaluative or the objective variables or the relations between evaluative and objective variables are independent of the events or categories of events being judged. If this position is a tenable one, then it should be possible to support the theoretic hypothesis given above.

Definitions

Given that a set of semantic differential responses are factor analyzed and two or more orthogonal factors are obtained; that factor with which the <u>good-bad</u> scale is most highly correlated is the <u>evaluative factor</u>. Any other factor is <u>non-evaluative</u>. The correlation between any

scale and the evaluative factor is the evaluative factor loading for that scale, and this correlation squared is an indication of the variance in the markings on that scale accounted for by the evaluative factor. Given the variance accounted for by the evaluative factor (e^2) and the total variance accounted for by all the factors (h^2); if the quantity $2\ e^2 - h^2$ for a given scale is greater than zero, that scale is predominantly evaluative. If that quantity is less than zero, that scale is predominantly non-evaluative.

Given sets of subjects' responses on a given scale to the "best imaginable" and the "worst imaginable" events in a category named by a concept, if a significant proportion of the sample of subjects agree on the polarity of the scale, the scale has an evaluative discrimination capacity for that concept. Since this definition is the key to this whole experiment, it is elaborated elsewhere in the design, but it is given here in summary form to help the reader understand those elaborations when they occur.

Design

Subjects. The sample consisted of all those people enrolled in the seven sections of Oral Communication Ia, Spring, 1963, at Kansas State University, who attended class

on a given day. Of the 159 enrollees in this class (73% male and 91% freshmen), 139 actually participated in the experiment.

Concepts. The concepts used are the same twenty used in the first analysis by Osgood et al. (1957, p. 34). The original rationale for the selection of these concepts can be found in the reference cited. They were selected for this study so the results would be directly comparable to the earlier work. The concepts are: LADY, BOULDER, SIN, FATHER, LAKE, SYMPHONY, RUSSIAN, FEATHER, ME, FIRE, BABY, FRAUD, GOD, PATRIOT, TORNADO, SWORD, MOTHER, STATUE, COP, and AMERICA.

Scales. The 75 scales used in this study include the 50 from the first analysis by Osgood et al. (1957, p. 37) and an additional 25 selected from their thesaurus study (pp. 53-61). The extra twenty-five scales were selected on the basis of their non-evaluative factor loadings in the thesaurus study to compensate for the fact that the first set was predominantly evaluative. The complete set of 75 scales may be obtained from Table 1.

Instructions. The instructions for this study were considerably different than the ones used in the earlier research (Osgood et al., 1957, pp. 82-84). The complete

instructions may be found in Appendix A, but the essential difference is this: Subjects were asked to mark, on each scale of a set, a "B" to indicate their feeling toward the "best imaginable" and a "W" to indicate their feeling toward the "worst imaginable" example of the class of things named by the concept at the top of the page.

Administration. Forty sets of scales were prepared for each concept, the scales appearing in a random, but constant, order for all concepts. Test booklets were compiled containing five concept sets and one instruction sheet. In preparing the test booklets, the concept sets of scales were arranged in an arbitrary order, and starting with the "first" set, five concepts were stapled together with an instruction sheet. The second booklet started with the "second" concept, the third booklet started with the "third" concept, and so on. Thus, any two contiguous booklets had four concepts in common, and all concepts occurred an equal number of times in the five possible positions. The test booklets were distributed, in the order of preparation, to subjects as they seated themselves in the classroom.

Given the manner of distribution, the fact that not all the booklets were used, and the fact that a few Ss failed to complete all the scales in a booklet, the number of

subjects actually responding to a given concept-scale pair ranges from 28 to 39.

After the subjects were seated, the experimenter entered the classroom with the regular instructor and was introduced as a fellow member of the speech faculty. The instructor encouraged the students to cooperate in the experiment and then, in most instances, left the room. The experimenter distributed the test booklets and allowed three minutes for the reading of the instructions. After this period, the experimenter answered any questions about the instructions and remained in the room to remind the subjects, periodically, of the passage of time.

Subjects were encouraged, but not required, to put their names on the test booklets. This procedure resulted in about 30% anonymous questionnaires. However, since a fairly adequate description of the population, of which the sample was 87%, was available in class records, it did not seem advisable to insist on identification after a member of the first group "identified" the questionnaire as a personality inventory.

Analysis. Given the data obtained by the method described above, the decision of whether or not a particular scale has an evaluative discrimination capacity was based on

the sign test, a non-parametric statistic. With this test, a null hypothesis was tested for each concept-scale item--that the number of subjects who placed their response to the "best example of the concept" to the left of their response to the "worst example of the concept" is equal to the number who indicated the opposite direction of preference. Given the kind of flip-flop tendency that had been anticipated, the two tailed test seemed most appropriate. It was decided to test this hypothesis at the 95% level of confidence.

The sign test is described in Siegel (1956, pp. 68-75). According to Siegel, "The only assumption underlying this test is that the variable under consideration has a continuous distribution." For this reason it was chosen for use in this study in preference to other tests which assume interval data or independence of samples.

Application of the sign test is a scoring task rather than a computational procedure. Looking at a set of responses to a scale-concept item, each subject who places "B" on the left of "W" but not in the same scale interval scores a plus. Each subject who places both marks in the same cell scores zero and drops out of the sample. A subject who places "W" on the left of "B" scores a minus. Starting with 20 Ss one might obtain ++++++++++++++ 00---.

The sign test is related to the binomial expansion, and the test of significance is a binomial test. The effect of a tie is to reduce N. The probability of obtaining an event as rare as some particular outcome can be determined by reference to a table of binomial probabilities. In the example just given, the probability of obtaining 3 or less of either sign, with an N of 18, is .008, and the null hypothesis would be rejected. If N is greater than 25, the normal approximation to the binomial distribution can be used. In this study, z (corrected for continuity) was computed for estimation purposes and for use in Experiment 2, although the reduced N was expected to be less than 25 in some cases. The actual decision to reject the null hypothesis, however, was based on the exact binomial probability.

If, for a given concept-scale item in this study, the sign test yields a significant value, the inferences made are (1) that the subjects can make evaluative distinctions among events in the concept category and (2) that these evaluative distinctions are related to discriminations that they make within the concept category on the dimension named by the test scale. A scale, then, will be said to have an evaluative discrimination capacity for a concept if, and

only if, the null hypothesis of the sign test is rejected at the 95% level of confidence.

This study set out to obtain, along with an estimate of the evaluative discrimination capacity, an objective criterion for determining scale polarity for each scale and concept. The sign test data provide just such an objective criterion. If the scale shows a significant evaluative discrimination capacity, then it may be said, also, that a significant proportion of the sample of subjects indicates a directional preference on the scale. Given a directional preference, then, the left end of the scale is preferred if the number of plus signs is greater, and the right end of the scale is preferred if the number of minus signs is greater.

Since, in this case, the sign test is applied to individual scale-concept items, in order to make a general statement about the discrimination capacity of a given scale, it is also necessary to show that a scale discriminates for more than 5% of the concepts tested (the alpha level on the individual tests being .05). Reference to a table of binomial probabilities shows that the probability of obtaining four or more significant values out of twenty, when the probability of each one is .05, is less than .05 (.016). Thus

if a scale discriminates for four or more of the twenty concepts, the null hypothesis that the scale does not discriminate for any concept can be rejected at the 95% level of confidence.

Controls. As discussed earlier under the heading of "administration," the normal controls--standard questionnaires, standard instructions, familiar surroundings, and a single experimenter--were observed in this study.

For reasons of convenience and available subject time, the subjects were tested in seven separate sessions, and each subject contributed data to only five of the twenty concepts. To control for subject and group differences that might result from these arrangements, the test booklets were systematically distributed so that no significant finding could be accounted for by any experience unique to a particular class section or peculiar to a single administration of the test.

Discussed earlier, under the heading of "definitions," was the proposition that a scale is labeled "evaluative" or "non-evaluative" on the basis of a factor analysis, by the formula $2\ \underline{e}^2 - \underline{h}^2$. Although values for $\underline{e}$ (the evaluative factor loading) and $\underline{h}^2$ (the proportion of the variance on a given scale accounted for by all the factors) were available

in The Measurement of Meaning for the 75 scales used in this study, all of the scales did not occur in any single factor analysis, and the factor loadings in the two studies in which they did occur are not exactly comparable. The subjects for this study are not exactly comparable to those used earlier, and the administrative procedures were somewhat different than those followed by Osgood, Suci, and Tannenbaum (1957). For these reasons, a standard form of the semantic differential, with single mark instructions, was administered in the same population of subjects, under similar conditions, approximately three weeks after the "best-worst" data were collected, and a new factor analysis was performed.

Since, for this analysis, equal Ns for all concepts was an important consideration, only 140 data forms were prepared and each data form was checked for missed pages of scales as it was turned in. It was, however, necessary to have six test booklets filled out by volunteers from the same population and to code an occasional "missed" scale "4" to obtain an N of 35 for every concept-scale item.

This control factor analysis was performed, due to the availability of computer programs, by the method of principal axes with varimax rotation. Although these methods

are not exactly the same as those employed by earlier researchers (which makes comparison between factor analyses difficult), according to Harman (1960), the principal axes solution is the rigorous mathematical solution to which the centroid method is a crude approximation, and the varimax rotation tends to provide mathematical precision for the intuitive notion of simple structure.

Although the task of making a comprehensive comparison between this factor analysis and the previous ones would be extremely difficult if not impossible, there is a very simple test of similarity that satisfies the needs of this study. That is a rank order correlation between the two sets of "evaluative dominance scores." It was reasoned that this correlation should be significantly large, and that such a finding would support the generality of this total study. But, in any case, the new factor analysis serves as a control, making it possible to compare the item by item analysis and the general factor analysis with data gathered from the same subjects under comparable conditions.

Experiment 2

Since it was felt that the results of the previous experiment would be more meaningful if it could be shown

that the evaluative discrimination capacity of a scale for a concept is related to the importance of the scale variable as an effector in evaluative decisions, it was decided to test the following hypothesis:

Hypothesis 2

There is a positive correlation between the ranks assigned to scales on the basis of the confidence in the scale's discrimination capacity (the absolute size of the sign test *z*) and ranks assigned by subjects instructed to rank the scales in order of importance to an evaluative decision about a given concept.

Rationale

Given that the sign test is, as applied in this study, a measure of agreement among subjects on the direction of the relation between a particular scale and the "best-worst" evaluative variable, and given that subjects are more likely to call important those variables on which there is high agreement, the hypothesis seems to follow.

Design

Concepts. Hypothesis 2 was tested for six concepts--the six common to the two studies from which the scales for this study were selected (Osgood et al., 1957, pp. 34 & 39).

The concepts are ME, SYMPHONY, AMERICA, MOTHER, BOULDER, and SIN. These concepts appear to be typical of the total set by the criteria of earlier experimenters. They also represent the "personal-impersonal" and the "well discriminated-poorly discriminated" dimensions of variability in this set of concepts which will be noted in the results of Experiment 1.

Subjects. Forty-eight subjects were selected (on the basis of availability) from a population similar to that sampled in the first experiment. The subjects were students at Kansas State University, Summer, 1963. Of these subjects, 19% were males and 70% were sophomores. None had participated in the discrimination experiment. Eight subjects were randomly assigned to each of the six concepts.

Administration. This experiment was administered in two class sessions. Each subject was given a mimeographed set of instructions (a sample appears in Appendix B), a concept card, and a set of 75 cards each with a pair of scale adjectives typed on it. The instructions directed the subject to sort the cards, ranking them in order of importance to an evaluative decision about the event named by his concept.

Analysis. For each of the six sets of eight rankings,

a Kendall Coefficient of Concordance was computed to determine if there was a significant relation among the rankings (Walker & Lev, 1953, pp. 283-386). Then, given a significant relation among rankings, the ranks of the sums of ranks (computed for the Kendall Coefficient) were assumed to be the best estimate of the true rank for each scale-concept (Walker & Lev, 1953, p. 286). Finally, a Spearman rank correlation was computed between these "importance ranks" and the discrimination capacity ranks to test, for each concept, the hypothesis that rho is equal to or less than zero.

Since the number of scales that had shown a significant discrimination capacity differed for each of these concepts, and, since there was no assurance that non-significant scales were in any sense rankable, it was decided in advance to perform the same analysis, as described above, considering only those scales which had shown a significant discrimination capacity for each of the concepts. The same data were used as in the complete analysis simply by disregarding the insignificant scales and assigning the new ranks accordingly.

Chapter 3

This chapter is a report of the results of two experiments.

In the preceding chapter, two experimental designs were presented. The first, a major experiment, tested the hypothesis that evaluative judgments are not independent of "non-evaluative" or objective judgments for particular events or categories of events. The second, a minor exploratory experiment, was intended merely to make the findings of the other more meaningful. It tested the hypothesis that the importance of a scale (dimension) to an evaluative decision about an event is not independent of the scale's evaluative discrimination capacity for that event.

Results of Experiment 1

The sign test was used to test, for each scale-concept item, the null hypothesis that the number of subjects who place "B" (the response to the best imaginable example of the concept) on the left side of "W" (the response to the worst imaginable example of the concept) is equal to the number who made the opposite choice. It was decided to reject this hypothesis at the 95% level of confidence, two tailed test.

Table 1 shows the results of the 1500 sign tests that

resulted from the combinations of 75 scales and 20 concepts. In Table 1 the scales are ordered from first to last by the number of concepts for which each discriminated "best" from "worst." The concepts are ordered from left to right by the number of scales that discriminated for each concept. Only significant values are indicated--those showing a preference for the adjective on the left by a plus and those showing a preference for the adjective on the right by a minus. Several of the scales have been reversed to simplify the reading of the table.

Since 20 sign tests were performed for each scale, each at the 95% level of confidence, reference was made to a table of binomial probabilities to determine the probability that N values out of 20 might be "significant" by chance. It was found that the probability of obtaining four or more significant values, when the probability of each one is .05, is less than .05. Thus, according to Table 1, all but three scales can be said to discriminate for some concept with greater than 95% confidence. Due to the arrangement of scales in Table 1, it is the last three scales that are of questionable value as evaluative discriminators for these concepts.

Table 1

Evaluative Discrimination Capacity and Polarity of 75 Scales for 20 Concepts

	j. LADY	m. MOTHER	i. GOD	f. FEATHER	t. TORNADO	
	a. AMERICA	e. FATHER	b. BABY	r. SWORD	g. FIRE	
	s. SYMPHONY	n. PATRIOT	d. COP	c. BOULDER	h. FRAUD	
	l. ME	k. LAKE	q. STATUE	p. SIN	o. RUSSIAN	

	j	a	s	l	m	e	n	k	i	b	d	q	f	r	c	p	t	g	h	o	
1. kind	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+		+		cruel
2. colorful	+	+	+	+	+	+	+	+		+	+	+	+	+	+		+	+	+		colorless
3. pleasant	+	+	+	+	+	+	+	+	+	+	+	+	+		+	+			+		unpleasant
4. clean	+	+	+	+	+	+	+	+	+	+	+	+	+	+				+	+		dirty
5. calm	+	+	+	+	+	+		+	+	+	+	+	+		+		+		+	+	agitated
6. honest	+	+	+	+	+	+	+	+	+	+	+	+	+	+		+				+	dishonest
7. bright	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+			+			dark
8. fragrant	+	+	+	+	+	+	+	+	+	+	+	+	+		+			+	+		foul
9. good	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+						bad
10. valuable	+	+	+	+	+	+	+	+	+	+		+	+	+	+			+			ugly
11. beautiful	+	+	+	+	+	+	+	+	+	+	+	+	+		+			+			awful
12. nice		+	+	+	+	+	+	+	+	+	+	+		+	+		-				weak
13. strong	-	+	+		-	+	+	+	+		+	+		+	+	-	-				small
14. large	+	+	+	+	+	+	+		+	+	+	+		+		+		+			cowardly
15. brave	+	+	+	+	+	+	+	+	+	+	+	+	+		+						profane
16. sacred	+	+	+	+	+	+	+	+	+	+		+	+		+						ferocious
17. peaceful	+	+	+	+	+	+	+	+	+	+	+	+							+		sad
18. happy	+	+	+	+	+	+	+	+	+	+	+	+	+								sick
19. healthy	+	+	+	+	+	+	+		+	+	+		+			-	-				constrained
20. free	+	+	+	+	+	+		+	+	+	+		+		+				+		tense
21. relaxed	+	+	+	+	+	+	+	+	+	+	+	+	+								stale
22. fresh	+	+	+	+	+	+	+	+	+	+	+	+		+							unfair
23. fair	+	+	+	+	+	+	+		+	+	+				-				-		passive
24. active	-	+	-		-		+	-		-	+	+	-	+	+						soft
25. hard																					

Table 1 (continued)

j. LADY				m. MOTHER				i. GOD				f. FEATHER				t. TORNADO				
a. AMERICA				c. FATHER				b. BABY				r. SWORD				g. FIRE				
s. SYMPHONY				n. PATRIOT				d. COP				c. BOULDER				h. FRAUD				
l. ME				k. LAKE				q. STATUE				p. SIN				o. RUSSIAN				
	j	a	s	l	m	e	n	k	i	b	d	q	f	r	c	p	t	g	h	o	
26. alive	+	+	+	+	+	+	+	+	+	+	+				-						dead
27. clear	+	+	+	+	+	+	+	+	+	+	+	+									hazy
28. sober	+	+	+	+	+	+	+		+	+	+	+							+		drunk
29. fast	+	+	+	+	+	+	+		+	+	+			+					+		slow
30. mature	+	+		+	+	+	+		+	-	+		+		+					+	youthful
31. full	+	+	+	+	+	+	+	+	+	+		+					-				empty
32. sweet	+	+	+	+	+	+	+	+	+	+		+	+								sour
33. tasty	+	+	+	+	+	+	+	+	+	+	+	+									distasteful
34. deep	+	+	+	+	+	+	+	+	+			+			+	-					shallow
35. rugged	-	+				+	+		+		+	+	-	+	+	-					delicate
36. tough	-	+	-	+		+	+		+		+		-	+	+						fragile
37. sharp	+	+	+	+	+	+	+		+	+	+			+							dull
38. smooth	+		+		+	-		+		+	-	+	+	+		+					rough
39. wide	-	+	+	-			+	+	+					-	+	-	-				narrow
40. objective	+	+	+	+	+	+	+	+			+	+								+	subjective
41. near	+	+			+	+	+	+	+						-		-		-	-	far
42. proud		+	+	+	+	+	+				+	+		+				+			humble
43. masculine	-	+			-	+	+		+	+	+			+		+					feminine
44. sharp	+	+	+	+	+		+						+	+	-			+			blunt
45. rich	+	+	+	+	+	+		+						+						-	poor
46. white	+			+	+	+	+	+	+	+			+								black
47. sweet	+	+	+	+	+	+		+	+	+											bitter
48. long								+			+	+	+	+	+	-	-		-		short
49. savory	+	+	+	+	+	+	+	+		+											tasteless
50. new	+	+	+	+						+			+	+	-						old

Table 1 (continued)

j. LADY				m. MOTHER			i. GOD			f. FEATHER			t. TORNADO							
a. AMERICA				e. FATHER			b. BABY			r. SWORD			g. FIRE							
s. SYMPHONY				n. PATRIOT			d. COP			c. BOULDER			h. FRAUD							
l. ME				k. LAKE			q. STATUE			p. SIN			o. RUSSIAN							
	j	a	s	l	m	e	n	k	i	b	d	q	f	r	c	p	t	g	h	o	
51. spacious		+	+	+				+	+			+					-	-			constricted
52. humorous	+			+	+	+			-	+			+	-							serious
53. soft	+				+		-	+	+	+			+	-							loud
54. light	+				+						-	-	+		-		+	+			heavy
55. curved	+				+		-	+			-	+		-	+						straight
56. complex	+	+	+	+					+			+		-							simple
57. ornate	+	+	+								-	+	+	+							plain
58. high				+		+		+	+			+			+		+				low
59. opaque		+	+	+				-				+	+	+							transparent
60. stable			-	-					+			+		+		-					changeable
61. aggressive			+	+		+	+				+	+									defensive
62. rounded	+		+							+				-	+		+				angular
63. young	+	+		+						+					-	+					old
64. thin	+	-	-							+					-		+				thick
65. bass	-		+		-	+	+				+										treble
66. calm			-			+		+	+		+										excitable
67. dry								-		+		+	+	+							wet
68. lenient	+			+				+								+					severe
69. unusual			+					+				+	+								usual
70. red								-		+			+					+			green
71. hot	+			+			+											+			cold
72. tenacious		+					+				+					-					yielding
73. rational		+			+		+														intuitive
74. blue		+						+					+								yellow
75. pungent																					bland

It can be said, then, that clear support has been obtained for the major theoretic hypothesis. That is, the hypothesis has been supported by 72 of 75 scales, including strong-weak and active-passive which are ranked 14.5 and 23, respectively, in number of concepts for which they discriminate, though they have been the lead scales in factors orthogonal to evaluation in repeated factor analyses.

Light-heavy (54), large-small (14), hard-soft (25), calm-excitable (66), and hot-cold (71), all of which have been identified with factors orthogonal to evaluation, also show an evaluative discrimination capacity for a significant proportion of the concepts.

In the rationale for the theoretic hypothesis, a flip-flop tendency, a change of polarity from concept to concept, was postulated as one means of accounting for the fact that certain scales for which evaluative discrimination capacity was predicted have not correlated well with the evaluative factor. In 37 of 75 scales (13, 14, 20, 24, 25, 30, 31, 34, 35, 36, 38, 39, 41, 43, 44, 45, 48, 50, 51, 52, 53, 54, 55, 56, 57, 59, 60, 62, 63, 64, 65, 66, 67, 70, and 72) this reversal of polarity occurs at least once, and in one case (hard-soft) the flip-flop is entirely sufficient (6 -'s and 6 +'s) to account for the apparent independence of

evaluation, although that scale discriminates for 12 of 20 concepts. Large-small, wide-narrow, and near-far, also show a significant number of discriminations in both directions.

Although the flip-flop tendency does not appear to be as pronounced as was anticipated, the evidence obtained here does show that an assumption of constant polarity is not universally tenable.

The factor analysis performed as a control in this study, using the same subjects, scales, and concepts as the other analysis gives no reason to think that these findings are in any way unusual. The rank correlation between evaluative dominance scale scores ($2e^2 - h^2$) from this analysis and those of Osgood et al. (1957) is .63 ($p < .05$). The patterns of factor loadings are only slightly different than those obtained by Osgood et al. ten years ago. By a principal axis analysis and varimax rotation, an eight factor solution was obtained that accounts for 47% of the variance. The factor accounting for the most variance is clearly evaluative (good-bad, nice-awful, fair-unfair, and kind-cruel are the highest loadings on this factor). The second factor in terms of variance accounted for is clearly the potency factor (rugged-delicate, tough-fragile, hard-soft, heavy-light, rough-smooth, and strong-weak are the

highest loaded scales on this factor). An activity factor, which accounts for about half as much variance as the potency factor, ranks sixth in order of importance (active-passive, dead-alive, aggressive-defensive, and fast-slow are the strongest scales on this factor). The third factor seems to be a secondary evaluative factor (clear-hazy, black-white, and happy-sad). The fourth factor, which accounts for about 5% of the variance, is different. Complex-simple is the highest loaded scale followed by spacious - constricted, savory-tasteless, unusual-usual, and rich-poor. Of the 12 scales which have their highest loadings on this factor, 10 were in the set of 25 non-evaluative scales taken from the thesaurus study. The fifth factor takes the largest loadings on empty-full, wide-narrow, thick-thin, and high-low and might be called a size factor except for the colorful-colorless and ornate-plain scales, which also have their highest loading on it. The seventh factor is identified by angular-rounded, curved-straight, dull-sharp, hot-cold, and bright-dark as a second activity or a sharpness factor. The eighth factor, which accounts for only 2.8% of the total variance, might (in desperation) be called the diaper factor, because the three scales which have their highest loadings on it are pungent-bland, wet-

dry, and changeable-stable. The complete set of factor loadings may be found in Table 2.

Although there are minor differences between this analysis and the earlier results by Osgood, they do not seem to be greater than would be expected from the changes in methodology. The evaluative, potency, and activity factors occur and account for the expected proportions of variance relative to each other (Osgood et al., 1957, pp. 72-73). The unexpected occurrence of the "complexity" factor is probably a result of a bias in selecting "non-evaluative" scales to add to the matrix, but it does point out a kind of scale-scale interaction operating on the factor structure.

A somewhat more objective comparison can also be made between these results and the earlier work. A separate factor analysis was performed of the fifty scales which had been used previously with the same concepts. A four factor solution was obtained using the principal axis and varimax methods. Then, the degree of factorial similarity (Harman, 1960, p. 257) was computed between each of these factors and its corresponding factor in the earlier analysis (Osgood, et al., 1957, p. 37). The degree of factorial similarity between the two "evaluative" factors is -.97 and between the

Table 2

A Principal Axis Factor Analysis with Varimax Rotation

Scales	1	2	3	4	5	6	7	8	h^2
1. angular-rounded	.06	.02	.13	-.04	-.17	-.05	.56	.10	.38
2. red-green	.24	-.04	.12	-.16	-.07	.22	.28	.19	.26
3. passive-active	.04	-.12	.02	-.08	.14	-.62	-.15	.23	.50
4. colorful-colorless	-.34	-.20	.01	.02	.44	.31	.20	.03	.49
5. worthless-valuable	.69	.06	.12	-.11	-.21	-.14	-.03	-.09	.58
6. ferocious-peaceful	.63	.14	.23	-.12	.05	.20	.28	-.05	.61
7. severe-lenient	.57	.18	.05	.03	.09	.29	.06	.10	.47
8. beautiful-ugly	-.67	-.24	.04	.02	.34	-.10	.09	-.01	.64
9. tough-fragile	.18	.73	.02	-.03	.16	.15	.13	-.06	.64
10. savory-tasteless	-.25	-.12	-.19	.51	.05	.07	-.04	.04	.39
11. changeable-stable	.24	-.23	.01	-.01	.00	.35	-.13	-.41	.42
12. wide-narrow	.00	.20	-.01	.01	.53	.07	-.28	.07	.41
13. sharp-blunt	.12	-.02	-.18	.41	.03	.19	.39	.16	.43
14. cowardly-brave	.47	-.27	.21	-.25	-.10	-.25	-.10	-.23	.54
15. hard-soft	.23	.69	.04	.13	.07	-.07	.16	-.06	.58
16. dull-sharp	.13	-.04	.08	-.17	-.19	-.24	-.48	-.04	.38
17. good-bad	-.82	-.02	-.04	-.12	.23	.01	.05	.07	.75
18. dirty-clean	.69	.16	.22	-.29	.06	-.02	-.04	-.10	.65
19. rational-intuitive	-.15	.20	.02	.00	-.14	.13	.04	.14	.12
20. tenacious-yielding	.25	.28	.08	-.29	.23	.25	.12	.15	.38
21. excitable-calm	.31	-.21	.29	-.05	.18	.38	.03	-.12	.41
22. fragrant-foul	-.44	-.26	-.25	.41	.00	-.01	-.15	.15	.53
23. curved-straight	.02	-.17	.10	.16	.06	.04	-.54	.02	.37
24. deep-shallow	-.03	.16	-.09	.42	.31	-.13	-.04	-.09	.34
25. wet-dry	.01	.01	-.01	-.05	.15	-.05	-.10	-.49	.27

Table 2 (Continued)

Scales	1	2	3	4	5	6	7	8	h^2
26. hot-cold	.01	.02	.14	.00	-.04	.23	.44	-.32	.37
27. ornate-plain	-.13	-.28	.14	.26	.35	.11	.21	.20	.40
28. small-large	-.14	-.48	.12	-.42	-.24	.01	.18	-.22	.58
29. honest-dishonest	-.77	.09	-.11	-.25	.18	.19	.07	.07	.75
30. poor-rich	.38	.11	.08	-.45	-.17	-.09	-.05	.04	.41
31. complex-simple	-.03	.05	-.03	.63	-.03	.13	-.08	-.11	.44
32. usual-unusual	-.01	-.04	-.12	-.46	.08	.10	-.16	.03	.27
33. healthy-sick	-.50	.04	-.39	.40	-.21	.22	-.07	-.04	.66
34. fast-slow	-.01	.23	-.05	.02	.37	.41	.21	.00	.40
35. dead-alive	.41	.11	.21	.00	-.06	-.57	.11	-.13	.58
36. masculine-feminine	.04	.53	-.12	-.08	.19	.03	.08	.19	.38
37. humorous-serious	-.36	-.32	-.08	-.37	.19	.05	-.01	-.06	.41
38. empty-full	.27	-.09	.20	-.07	-.60	.03	.00	.10	.50
39. thick-thin	-.04	.25	.07	-.13	.58	.04	-.23	-.03	.48
40. bass-treble	.09	.45	-.18	.27	-.14	-.14	.07	-.14	.38
41. yellow-blue	.20	.06	.31	-.22	-.20	.06	.20	-.04	.27
42. pleasant-unpleasant	-.72	-.10	-.27	.31	-.08	-.06	-.10	-.03	.72
43. high-low	-.25	.07	-.05	-.16	.47	-.04	.07	-.07	.33
44. opaque-transparent	.07	.12	-.07	.32	-.11	-.02	-.08	.13	.16
45. sacred-profane	-.65	-.07	-.07	-.13	.23	.03	.08	-.07	.51
46. mature-youthful	-.41	.46	-.03	-.11	.12	.07	.10	.15	.45
47. weak-strong	.26	-.58	.03	-.01	-.32	-.25	-.14	-.11	.60
48. constricted-spacious	.13	-.09	.26	-.55	.07	-.10	.05	.23	.46
49. pungent-bland	.00	.03	.02	.15	-.16	.08	.08	-.57	.39
50. rugged-delicate	.16	.73	-.06	.26	.04	.09	-.02	.01	.64
51. aggressive-defensive	.25	.25	-.14	.14	-.01	.43	.07	.14	.37

Table 2 (Continued)

Scales	1	2	3	4	5	6	7	8	h^2
52. soft-loud	-.36	-.39	-.08	-.06	-.10	-.27	.04	-.19	.42
53. proud-humble	-.18	.14	-.24	.37	.02	.30	-.02	.33	.44
54. nice-awful	-.81	-.14	-.01	.05	.01	-.11	.01	-.20	.73
55. objective-subjective	-.05	.13	-.56	.12	-.06	.26	-.08	.06	.42
56. unfair-fair	.79	-.06	-.03	-.11	-.01	-.11	-.01	-.08	.66
57. tasty-distasteful	-.44	-.08	-.39	.12	.09	-.13	.12	-.31	.50
58. sad-happy	.48	.11	.59	-.14	-.09	-.13	.07	-.01	.64
59. relaxed-tense	-.39	-.17	-.31	-.08	.06	-.17	.10	-.23	.38
60. rough-smooth	.25	.60	.25	-.04	-.03	.21	-.07	-.05	.54
61. new-old	-.07	-.39	-.24	-.02	.07	.19	.21	-.27	.38
62. sweet-sour	-.53	-.26	-.25	.12	.10	-.03	.07	-.18	.48
63. old-young	.06	.47	.33	.01	-.12	-.29	-.07	.01	.44
64. calm-agitated	-.47	-.08	-.47	.04	.00	-.28	-.14	.12	.56
65. bitter-sweet	.53	.30	.20	-.16	-.15	-.05	.04	-.06	.47
66. long-short	.14	.09	-.34	.26	.32	-.12	.22	.25	.44
67. black-white	.31	.06	.67	-.08	.03	-.12	-.02	-.04	.57
68. clear-hazy	-.22	-.05	-.70	.09	.11	-.12	.05	-.04	.58
69. constrained-free	.29	-.06	.29	-.30	-.06	-.21	-.16	.03	.34
70. sober-drunk	-.51	.10	-.33	-.04	-.10	.16	-.04	-.03	.42
71. bright-dark	-.38	-.20	-.29	.10	.17	-.05	.39	-.09	.46
72. stale-fresh	.40	.18	.58	-.17	-.08	-.06	-.08	.08	.58
73. kind-cruel	-.78	-.04	-.27	.06	-.07	-.03	-.02	.03	.69
74. heavy-light	.14	.60	.24	-.01	.07	-.12	-.20	-.14	.53
75. far-near	.10	.03	.20	-.14	-.13	-.24	.00	.20	.19
Per cent total variance	14.55	7.37	5.95	5.18	4.07	3.98	3.16	2.81	

two "potency" factors .91. The coefficient between the two "activity" factors is .27 and between the two "fourth" factors .31. Table 3 shows the rotated factor loadings for the four factors obtained in this study. The scales and factors are arranged for easy comparison with the corresponding set of loadings in The Measurement of Meaning (p. 37).

There is no test of significance of this measure of factor similarity, but it is evident that the agreement between the two analyses on the first two factors is quite high, and that the agreement on the other two factors is somewhat less satisfactory. This seems, however, to be adequate support for the similarity of the two sets of subjects.

There are three indices of scale evaluation capacity reported in Table 4, which summarize the results of this experiment. Column I is simply the proportion of the 20 concepts for which each scale shows an evaluative discrimination capacity in the "best-worst" analysis. Column II shows the evaluative dominance scores (twice the variance accounted for by the evaluative factor minus the total variance accounted for) computed from the factor analyses by Osgood, Suci, and Tennenbaum. Column III shows the same evaluative dominance scores based on the new factor analysis.

Table 3

A Principal Axis Factor Analysis with Varimax Rotation

	Scales	1	2	3	4	h^2
1.	good-bad	-.81	-.05	-.19	-.02	.70
2.	large-small	.09	.71	.24	-.17	.60
3.	beautiful-ugly	-.69	-.12	.15	-.18	.54
4.	yellow-blue	.32	-.16	-.22	.35	.30
5.	hard-soft	.21	.70	-.01	.05	.54
6.	sweet-sour	-.68	-.14	.22	-.08	.54
7.	strong-weak	-.26	.66	.15	.21	.57
8.	clean-dirty	-.75	-.06	.21	-.03	.61
9.	high-low	-.35	.24	-.10	-.13	.21
10.	calm-agitated	-.59	-.11	.01	-.34	.47
11.	tasty-distasteful	-.62	.10	.26	-.19	.50
12.	valuable-worthless	-.74	.03	.02	-.02	.55
13.	red-green	.24	-.08	.09	.50	.32
14.	old-young	-.28	-.32	.01	.29	.27
15.	kind-cruel	-.77	-.13	-.39	.03	.76
16.	loud-soft	.37	.41	-.10	.23	.37
17.	deep-shallow	-.15	.39	.00	-.16	.20
18.	pleasant-unpleasant	-.76	.00	.27	-.18	.69
19.	black-white	.56	-.04	-.22	.02	.36
20.	bitter-sweet	.62	.06	-.25	.18	.48
21.	happy-sad	-.76	-.08	-.09	.06	.60
22.	sharp-dull	-.33	.20	.19	.38	.33
23.	empty-full	.43	-.22	.08	.02	.24
24.	ferocious-peaceful	.64	.16	-.15	.37	.59
25.	heavy-light	.19	.40	-.73	.00	.72
26.	wet-dry	-.05	.03	-.03	-.28	.08
27.	sacred-profane	-.65	.01	.30	-.05	.52
28.	relaxed-tense	-.53	-.23	-.17	.00	.36
29.	brave-cowardly	-.55	.41	.04	.24	.53
30.	long-short	-.05	.37	.48	-.10	.38
31.	rich-poor	-.48	.11	.21	.01	.29
32.	clear-hazy	-.56	.03	.01	-.17	.34
33.	hot-cold	-.02	.00	-.13	.64	.43
34.	thick-thin	-.10	.35	-.52	-.09	.41
35.	nice-awful	-.83	-.10	.09	-.15	.73
36.	bright-dark	-.59	-.07	.25	.15	.43
37.	bass-treble	.01	.49	.09	-.11	.26
38.	angular-rounded	.07	-.07	.03	.44	.21
39.	fragrant-foul	-.62	-.21	-.10	-.04	.44
40.	honest-dishonest	-.71	.17	.27	.05	.61
41.	active-passive	-.07	.23	.51	.38	.46
42.	rough-smooth	.36	.58	-.14	.17	.51
43.	fresh-stale	-.57	.05	.58	-.12	.67
44.	fast-slow	-.11	.35	.15	.47	.38
45.	fair-unfair	-.81	.06	-.12	.08	.67
46.	rugged-delicate	.12	.77	-.08	.04	.62
47.	far-near	.23	-.16	-.64	.01	.49
48.	pungent-bland	.07	-.08	-.15	.24	.09
49.	healthy-sick	-.65	.22	.29	.03	.55
50.	wide-narrow	-.08	.40	-.08	-.29	.26
Per cent total variance		24.69	8.96	6.68	5.19	

Table 4

Three Indices of Scale Evaluation Capacity

	Scales	I	II	III
1.	kind-cruel	.90	+.61	+.52
2.	colorful-colorless	.85	-.10	-.25
3.	pleasant-unpleasant	.80	+.57	+.32
4.	clean-dirty	.80	+.66	+.30
5.	calm-agitated	.80	+.24	-.12
6.	honest-dishonest	.80	+.70	+.43
7.	bright-dark	.80	+.39	-.18
8.	fragrant-foul	.80	+.69	-.15
9.	good-bad	.75	+.78	+.59
10.	valuable-worthless	.75	+.61	+.37
11.	beautiful-ugly	.75	+.66	+.25
12.	nice-awful	.75	+.69	+.58
13.	strong-weak	.70	-.39	-.46
14.	large-small	.70	-.50	-.54
15.	brave-cowardly	.70	+.23	-.10
16.	sacred-profane	.70	+.54	-.33
17.	peaceful-ferocious	.65	+.28	+.18
18.	happy-sad	.65	+.57	-.18
19.	healthy-sick	.65	+.37	-.16
20.	free-constructed	.65	-.05	-.16
21.	relaxed-tense	.65	+.14	-.08
22.	fresh-stale	.65	+.40	-.25
23.	fair-unfair	.65	+.67	+.59
24.	active-passive	.60	-.33	-.50
25.	hard-soft	.60	-.33	-.48
26.	alive-dead	.60	-.94	-.24
27.	clear-hazy	.60	+.32	-.48
28.	sober-drunk	.60	-.68	+.10
29.	fast-slow	.60	-.50	-.40
30.	mature-youthful	.60	-.11	-.10
31.	full-empty	.60	+.23	-.35
32.	sweet-sour	.60	+.66	+.09
33.	tasty-distasteful	.60	+.38	-.11
34.	deep-shallow	.60	-.22	-.34
35.	rugged-delicate	.55	-.33	-.59
36.	tough-fragile	.55	-.88	-.57
37.	sharp-dull	.55	-.23	-.35
38.	smooth-rough	.55	-.02	-.41
39.	wide-narrow	.55	-.11	-.41

Table 4 (continued)

	Scales	I	II	III
40.	objective-subjective	.55	-.02	-.42
41.	near-far	.55	+.14	-.17
42.	proud-humble	.50	-.04	-.37
43.	masculine-feminine	.50	-.23	-.38
44.	sharp-blunt	.50	-.16	-.40
45.	rich-poor	.50	+.32	-.27
46.	white-black	.45	+.31	-.38
47.	sweet-bitter	.45	+.59	+.10
48.	long-short	.45	-.15	-.40
49.	savory-tasteless	.45	-.89	-.26
50.	new-old	.40	-.92	-.37
51.	spacious-constricted	.40	-.07	-.43
52.	humorous-serious	.40	-.07	-.16
53.	soft-loud	.40	-.15	-.16
54.	light-heavy	.40	-.27	-.19
55.	curved-straight	.40	-.11	-.36
56.	complex-simple	.35	-.06	-.44
57.	ornate-plain	.35	-.08	-.37
58.	high-low	.35	+.30	-.20
59.	opaque-transparent	.35	-.06	-.15
60.	stable-changeable	.30	-.05	-.31
61.	aggressive-defensive	.30	-1.00	-.25
62.	rounded-angular	.30	-.17	-.38
63.	young-old	.30	-.10	-.43
64.	thin-thick	.30	-.20	-.48
65.	bass-treble	.30	-.11	-.36
66.	calm-excitable	.25	-.08	-.22
67.	dry-wet	.25	-.02	-.27
68.	lenient-severe	.20	-.15	+.19
69.	unusual-usual	.20	-.08	-.27
70.	red-green	.20	-.06	-.15
71.	hot-cold	.20	-.22	-.37
72.	tenacious-yielding	.20	-.13	-.25
73.	rational-intuitive	.15	-.04	-.08
74.	blue-yellow	.15	+.05	-.20
75.	pungent-bland	.00	-.33	-.38

Rank correlation between I & II = .46 ($p < .05$)
Rank correlation between I & III = .33 ($p < .05$)
Rank correlation between II & III = .63 ($p < .05$)

Rank correlations were computed between all pairs of these three indices of scale evaluation capacity. As indicated before, the correlation of .63 between indices II and III is an indication of the similarity between two (actually three) factor analyses on this criterion. This result tends to support the idea that college students at Kansas State University in 1963 are not too different from college students at The University of Illinois in the mid 1950's. In other words it supports the comparability of the SD technique.

The correlations between I and II (.46) and between I and III (.33) indicate a tendency for scales that are evaluatively dominant (+ signs in columns II and III) to discriminate between "best" and "worst" for a larger proportion of the concepts (column I of Table 4). However, the contrast in the two techniques is also noticeable.

Of the 46 scales identified in column II of Table 4 as predominantly non-evaluative, 44 show an evaluative discrimination capacity for a significant proportion of the concepts. This ratio is 59/62 in the new factor analysis (column III). It is also true that a significant evaluative discrimination capacity has not been demonstrated for any scale for all concepts. Even the <u>good</u>-<u>bad</u> scale does not appear to differentiate among TORNADOS, RUSSIANS, FIRES, SINS, or FRAUDS.

The results of experiment 1, then, indicate that factor analysis of these scales across a set of concepts seriously underestimates the predictive power of "non-evaluative" scales for particular concepts and probably overestimates the predictive power of "evaluative" scales for certain concepts. The nature of the concept-scale interaction noted in earlier research with the semantic differential is now reasonably explicit, and it is clear that ignoring this interaction may lead to serious misinterpretation of SD results.

There is a "postscript" result to this study. On seeing the results of the "best-worst" analysis (Table 1), the writer noticed that the most frequently used scales, representing all the major factors, seemed to discriminate for more concepts than the less frequently used ones. This suggested the "hypothesis" that all of the regularly obtained factors are "evaluative." That is, the factors obtained in an SD analysis might be viewed as "modes of evaluation" (see Osgood et al., 1957, p. 62) rather than "evaluative" and "non-evaluative" dimensions of "meaning." When the new factor analysis was completed, a rank order correlation was computed between the proportion of the 20 concepts discriminated for by each scale and the communality (the

proportion of the total variance accounted for by all the factors) for each scale. The coefficient was .70, higher than any of the other interrelations in this set of data. If this result is not coincidence, it provides a basis for reinterpreting certain factor analytic results already obtained.

Results of Experiment 2

In the second experiment, eight subjects (for each of six concepts) ranked the 75 scales in order of importance to an evaluative decision about a concept. The six concepts which had occurred in both of the earlier studies from which the scales were drawn (Osgood et al., 1957, pp. 34 & 49) were chosen to represent the total set for this ranking task.

The Kendall coefficient of concordance (W) was computed for each set of eight rankings. As can be seen in Table 5, there was a significant agreement among the eight rankings in each set, since the six Ws are all significant beyond the .01 level.

Given significant Ws, the ranks of the sums of ranks were taken as the best estimate of the true importance ranks, for each concept. A second set of rankings was obtained from the results of Experiment 1 by ranking the

scales on the estimated confidence in their discrimination capacity (sign test z values) for each concept. Table 5 (column R_t) shows that there is a significant correlation ($p < .01$) between these two measures for all six concepts.

Table 5

Relations Between Importance and Discrimination Capacity

Concepts	W	R_t	R_d
MOTHER	.59	.65	.82
BOULDER	.34	.50	.52
ME	.64	.61	.56
AMERICA	.66	.59	.71
SYMPHONY	.59	.43	.44
SIN	.45	.36	.35*

*Probability greater than .05 all others less than .01.

The same correlations were calculated for just those scales which had shown a significant discrimination capacity for each of the concepts. Column R_d of Table 5 shows that five of the six correlations for discriminating scales are significant ($p < .01$). Two are higher and four do not differ from those calculated for the total set of scales (column R_t). The only insignificant one (SIN) was based on an N of sixteen.

It was decided to advance to perform this second analysis on the discriminating scales only, because there was no rationale available to predict that subject's dispositions of non-discriminating scales would not be random. The data show, however, that subjects do agree on not-important scales.

The size of these correlations, which probably approach a limit set by the reliability of the two measures, indicates that both the "best-worst" technique and the importance ranking technique are measuring, to a large extent, the same variable. This means that one can select scales for a particular measuring task by using either the group technique (best-worst) or the other, which is essentially an individual technique.

Summary

In the two experiments reported here, strong support has been obtained for both theoretic hypotheses. In Experiment 1, 72 of 75 scales were shown to have an evaluative discrimination capacity for a significant proportion of the 20 concepts. Among those scales showing an evaluative discrimination capacity were strong-weak, large-small, active-passive, and hot-cold, all of which have been consistently identified with factors said to be independent of evaluation.

These results are believed to clarify the concept-scale interaction that has been noted in research with the semantic differential.

In obtaining these results, a technique has been demonstrated which has practical value as a means of selecting scales for semantic differential analysis of specific concepts and, in many applications, can be substituted for the more complex factor analytic technique. Perhaps more important, this technique does not require the same assumptions as the SD technique and, therefore, provides a means of testing those assumptions in specific content areas.

In experiment 2, it was found that evaluative discrimination, by a scale for a concept, is related to the importance of that scale as a criterion of evaluation for that concept. This result suggests that an individual ranking technique may be used for the selection of evaluative discriminators and indirectly supports the validity of the "best-worst" method.

Chapter 4

In this chapter, the background of this research, two experiments, and their results are reviewed; the writer's conclusions drawn from those results are summarized; then, further implications are discussed for (a) the semantic differential, (b) meaning, and (c) new directions in research.

Background

This research developed out of the semantic differential, a technique called a measure of meaning. The SD has been frequently used as a research tool since its introduction in 1952, and has been applied in a wide variety of research studies. However, the SD was not adaptable to answering the question, what are the objective criteria on which people base their evaluations of events? On the contrary, results with the instrument seemed to indicate that evaluations are independent of the objective attributes of events.

Unable to accept this conclusion, the writer undertook a reassessment of the SD technique, hoping to find another way of interpreting the empirical results that would not conflict with what seemed a reasonable question. The results of that reassessment are reported in Chapter 1 of this thesis. In brief, it was found that the apparent independence

of evaluative judgments and such objective attributes as "activity" and "potency" can be reasonably explained as an artifact of the statistical method that produced this result. To put it another way, the SD technique assumes that any relation which does obtain between these "dimensions" of judgment is linear and constant across concepts. Applications of the SD under conditions in which either of these assumptions is untenable would, then, quite predictably result in a failure to reject the null hypothesis of independence among the dimensions of judgment. To state the conclusion still another way, the practice of summing across concepts obscures any differences that may obtain among the concepts. And, to the extent that there are differences, the final result is not necessarily descriptive of any concept in the set. Conclusions based on these results have, however, been assumed to apply to every concept in the set, on the assumption that no such differences obtain. Going even further, with an assumption of representative sampling, the conclusions have been generalized to concepts outside the sample. Prior to this research there was extensive evidence of concept-scale interaction available--evidence that the basic assumption is not tenable.

In spite of the fact that the interpretation of SD results seemed to be in error, the instrument itself--seven-interval scales bounded by polar adjectives--appeared to be a highly reliable and efficient means of obtaining judgments from subjects about events or categories of events. So it was decided to take this as a starting point for developing a method of finding the objective criteria on which evaluative judgments are based.

This study, then, was designed to provide a new basis for interpreting existing SD data and to provide a foundation for further research that could take advantage of the SD scaling technique. In line with this purpose, the scales and concepts used in this study were selected from earlier research so that it would be possible to make direct comparisons between the two methods of analysis.

Experiment 1

On the assumption that bipolar adjectival scales like those used in the semantic differential do measure evaluative and objective variables, and that evaluations are to some extent influenced by the objective attributes of events, the following hypothesis was tested.

Hypothesis 1. Bipolar adjectival scales, such as those used in the semantic differential and including those

identified in factor analysis as "non-evaluative," have an evaluative discrimination capacity for some concepts.

This hypothesis was tested for each of 75 scales by testing the null hypothesis, for each concept, that the number of Ss who indicated one polarity for the scale equals the number of Ss who indicated the opposite polarity (alpha = .05) and, then, testing a second null hypothesis that the first was not rejected for a significant proportion of the 20 concepts (alpha = .05).

Experiment 2

On the assumptions that the sign test, as applied here to subjects' responses to the "best" and "worst" examples of a concept, measures the level of agreement among subjects on the polarity of a scale for a concept and that the probability of agreement is greater on important criteria, a second hypothesis was tested.

Hypothesis 2. There is a positive relation between the confidence in a scale's evaluative discrimination capacity and its importance as an evaluative criterion.

The test of this hypothesis was based on the rank correlation between the absolute size of the sign test z values and importance ranks assigned by subjects.

Results

The results of the first experiment strongly support hypothesis 1. Of the 75 scales, 72 show an evaluative discrimination capacity for a significant proportion of the concepts. Of the 46 scales that were identified by earlier factor analysis as predominantly non-evaluative, 44 showed an evaluative discrimination capacity for a significant proportion of the 20 concepts. Only one scale (pungent-bland) did not show a single significant discrimination for these concepts. No scale (including good-bad) showed a significant discrimination capacity for all of the concepts. The number of scales discriminating for a particular concept ranged from 59 to 6, and the number of concepts differentiated by a particular scale ranged from 0 to 18.

Factor analyses performed as a check on the subjects and methods of data collection employed in this study showed the expected amount of agreement with earlier work. In the 75 scale analysis, evaluative, potency, and activity factors appeared in the right order of variance accounted for, and the correlation between evaluative scores computed from the new data and earlier factor studies was .63 ($p < .01$). A separate analysis of the 50 scales previously employed with these same concepts was performed, and computed indices of

factor similarity showed high agreement between this and previous data on the first two factors (.97 and .91 respectively). The less satisfactory results with the third and fourth factors (.27 and .31) are attributed to differences in the techniques of analysis.

It was noted that the frequently used scales, that consistently load high on some factor but not necessarily the evaluative factor, seemed to discriminate for more concepts than the less frequently used scales. It was deduced that the proportion of the set of concepts for which a scale discriminates should, then, be related to the communality of the scale (the proportion of total variance accounted for on that scale in factor analysis). A rank order correlation was computed between these two variables for the 75 scales, and the coefficient obtained was .70, significant well beyond chance expectations. It is appreciably larger than the correlation of .33 obtained between the first variable and the evaluative dominance of the scales calculated from the same data. Viewed as a descriptive statistic, it summarizes the relation between the two techniques of analysis.

In the second experiment, strong support was obtained for hypothesis 2. Over 75 scales, for six concepts, the

correlations between discrimination capacity and importance were all greater than .36 and significant beyond the 99% level of confidence.

Conclusions

1. The evaluative judgments that people make about events are related to their "objective" judgments of those events.

2. The objective criteria on which people base their evaluations of particular events are discoverable, using the "best-worst" technique.

3. The greater the evaluative discrimination capacity of a given scale the more likely it is to be an important criterion of evaluation.

4. The fact that a particular scale discriminates evaluatively (or does not) for a particular concept cannot be generalized to other, unrelated concepts.

5. Using the same scales and concepts, the replication of Osgood's factor analysis produced a similar factor structure.

Implications for the Semantic Differential

Prior data, plus the evidence of concept-scale interaction obtained in this study, is sufficient to reject the

assumption that such interaction does not occur and to invalidate inferences which employ that assumption. This statement is chiefly concerned with the acceptance of the independence of factors, but it applies as well to all aspects of a factor analysis based on scores that have been summed over a set of unrelated concepts.

As pointed out earlier (Chapter 1), given evidence of concept-scale interaction, the factor structure obtained from a set of concepts does not necessarily describe any concept in the set. By the same token, results obtained from a set of concepts individually do not necessarily describe the set as a whole. That is, the inclusion or exclusion of the variance among concepts has an unknown effect on the obtained factor structure which makes it impossible to generalize from either the specific or the general factor analysis to the other.

If, on the other hand, factor analysis of SD data were performed on single concepts, the problem of concept-scale interaction would be eliminated, but there would then be the problem of low variance and the indeterminate correlations that result from it. It would also still be necessary to consider the fact that SD data may not meet the interval data assumption of the product moment

correlation, and the evidence that the linear assumption is not universally tenable. The problems of polarity and relevance would not interfere with the single concept factor analysis, but the information provided by such an analysis, over and above that provided by the simpler "best-worst" technique, is probably not worth the additional effort required to obtain it.

In some cases, factor analysis might reasonably be employed with the "best-worst" method to assist in the development of categories of evaluative criteria for particular concepts or categories. Even in this restricted application, however, the results must be interpreted with extreme caution.

Implications for Meaning

The most important implication of this study for <u>meaning</u> is that there may not be any non-evaluative dimensions of meaning. This study has provided support for the contention that a scale is either evaluative or irrelevant to a particular concept. Nearly all the scales have shown an evaluative discrimination capacity for some concept, and with a larger sample of concepts, this result would probably have been unanimous.

The idea that we attend to those attributes of events which have a sign relation to some form of reinforcement, in some context, seems to fit better with commonly accepted theories of human behavior than the idea that we attend to some, but not all, attributes which are independent of evaluation. That is, the finding that all the dimensions are evaluative provides a closer tie with learning theory, particularly the reward principle, than the "tenuous assumption" on which interpretation of the semantic differential previously depended (see Osgood et al., 1957, p. 27).

These results do not imply that the SD does not measure meaning. For to infer, from, the observation that all the "dimensions" appear to be evaluative, that the SD is an "attitude" measure and, therefore, not a measure of meaning, is to commit the same fallacy of two-valued logic that led to "non-evaluative" factors in the first place.

To argue over whether the SD technique and the "best-worst" technique measures <u>meaning</u> or <u>attitude</u> does not seem to be a useful expenditure of energy, for whether or not the only "dimension of meaning" is evaluative, it certainly is reasonable to say that the ability to discriminate evaluatively among members of a category and the ability to discriminate evaluatively among categories may <u>both</u> be

considered evidence that the categories are meaningful to the discriminator. It does seem desirable to be able to separate these two abilities and to explore them separately in that they appear to index two quite different levels of meaningfulness. The "best-worst" technique seems admirably adapted to this task.

Implications for New Directions in Research

It has already been suggested that the technique developed here has application in consumer market research. Image research has been done with the semantic differential by comparing the "ideal" of a particular product with some specific brand of the product. This approach did not, however, include an objective method for determining scale polarity (which this study has shown is a variable) nor show the evaluative significance of variables that could be manipulated in the product. The "best-worst" technique, on the other hand, is a method for discovering the dimensions that have an evaluative discrimination capacity for a specific kind of product (whether it's automobiles or toothpaste), it establishes the polarity of the scales for that product, and by anchoring both ends of the scale, gives an idea of how much difference makes a difference.

Most importantly, it shows the evaluative significance of objective variables. A profile of a product on scales selected by this technique should be directly interpretable as suggestions for modification in the product and the advertising of the product.

The image research idea, though, is only a specific example of the kind of research to which the B-W technique is appropriate. As mentioned earlier, this instrument was designed for (and seems to be suited to) asking why one event or object is preferred to another of its kind, in regard to any category of events whatsoever. Broad application of this instrument--perhaps in conjunction with the standard SD--might lead, eventually, to a description of the system of values which controls the majority of human behavior.

At this point, it seems reasonable to assume that there are both general and specific values involved in the decisions that people make, and knowledge gained about any of those values would increase the overall predictability of human behavior. In other words, the behaviors of people are probably controlled by their evaluative judgments, evaluative judgments are probably related to the objective attributes of events, and the B-W technique is a crude but

usable technique for extending our knowledge of that intricate network of relationships.

Perhaps the most exciting area of research in which the B-W technique may be applied is the general area of persuasion. Is a speaker perceived as more knowledgeable when he restricts his arguments to dimensions that are "relevant?" Is a persuader more effective when he presents "factual information" that fits the audience's predetermined value system or when he presents and interprets "facts" they didn't know were supposed to have an effect on their judgment of value? Is an advertisement less effective if it bases a claim of value on a dimension which does not have an evaluative discrimination capacity for a given audience and a given product? Why does an argument that works with one audience fail with another--is it because they have different criteria of evaluation?

All of these questions have been asked before, and their answers seem obvious, but the technique presented here offers innumerable possibilities for refining both the questions and the answers.

The advertiser or any other persuader certainly has reason to want to accomplish his purpose with as little

effort and expense are possible. It seems very likely that, by determining in advance exactly what variables the members of his audience perceive as relevant to the decision he requires of them, he can not only decrease his cost but increase his effect as well.

The results of this study and the viewpoint evolved in this discussion have some very important implications for the measurement of attitudes per se. Kerrick and McMillan (1961) found that subjects responded differently to news stories on SD scales when they were informed that they were engaged in attitude research. The following quotation is from the summary of their report.

> The informed group showed much less tendency to change their attitudes in response to the news stories. In addition, when members of the informed group did show change in response to the stories, they were more likely to change in the direction opposite to that advocated in the stories than were members of the naive group.
>
> The naive group's attitude change was predictable from the principle of pressure toward congruity. The informed group's was not.
>
> Instructions inhibited only evaluative change; non-evaluative change in response to the communication was no different for informed and naive groups.

This suggests that if subjects are told or guess that the study in which they are participating is an "attitude study," a more accurate picture of the effect of an independent variable can be obtained by discarding the

"evaluative" scales and basing one's conclusions on the "non-evaluative" scales that are used for "masking" purposes.

Summary

In short, there are two important outcomes of this research project: (a) The results of research that has been done with the semantic differential must be re-evaluated, because at least one basic assumption in that technique is untenable, and results have been obtained here that are incompatible with inferences based on that assumption. (b) An alternate method of analysis for the SD has been demonstrated which does not require as many tenuous assumptions as the factor analytic method, can be more parsimoniously interpreted, can be applied in a wide range of research designs, and is specifically suited to the task of finding out why people prefer one example of a category to another example of the same category.

Bibliography

Baxter, J. C. Mediated generalization as a function of semantic differential performance. Dissertation Abstr., 1959 (Nov.), 20, 1957. (Abstract)

Berlo, D. K. The process of communication. New York: Holt, Rinehart, & Winston, 1960.

Berlo, D. K., & Gulley, H. E. Some determinants of the effect of oral communication in producing attitude change and learning. Speech Monogr., 1957, 24, 10-20.

Bettinghaus, E. P. The application of the principle of congruity to certain aspects of language behavior. East Lansing: Communication Research Center, Michigan State University, 1961. (a)

Bettinghaus, E. P. The operation of congruity in an oral communication situation. Speech Monogr., 1961, 28, 131-142. (b)

Block, J. An unprofitable application of the semantic differential. J. consult. Psychol., 1958, 22, 235-236.

Brown, R. W. Review of: "Osgood, Suci, & Tannenbaum, The measurement of meaning." Contemp. Psychol., 1958, 3, 113-115.

Brown, R. W. Words and things. Glencoe, Ill.: The Free Press, 1959.

Carroll, J. B. Review of: "Osgood, Suci, & Tannenbaum, The measurement of meaning." Language, 1959, 35, 58-77.

Church, J. Language and the discovery of reality. New York: Random House, 1961.

Dicken, C. F. Connotative meaning as a determinant of stimulus generalization. Psychol. Monogr., 1961, 75, No. 1 (Whole No. 505).

Donahoe, J. W. Changes in meaning as a function of age. J. genet. Psychol., 1961, 99, 23-28.

Eisdorfer, C., & Altrocchi, J. A comparison of attitudes toward old age and mental illness. J. Gerontol., 1961, 16, 340-343.

Endler, N. S. Changes in meaning during psychotherapy as measured by the semantic differential. J. counsel. Psychol., 1961, 8, 105-111.

Ferguson, G. A. Statistical analysis in psychology and education. New York: McGraw-Hill, 1959.

Flavell, J. H. Meaning and meaning similarity: I. A theoretical reassessment. J. gen. Psychol., 1961, 307-319. (a)

Flavell, J. H. Meaning and meaning similarity: II. The semantic differential and co-occurrence as predictors of judged similarity in meaning. J. gen. Psychol., 1961, 64, 321-335. (b)

Greenburg, B. S., & Tannenbaum, P. H. The effect of bylines on attitude change. Journ. Quart., 1961, 38, 535-537.

Grigg, A. E. A validity study of semantic differential technique. J. clin. Psychol., 1959, 15, 179-181. (a)

Grigg, A. E. A validity test of self-ideal discrepency. J. clin. Psychol., 1959, 15, 311-313. (b)

Gulliksen, H. Review of: "Osgood, Suci, & Tannenbaum, The measurement of meaning." Contemp. Psychol., 1958, 3, 115-119.

Harman, H. H. Modern factor analysis. Chicago: University of Chicago Press, 1960.

Jenkins, J. J., Russell, W. A., & Suci, G. J. An atlas of semantic profiles for 360 words. Amer. J. Psychol., 1958, 71, 688-699.

Jenkins, J. J., Russell, W. A., & Suci, G. J. A table of distances for the semantic atlas. Amer. J. Psychol., 1959, 72, 623-625.

Kelly, Jane A., & Levy, L. H. The discriminability of concepts differentiated by means of the semantic differential. Educ. psychol. Measmt., 1961, 21, 53-58.

Kerrick, Jean S., & McMillan, D. A., III. The effects of instructional set on the measurement of attitude change through communications. J. soc. Psychol., 1961, 53, 113-120.

Kjeldergaard, P. M. Attitudes toward newscasters as measured by the semantic differential. J. appl. Psychol., 1961, 45, 35-40.

Kraus, S. Modifying prejudice: Attitude change as a function of race of the communicator. AV comm. Rev., 1962, 10, 14-22.

Kumata, H. A factor analytic investigation of the generality of semantic structure across two selected cultures. Unpublished doctoral dissertation, Univer. of Illinois, 1957.

Lambert, W. E., & Jakobovits, L. A. Verbal satiation and changes in the intensity of meaning. J. exp. Psychol., 1960, 60, 376-383.

Lyle, J. Semantic differential scales for newspaper research. Journ. Quart., 1960, 37, 559-562, 646.

McNelly, J. T. Meaning intensity as related to readership of foreign news. Unpublished doctoral dissertation, Michigan State Univer., 1961.

Manis, M. Assessing communication with the semantic differential. Amer. J. Psychol., 1959, 72, 111-113.

Messick, S. J. Metric properties of the semantic differential. Educ. psychol. Measmt., 1957, 17, 200-206.

Messick, S. J., & Solley, C. M. Word-association and semantic differentiation. Amer. J. Psychol., 1957, 70, 586-593.

Michon, J. A. An application of Osgood's "semantic differential" technique. Acta psychol. Amst., 1960, 17, 377-391.

Mindak, W. A. Fitting the semantic differential to the marketing problem. J. Market., 1961, 25 (April), 28-33.

Mitsos, S. B. Personal constructs and the semantic differential. J. abnorm. soc. Psychol., 1961, 12, 433-434.

Moss, C. S. Current and projected status of semantic differential research. Psychol. Rec., 1960, 10, 47-54.

Moss, C. S. Experimental paradigms for the hypnotic investigation of dream symbolism. Int. J. clin. exp. Hypnosis, 1961, 9, 105-117.

Moss, C. S., & Waters, T. J. Intensive longitudinal investigation of anxiety in hospitalized juvenile patients. Psychol. Rep., 1960, 60, 262-265.

Mogar, R. E. Three versions of the F scale and performance on the semantic differential. J. abnorm. soc. Psychol., 1960, 60, 262-265.

Nebergall, R. E. An experimental investigation of rhetorical clarity. Speech Monogr.

Norman, W. T. Stability-characteristics of the semantic differential. Amer. J. Psychol., 1959, 72, 581-584.

Osgood, C. E. Method and theory of experimental psychology. New York: Oxford Univer. Press, 1953.

Osgood, C. E., and Luria, Zella. A blind analysis of a case of multiple personality using the semantic differential. J. abnorm. soc. Psychol., 1954, 49, 579-591.

Osgood, C. E., & Suci, G. J. A measure of relation determined by both mean difference and profile information. Psychol. Bull., 1952, 49, 251-262.

Osgood, C. E., & Suci, G. J. Factor analysis of meaning. J. exp. Psychol., 1955, 50, 325-338.

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. The measurement of meaning. Urbana, Ill.: Univer. of Illinois Press, 1957.

Osgood, C. E., & Tannenbaum, P. H. The principle of congruity in the prediction of attitude change. Psychol. Rev., 1955, 62, 42-55.

Osgood, C. E., Ware, E. E., & Morris, C. Analysis of the connotative meanings of a variety of human values as expressed by American college students. J. abnorm. soc. Psychol., 1961, 62, 62-73.

Rabin, A. I. A contribution to the "meaning" of Rorschach's inkblots via the semantic differential. J. consult. Psychol., 1959, 23, 368-372.

Resnick, J., & Landfield, A. W. The oppositional nature of dichotomous constructs. Psychol. Rec., 1961, 11, 47-55.

Rosen, E. A cross cultural study of semantic profiles and attitude differences: Italy. J. soc. Psychol., 1959, 49, 137-144.

Seman, Catharine B. Use of the semantic differential with lobotomized psychotics. J. consult. Psychol., 1957, 21, 264.

Senders, Virginia L. Measurement and statistics. New York: Oxford Univer. Press, 1958.

Siegel, S. Nonparametric statistics. New York: McGraw-Hill, 1956.

Smith, R. G. Development of a semantic differential for use with speech related concepts. Speech Monogr., 1959, 26, 263-272.

Smith, R. G. A semantic differential for theater concepts. Speech Monogr., 1961, 28, 1-8.

Smith, R. G. A semantic differential for speech correction concepts. Speech Monogr., 1962, 24, 32-37.

Solley, C. M., & Messick, S. J. Probability-learning, the statistical structure of concepts, and the measurement of meaning. Amer. J. Psychol., 1957, 70, 161-173.

Solomon, L. N. Semantic reactions to systematically varied sounds. J. Acoust. Soc. Amer., 1959, 31, 986-990.

Springbett, B. M. The semantic differential and meaning in non objective art. Percept. mot. Skills, 1960, 10, 231-240.

Tannenbaum, P. H., & Lynch, M. D. Sensationalism: The concept and its measurement. Journ. Quart., 1960, 37, 381-392.

Tolor, A. The "meaning" of the Bender-Gestalt test designs: A study in the use of the semantic differential. J. proj. Tech., 1960, 24, 433-438.

Triandis, H. C. Differential perception of certain jobs and people by managers, clerks, and workers in industry. J. appl. Psychol., 1959, 43, 221-225.

Triandis, H. C. A comparative factorial analysis of job semantic structures of managers and workers. J. appl. Psychol., 1960, 44, 297-302.

Triandis, H. C., & Osgood, C. E. A comparative factorial analysis of semantic structures in monolingual Greek and American college students. J. abnorm. soc. Psychol., 1958, 57, 187-196.

Walker, Helen M., & Lev J. Statistical inference. New York: Holt, Rinehart, & Winston, 1953.

Weinreich, U. Review of: "Osgood, Suci, and Tannenbaum, The measurement of meaning." Word, 1958, 14, 346-366.

Winter, W. D. Values and achievement in a freshman psychology course. J. educ. Res.,

Wiener, D. N., & Ehrlich, D. "Goals" and "Values." Amer. J. Psychol., 1960, 73, 615-617.

Wohl, J. A note on the generality of constriction. J. proj. Tech., 1957, 21, 410-413.

Zax, M. Loiselle, R. H., & Karras, A. Stimulus characteristics of Rorschach inkblots as perceived by a schizophrenic sample. J. proj. Tech., 1960, 24, 439-443.

Appendix A

A Survey of Semantic Relations

The purpose of this study is to develop a method for finding out what criteria people use for making judgments about things--what kinds of questions would they want to ask about a particular thing before they could decide whether it was a "better" or "worse" thing of its kind. For example, you probably don't care whether your friends are large or small, but that's the first question you would ask about a pay check. You may not care whether your automobile is red or green, but it makes a difference in apples. If we had a hundred years to spare, we might be able to answer this question by discussion, but this study is an attempt to get an answer more quickly than that.

On the following pages are scales with adjectives at each end that look like this:

left ____:____:____:____:____:____:____ right

The intervals on these scales may be interpreted as extremely left, quite left, slightly left, neither or both, slightly right, quite right, and extremely right. Of course you are to substitute whatever words occur at the left and right ends of the scales.

At the top of each page is a concept, such as DOG. What you are to do is to think of the best imaginable and the worst imaginable examples of the class of things named by that concept (in this case, the best imaginable DOG and the worst imaginable DOG) and indicate where you think the best and the worst examples fall on each of the scales on that page. For example, if you happen to like large DOGS, and you don't care much for small DOGS, you might indicate that the best imaginable DOG is extremely large and the worst imaginable DOG is quite small. Of course, your best DOG may be "gentle" and your worst DOG "mean," but you will have an opportunity to indicate that on another scale.

Indicate your feeling for the "best" example by marking a "B" on the scale in the appropriate place. Indicate

"worst" by marking a "W" in the appropriate place. Your responses might look like this:

DOG

large __B__:____:____:____:____:__W__:____ small

mean __W__:____:____:____:____:____:__B__ gentle

green ____:____:____:__BW__:____:____:____ red

The latter mark indicates that you don't really care whether a DOG is green or red or that this scale just doesn't apply to DOGS. With concepts such as ELEPHANT, MONSTER, or RUBY you might feel that one of the extreme positions on the scale describes all the members of the class, in which case you should mark "BW" in the extreme position. Just make sure that you have two marks on every scale.

(Concept)

hot ____:____:____:____:____:____:____ cold

ornate ____:____:____:____:____:____:____ plain

small ____:____:____:____:____:____:____ large

honest ____:____:____:____:____:____:____ dishonest

poor ____:____:____:____:____:____:____ rich

complex ____:____:____:____:____:____:____ simple

usual ____:____:____:____:____:____:____ unusual

healthy ____:____:____:____:____:____:____ sick

fast ____:____:____:____:____:____:____ slow

dead ____:____:____:____:____:____:____ alive

masculine ____:____:____:____:____:____:____ feminine

(Concept)

humorous ____:____:____:____:____:____:____ serious

empty ____:____:____:____:____:____:____ full

thick ____:____:____:____:____:____:____ thin

bass ____:____:____:____:____:____:____ treble

yellow ____:____:____:____:____:____:____ blue

pleasant ____:____:____:____:____:____:____ unpleasant

high ____:____:____:____:____:____:____ low

opaque ____:____:____:____:____:____:____ transparent

sacred ____:____:____:____:____:____:____ profane

mature ____:____:____:____:____:____:____ youthful

weak ____:____:____:____:____:____:____ strong

constricted ____:____:____:____:____:____:____ spacious

pungent ____:____:____:____:____:____:____ bland

rugged ____:____:____:____:____:____:____ delicate

aggressive ____:____:____:____:____:____:____ defensive

soft ____:____:____:____:____:____:____ loud

proud ____:____:____:____:____:____:____ humble

nice ____:____:____:____:____:____:____ awful

objective ____:____:____:____:____:____:____ subjective

unfair ____:____:____:____:____:____:____ fair

tasty ____:____:____:____:____:____:____ distasteful

sad ____:____:____:____:____:____:____ happy

(Concept)

relaxed ____:____:____:____:____:____:____ tense

rough ____:____:____:____:____:____:____ smooth

new ____:____:____:____:____:____:____ old

sweet ____:____:____:____:____:____:____ sour

old ____:____:____:____:____:____:____ young

calm ____:____:____:____:____:____:____ agitated

bitter ____:____:____:____:____:____:____ sweet

long ____:____:____:____:____:____:____ short

black ____:____:____:____:____:____:____ white

clear ____:____:____:____:____:____:____ hazy

constrained ____:____:____:____:____:____:____ free

sober ____:____:____:____:____:____:____ drunk

bright ____:____:____:____:____:____:____ dark

stale ____:____:____:____:____:____:____ fresh

kind ____:____:____:____:____:____:____ cruel

heavy ____:____:____:____:____:____:____ light

far ____:____:____:____:____:____:____ near

angular ____:____:____:____:____:____:____ rounded

red ____:____:____:____:____:____:____ green

passive ____:____:____:____:____:____:____ active

colorful ____:____:____:____:____:____:____ colorless

worthless ____:____:____:____:____:____:____ valuable

(Concept)

ferocious ____:____:____:____:____:____:____ peaceful

severe ____:____:____:____:____:____:____ lenient

beautiful ____:____:____:____:____:____:____ ugly

tough ____:____:____:____:____:____:____ fragile

savory ____:____:____:____:____:____:____ tasteless

changeable ____:____:____:____:____:____:____ stable

wide ____:____:____:____:____:____:____ narrow

sharp ____:____:____:____:____:____:____ blunt

cowardly ____:____:____:____:____:____:____ brave

hard ____:____:____:____:____:____:____ soft

dull ____:____:____:____:____:____:____ sharp

good ____:____:____:____:____:____:____ bad

dirty ____:____:____:____:____:____:____ clean

rational ____:____:____:____:____:____:____ intuitive

tenacious ____:____:____:____:____:____:____ yielding

excitable ____:____:____:____:____:____:____ calm

fragrant ____:____:____:____:____:____:____ foul

curved ____:____:____:____:____:____:____ straight

deep ____:____:____:____:____:____:____ shallow

wet ____:____:____:____:____:____:____ dry

Appendix B

A Study in Semantic Relationships

You are about to participate in an experiment. This experiment is part of a larger study that is attempting to find out the WHY behind statements like "This one's better than that one," or "This one's not as bad as that one."

You will be given a set of 76 cards. On the first one is a concept (MOTHER, ME, SIN, AMERICA, SYMPHONY, or BOULDER). It is assumed that this concept names a class of objects or events, some of which you like better than others or dislike less than others. It is also assumed that you have some reasons for your likes and dislikes. For example, you probably prefer a capitalistic AMERICA to a communistic AMERICA, <u>because</u> it is capitalistic rather than communistic.

The next 75 cards in your set (numbered sequentially in the lower right hand corner) each has a pair of adjectives typed on it, like communistic-capitalistic. Each pair of adjectives is assumed to name a dimension on which events named by your concept might differ. You must decide which of these dimensions are <u>most</u> <u>important</u> to you--

which ones carry the most weight--when you are trying to decide that "This one is better," or "That one is not as bad."

RANK THE 75 ADJECTIVES CARDS IN ORDER OF IMPORTANCE TO YOUR FINAL EVALUATION OF A PARTICULAR EXAMPLE OF THE CONCEPT.

You might begin by dividing the set of cards into three stacks--important, maybe important, not important--and then divide them again and again until you have the set in rank order with the most important on top. Put the concept card last.

Appendix C

A Survey of Semantic Relations

The purpose of this study is to measure the meanings of certain things to various people by having them judge them against a series of descriptive scales. In taking this test, please make your judgments on the basis of what these things mean to you. On each page of this booklet you will find a CONCEPT to be judged and beneath it a set of scales. You are to rate the concept on each of the scales in order.

Here is how you are to use these scales:

If you feel that the concept at the top of the page is very closely related to one end of the scale, you should place your check-mark as follows:

fair __X__:____:____:____:____:____:____ unfair

OR

fair ____:____:____:____:____:____:__X__ unfair

If you feel that the concept is quite closely related to one or the other end of the scale (but not extremely), you should place your check-mark as follows:

fair ____:__X__:____:____:____:____:____ unfair

OR

fair ____:____:____:____:____:__X__:____ unfair

If the concept seems only slightly related to one side as opposed to the other side (but not really neutral), then you should mark as follows:

fair ____:____:__X__:____:____:____:____ unfair

fair ____:____:____:____:__X__:____:____ unfair

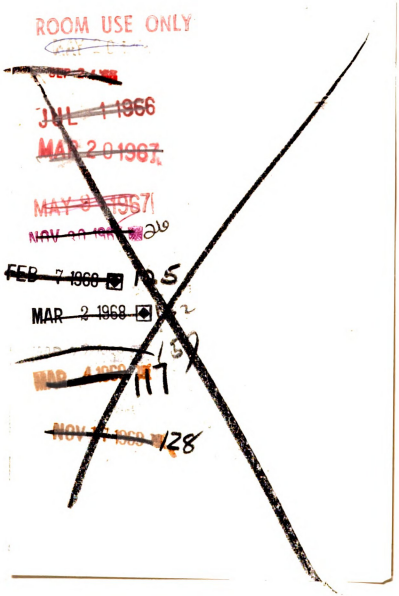
The direction twoard which you check, of course, depends upon which of the two ends of the scale seems most characteristic of the thing you're judging.

If you consider the concept to be neutral on the scale, both sides of the scale equally associated with the concept, or if the scale is completely irrelevant, unrelated to the concept, then you should place your check-mark in the middle space:

fair ____:____:____:__X__:____:____:____ unfair

IMPORTANT:
(1) Be sure you check every scale for every concept--do not omit any.
(2) Make one and only one check mark on each scale.
(3) Make each item a separate and independent judgment.

Work at a fairly high speed through this test. Do not worry or puzzle over individual items. It is your first impressions, the immediate "feelings" about the items, that we want. On the other hand, please do not be careless, because we want your true impressions.


ROOM USE ONLY
JUL 1 1966
MAR 2 0 1967
MAY 8
NOV
FEB 7 1968
MAR 2 1968
MAR 4
NOV
219
105
117
128