Psychological Bulletin
1969, Vol. 72, No. 6, 406-422

SOME METHODOLOGICAL ISSUES IN SEMANTIC DIFFERENTIAL RESEARCH [1]

DAVID R. HEISE [2]

Department of Sociology, Queens College of the City University of New York

Methodological research on the semantic differential (SD) is reviewed concerning metric, sources of rating variation, and the structure of ratings. Major conclusions are as follows: (*a*) The metric assumptions involved in SD scales are in some ways inaccurate but adequate for many applications. (*b*) Biased errors may arise in SD data because of social desirability effects or because of scale-checking styles. (*c*) A substantial portion of variation in SD ratings is due to individual differences and temporal variations in responses. (*d*) The basic dimensions of average response on SD scales are Evaluation, Potency, and Activity, and no extensive proliferation of basic dimensions beyond these can be expected. (*e*) There are individual differences in the size and character of the semantic space. (*f*) The appearance of scale-concept interactions frequently is a methodological artifact which would not occur in adequately designed studies. (*g*) The existence of real scale-concept interactions demands tailoring the SD to different stimulus domains, but the studies required for this must be carried out with considerable care.

Of the more than 1,000 articles and books dealing with the semantic differential, many focus on methodological problems associated with the instrument. Methodology was discussed in detail in *The Measurement of Meaning* (Osgood, Suci, & Tannenbaum, 1957), but the cumulation of research since then warrants a reexamination of some major methodological issues. Some ambiguous areas have been clarified by research and criticism, while other topics remain problematic.

This review deals with four primary topics: the metric of semantic differential scales, sources of variance in semantic differential ratings, the structure of semantic differential ratings, and the development of special instruments for use with particular content domains. Methodological considerations in constructing and administering semantic differentials have been reviewed elsewhere (Heise, 1970).

On the following pages references to *The Measurement of Meaning* (Osgood, Suci, & Tannenbaum, 1957) are abbreviated *MM*; semantic differential is abbreviated SD; and the three major dimensions of rating variation—Evaluation, Potency, and Activity—are abbreviated E, P, and A.

[1] This review was facilitated greatly by "A Bibliography of Literature Relevant to the Semantic Differential Technique," made available to the author in mimeographed form by Charles E. Osgood. The bibliography has been published recently in Snider and Osgood (1969).

[2] Requests for reprints should be addressed to David R. Heise, Department of Sociology, Queens College of the City University of New York, Flushing, New York 11367.

METRIC

The SD measures people's reactions to things in terms of ratings on scales defined with contrasting adjectives at each end. An example of an SD scale is:

Good	—	—	—	—	—	—	—	Bad
	3	2	1	0	−1	−2	−3	

Specifying the positions on the scale with a set of adverbs facilitates the rating task (Wells & Smith, 1960), and usually the position coded 0 is labeled "neutral," the positions coded 1 or −1 are labeled "slightly," the 2 positions "quite," and the 3 positions "extremely." Seven-point scales are customary in SD research, because of early methodological research by Osgood and his colleagues (*MM*, p. 85) and more recently in deference to Miller's (1956) argument that not many more than seven discriminations can be made simultaneously.

SD data are coded numerically as if the scales were bipolar, equal interval scales each passing through the origin of the SD space. Thus, a number of metric assumptions are involved in the use of SD scales.

Bipolarity

Are SD adjectives really bipolar in the sense of being at about opposite points in the SD space? Most of the adjective pairs used in SD work are true linguistic contrasts (Deese, 1964), and it is assumed in SD work that linguistic contrasts provide a means for making up scales which define basic affective contrasts. That is, it is assumed that two contrasting adjectives plotted in the SD space would be about equidistant from the neutral center point, and they also would be opposite one another so that a line passing between them would also pass through the center. Mordkoff (1963, 1965), using an approximate technique to test for "functional antonymy," found that some scales—typically those used most in SD research—do meet these criteria, but that others do not. Two scales that clearly are not true affective contrasts are masculine-feminine and hard-soft.

Green and Goldfried (1965) argued that if the adjectives used in SD scales are real contrasts, then unipolar ratings using each adjective separately should correlate negatively and thereby define bipolar factors. They did find bipolar factors for Evaluation adjectives but not for the other dimensions. Thus, they concluded that the assumption of bipolarity generally is unwarranted. However, unipolar ratings probably have more sources of variance than ratings on scales which are doubly anchored, so the Green and Goldfried procedure may not be a sufficient test of bipolarity. Furthermore, when ratings on unipolar scales are summed, the results do correlate with the corresponding bipolar scales (MM, p. 153, footnote). In other words, if the extraneous sources of variance in unipolar ratings are partially canceled by averaging, then the unipolar ratings do correspond to bipolar ratings.

On the whole, the bipolarity assumption is probably justified for most scales used in SD research. Some scales do exist, however, which do not meet the assumption of true bipolarity. Use of such scales can distort measurements of the EPA structure.

Equal Intervals

The metric characteristics of adverbial quantifiers have been investigated in a number of studies (Cliff, 1959; Howe, 1962, 1966a, 1966b). The results indicate that the adverbs "extremely," "quite," and "slightly" should define rating positions which are about equidistantly spaced.

Messick (1957) applied the method of successive intervals to SD data to determine if the assumption of equal intervals is warranted. He found that category boundaries were similarly spaced on all of the nine scales he considered, but not exactly in the proper positions for equal intervals. One side of a scale generally has slightly larger intervals than the other side. Unfortunately, there is ambiguity about the meaning of this—in Messick's data the larger intervals are always on the left side of a scale, but the left side also is always the side on which the good, strong, or active poles appear. Messick also found that the end intervals tend to be larger than those toward the middle. This arises because ratings on SD scales are bounded, that is, there is no position beyond "extremely" and so ratings of very extreme concepts tend to pile up in the end category (Cliff, 1959, also noted this effect).

Messick indicates that despite the deviations from the equal interval assumption, one does not go far wrong in making this assumption. The correlations between the assumed and scaled boundaries were greater than .98 for every scale considered.

Zero Point

Messick's (1957) scaling study revealed that the center point of SD scales is not true zero, but rather a point lying about .2 scale units away from true zero. Unfortunately, it is again impossible to tell whether this effect arises from a left-right bias or whether it is determined by the orientation of the good, strong, or active pole of the scale.

Conclusions

While a few studies are available concerning the metric of SD scales, this area has remained one of the least studied topics in SD methodology. The information available suggests that the basic metric assumptions for

the SD are not quite accurate, but also that violations of the assumptions are not serious enough to interfere with many present applications of the SD. Furthermore, some metric errors would be expected to counteract one another when ratings on several different scales are added together to form factor scores.

While lack of a refined metric may not interfere with most uses of the SD, this inadequacy could create problems in certain kinds of work. For example, studies of semantic satiation (e.g., Messer, Jakobovits, Kanungo, & Lambert, 1964) postulate that when words are repeated continuously, they become meaningless, and therefore, ratings on SD scales should drift toward center. Tests of this idea depend directly on the assumption that the true zero point of each scale is located at the middle check position. If the assumption is wrong, the index of meaninglessness is inaccurate, and results of experiments cannot be interpreted unambiguously. Failure to meet the equal intervals assumption may produce confounding of results in studies aimed at developing precise models for predicting attitudinal processes (e.g., Gollob, 1968; Heise, 1969a). In particular, wider intervals at the endpoints of scales conceivably could systematically distort data so that the empirically derived models do not correspond directly to psychological processes. Problems such as these emphasize the need for more research on SD metric.

Sources of Rating Variance

Variance in SD ratings derives from a number of sources. A basic division is between error variance and true variance. True variance is that derived from actual variations in affective responses, and error variance is that due to other factors. Error variance itself can be divided into two classes: biased error and random error.

Biased Error—Social Desirability

In a study by Nickols and Shaw (1964) dealing with the validity of SD ratings as attitude measurements, it was found that the relationship between SD measurements and Thurstone measurements varies under certain conditions. Attitudes toward *college professors* and toward *the church* were assessed using Evaluative factor scores and two Thurstone scales. When the attitude object was nonsalient for subjects, the relationship between the types of measurement was high: $r = .71$ for noncollege subjects rating *professor* and $r = .76$ for college students rating *the church*. However, when attitude object was salient for the subjects, the relationship between the two types of measures dropped to low values: $r = .29$ for college students rating *professors* and $r = .39$ for church attenders rating *the church*. Nickols and Shaw noted that even though variations in attitude were less in the high-saliency groups, this difference was not enough to explain the drops in correlation. They also presented evidence that the Thurstone scales retained their reliability as measurements of individual differences in the high-saliency groups. An ad hoc analysis by this author (using data summarized by Heise, 1965) indicated that saliency of attitude objects does not affect the reliability of SD ratings either.

Nickols and Shaw hypothesized that subjects are more sensitive to the social repercussions of their ratings when dealing with salient objects, and that the SD is more transparent as a measure of attitude. Thus, social desirability may enter as a factor in SD ratings of salient objects. This interpretation receives indirect support from a study by Ford and Meisels (1965) which showed that the social desirability of SD scales corresponded directly to their loading on the Evaluation dimension. (The Potency and Activity dimensions are unrelated to social desirability.) Thus, it may be true that direct SD ratings of objects may not be an efficient approach to measurement when salient or delicate topics are involved, because subjects can distort their responses in the direction of social desirability. However, before concluding this firmly, one would like to see replications which involve more than two attitude objects, in which subject's need for approval is an actual control variable, in which a criterion is used to show that the SD measurements are the less valid, and in which SD measurements are made on all three dimensions.

Biased Error—Scale-Checking Styles

There appear to be important differences between persons in scale-checking styles. In particular, some subjects appear to use the end points of scales more often and to avoid the intermediate discriminatory positions (Peabody, 1962). This propensity is a stable trait of individuals over time and over different sets of concepts: When measurements of scale-checking style are made, the test-retest and split-half correlations of the measurements are above .70 (Arthur, 1966).

Scale-checking styles introduce biased error by moving measurements toward the end points or midpoint of a scale. Furthermore, it seems likely that the consistency and extent of end-point checking is related to the true scores for concepts. For example, if a subject with such a bias rates a "quite good' concept on five evaluation scales, he is likely to check the positive end on all five scales; if he rates a "slightly good" concept, he is more likely to alternate between the end points and midpoints of the scales. The result would be that when averaging his ratings to get factor scores, we would find systematic deviations from true scores, and generally speaking, the amount of deviation would be larger for the more polarized concepts. This amounts to exaggeration error (Kahneman, 1963), a nonrandom deviation of a subject's rating from the true score which varies with the true score for the concept and the individual's propensity to exaggerate when using rating scales. Exaggeration biases introduce problems in SD methodological research (see section on concept-scale interaction), and they complicate cross-person comparisons of attitudes by raising the question, Does the obtained difference in scores reflect true differences in affective responses or merely differences in propensity to exaggerate?

Some studies of scale-checking styles are available, and these suggest that certain classes of people are more likely to overemphasize scale end-points. *Age*: Children tend to use the end points and center more than adults (*MM*, p. 85); *IQ*: Low IQ is associated with greater use of end points but mainly among young children (*MM*, pp. 227–228; Stricker & Zax, 1966); *F scores*: High scorers on the F scale tend to use the extremes more often (Mogar, 1960); perhaps related is the tendency of American Legionnaires to use the extremes and middle more (*MM*, p. 85); *Sex*: Some studies suggest that females use the extremes more (Dixon & Dixon, 1964; Goldfried & Kissel, 1963); another study reports no sex difference (Stricker & Zax, 1966); *Neurosis*: Some studies indicate that neurotics use the extremes more (Arthur, 1965; Zax, Gardiner, & Lowy, 1964); there is one report of no differences between neurotics and controls in the use of scales (Luria, 1959); *Psychosis*: A number of studies indicate that psychotics use the extremes more than normals (Arthur, 1965; *MM*, pp. 85, 227; Zax, Gardiner, & Lowy, 1964), and other studies provide additional inferential evidence for this (Beitner, 1961; Neuringer, 1963).

The available research documents the existence of basic differences between people in the extent to which extremes are checked. Therefore, exaggeration bias should be a considered variable in experiments, controlled either by random assignment of subjects into experimental and control groups or by attempting to measure the effect in order to control for it statistically.

It should be noted that available research has little to say about the possibility that use of extremes may reflect a true personality condition in which a person has intense responses to nearly everything. Research is needed to determine if such personality states exist and if they do, to provide means for distinguishing the personality state from the scale-checking style.

Random Error—Reliability Studies

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

that Evaluation scales tend to be somewhat more reliable at all grade levels.

A reliability study by Norman (1959) gives information on how much shift in ratings occurs relative to what might be expected if the ratings were purely random. Norman had 30 college subjects rate 20 concepts on 20 scales in a test and retest spaced 4 weeks apart. On the average, he found that the amount of shift in ratings was about 50% of what would be expected if the ratings were completely random. More specifically, his results showed that 40% of the scale ratings do not shift at all from test to retest, 35% of the ratings shift by 1 scale unit, and 25% of the ratings shift 2 or more scale units. Norman found that ratings are more stable for some concepts than for others, and this seems to be related to the number of meanings for a concept; for example, *leper* and *tornado* are relatively stable concepts, whereas *stars* is unstable. (This also may be a function of how extremely the word is rated; other studies suggest that concepts whose true values are neutral are rated with less reliability—Luria, 1959; Peabody, 1962.) Norman also found that some subjects were more stable than others in making their ratings; in particular, there is a tendency for those who use the end points of the scales more often to have lower test-retest stability. Finally, he found that certain scales are associated with greater stability; in particular, Evaluation scales evoke fewer shifts.

Factor scores. A study is reported in *MM* (p. 192) in which several controversial topics were rated on six evaluation scales; and factor scores, representing each subject's evaluative reaction to a given topic, were obtained by summing the ratings on the six scales. The correlations between test and retest factor scores ranged from .87 to .97 with a mean of .91. DiVesta and Dick (1966), in their study of SD reliability among children, made up factor scores by averaging ratings on two scales for a given dimension and correlating the measurements from the first test with those from the second test 1 month later. For children in the fourth grade or later, the correlations ranged between .5 and .8 and were highest for Evaluation factor scores; for students in the third or earlier grades, test-retest correlations ranged between .4 and .5. (Reduced reliability of factor scores among younger children also were found by Maltz, 1963.) DiVesta and Dick found that test-retest correlations were somewhat higher when the test followed the first test immediately: In this case the *r*'s ranged between .6 and .8. Norman (1959) examined the effect of making up factor scores from various numbers of scales. His results indicated that factor scores are more reliable than single ratings and that most of the gain in test-retest stability is accomplished by averaging just three or four scales; going up to an eight-scale factor score seems to add very little additional stability when looking at data from a test and retest spaced 1 month apart.

The various studies indicate that there is a gain in test-retest correlations when factor scores are used rather than individual scale ratings. Furthermore, it appears that most of the possible improvement can be obtained using relatively few scales in making up the factor scores. The first of these empirical findings is exactly what one would expect in terms of classical test theory: Basing a score on multiple items that load on a factor increases the precision with which that factor is measured (Cronbach, 1951). The second finding is somewhat surprising since ordinarily one expects continual gains in reliability as more relevant items are added into the total score. The anomaly can be explained by proposing that combining relatively few SD scales produces a factor score of high reliability which, however, has somewhat low stability in tests-retests spaced a month apart (see Heise, 1969b).

Group means. Many SD studies do not focus on an individual's rating of a concept but on a group mean. In such a case, there is averaging both across scales (factor scores) and across persons, and test-retest correlations should be higher.

DiVesta and Dick (1966) calculated factor score means for groups of from three to five children. The immediate test-retest correlations ranged from .73 to .94, figures that are significantly greater than the correlations based on individual subjects. Norman (1959) calculated scale means for 20 concepts using groups of 30 raters. The test-retest correla-

tion between means was .96, and the correlation between means produced by two different samples of student respondents was .94. Miron (1961) averaged the factor scores for 20 concepts across 112 subjects and obtained test-retest correlations of .98 or more.

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

True Variance

An SD measures a *person's* reaction to a given *stimulus* at a given *time*. Thus, there are three major sources of true-score variance. Variance due to stimuli is the focus of most SD studies. The other two sources of variance—individual differences and temporal instability—sometimes are ignored and thus warrant some comment here.

When ratings of a concept are obtained from a sample of persons, one finds variability around the group mean. The reliability studies suggest that part of this variance is random and due to imprecision in the instrument. Certainly, however, some other part of it is due to individual differences in reactions. The existence of individual differences is verified in studies that use SD ratings of a single concept for predictive purposes, and such studies are too numerous for review here. The nonrandomness of individual deviations also is indicated in a methodological study by Kahneman (1963). He showed that correlations between SD scales are essentially the same whether based on mean scores for concepts or on the deviations from the mean for a single concept. This implies that the EPA structure can be derived from the deviations, and thus the deviations themselves are meaningful, nonrandom data. (This contradicts a statement in *MM*, p. 137, to the effect that deviations are independent, but Kahneman points out that the data presented by Osgood et al. 1957, actually support the thesis of nonindependence. The data of concern are change scores and must be interpreted with reference to the problems in analyzing change—Harris, 1963.)

Another part of the overall variation in SD ratings should be due to temporal instability of reactions. For example, a male may rate *girl* as quite good one day and as slightly bad on another day following a fight with his fiancée; this discrepancy between ratings is not random error, but due to an actual change. There are a number of clues in the various reliability studies which suggest that temporal instability does account for a substantial proportion of the variation in SD ratings.

It was reported in *MM* (p. 136) that ratings of subjective concepts like *me* or *my mood* involve larger test-retest deviations than ratings of more objective concepts, so concepts which one would suppose are temporally unstable in fact are associated with less "reliable" ratings. DiVesta and Dick (1966) found higher test-retest correlations when the retest was immediate than when it was delayed by a month: This would be expected if the true scores do tend to change over time. Norman's (1959) finding that adding together more scales improves test-retest reliability of factor scores only up to a point also suggests that instability contributes to low test-retest correlations since this is not what one would expect if the variations around the mean rating were all random and meaningless. In such a case, adding in more related scales should continually improve precision. On the other hand, the limit on improvement in precision is what one would expect if some of the rating variance represents true, but temporal, variations in reactions.

Partitioning the Variance in SD Ratings

SD ratings, then, are determined partly by the object being rated, partly by the idiosyncratic and momentary reactions of the particular rater, and partly by random error. Kahneman (1963) has presented a model of SD scaling in which these and other determinants of rating variance are considered systematically.

In Kahneman's model, the sources of variation in ratings on a single scale are as follows (holding the subject population constant). (*a*) The true score of Concept x when rated on Scale p: this is estimated by the mean rating of x on p calculated over subjects. (*b*) Subject a's personal bias in using Scale x: this is estimated by finding the

difference between *a*'s rating and the true score and then averaging these differences over all the concepts considered. (*c*) The unique deviation that occurs when Subject *a* rates Concept *x* on Scale *p*: this is estimated as the difference between *a*'s actual rating of *x* on *p* and the true score with a correction for the subject-scale interaction. The unique deviation can be partitioned further into the following. Bias—Subject *a*'s tendency to give exaggerated ratings on scales where the degree of exaggeration is proportional to the true score. Constant individual deviation—this would be measured as the difference between a subject's own true score (obtained by averaging over repeated measurements) and the population true score. Momentary deviations by subjects—this corresponds to a subject's real but temporal deviation from his own true score. Random error—the amount of varation left when the other sources have been removed.

A very crude estimate from information presented by Kahneman and others mentioned in the sections above indicates that often the variance of ratings of a given concept on a given SD scale splits up as follows: one-tenth due to subject-scale interaction, that is, due to differences between subjects in the use of scales, one-quarter due to bias and/or deviations of subjects' true scores from the population true scores, one-quarter due to momentary deviations of subjects from their own true scores and two-fifths due to random error. The two-fifths random error is for ratings on a single scale, and this component of the variance would be considerably less for factor scores. What is especially notable is that as much as one-half of the rating variance may be due to individual and temporal variations among subjects.

This partitioning of the variance was inferred from a variety of correlations reported in several different studies, and an experiment addressed specifically to the problem of partitioning the variance in SD ratings has yet to be done. Having been pieced together from studies focusing on other matters, the above estimates should be viewed as rough and very tentative average figures. One would expect the proportions to differ across concepts since there is clear evidence (Snider & Osgood, 1969, Appendix) that individual variations in ratings are much greater for some concepts (e.g., *atheists*) than others (such as *arm*). One also would expect variations across scales and subject populations.

Structure of Ratings

EPA Dimensions

Factor analyses of SD data consistently show that there are three major dimensions of rating response—Evaluation, Activity, and Potency. Studies dealing with a great variety of scales, stimuli, and subjects have demonstrated the prominence and significance of the EPA structure in SD data.

Among the studies reported in *MM* (pp. 47–66) was the thesaurus study in which 76 adjective contrasts were chosen from *Roget's Thesaurus* and the corresponding bipolar scales were used by 100 college students to rate 20 different concepts. Correlations between the ratings on different scales were calculated and factored. The EPA structure was clearly evident. Bopp (reported in *MM*, pp. 223–226) had 40 schizophrenics rate 32 words on a 13-scale form; the usual EPA structure was recognizable. Wright (1958) had 40 concepts rated on a 30-scale SD by a survey sample of 2,000 men and women spread over the spectrum of socioeconomic status. Wright found four factors in his data, the first three of which clearly were EPA. Heise (1965) had 1,000 concepts rated on 8 scales by Navy enlistees; factor aanlyses of the data based on mean ratings for the 1,000 different words yielded the usual EPA structure. DiVesta (1966) had 100 concepts rated on 27 scales by subjects in Grades 2 through 7 (20 subjects were used for each concept). The usual EPA structure emerged, though there was some tendency for Potency and Activity to merge into a single Dynamism dimension up until the fifth grade. DiVesta also reports another study in which grade school children used 21 scales to rate 100 different concepts (this time with 100 subjects rating each concept) and combining all the data for all grades, the usual EPA structure was found.

Osgood (1962) reviewed several early studies aimed at determining whether the

EPA structure is idiosyncratic to English or whether it holds up within other languages and other cultures. G. J. Suci (1960) had illiterate Navajo, Hopi, and Zuni respondents make ratings by pointing; the data obtained revealed Evaluation and Potency factors; Activity did not appear separately, possibly because not enough Activity scales were included, or possibly because the set of concepts did not introduce enough Activity variance. H. Akuto (reported in Osgood, 1962) had 100 Japanese subjects rate 90 concepts on 50 scales in Japanese and found that the EPA structure was clearly evident in the factor structure.

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

Dimensionality

The stress in most SD research is on the three EPA dimensions, as if additional dimensions do not exist. However, the dimensionality of SD ratings still is not a completely settled issue. Following are some studies and arguments that suggest that there may be more than three dimensions of rating response.

1. A number of dimensions besides EPA were found in the Thesaurus study, and throughout *The Measurement of Meaning*, the presumption was that additional dimensions do exist.

2. Carroll, in his reviews of SD technology (1959a, 1959b), points out that early SD factor analyses were based on ratings of no more than 20 concepts, and if ratings were taken over more concepts, or more systematic samples of concepts, more factors might appear. The sample of scales used in an SD study also is of critical importance in determining dimensionality since a factor can appear only if several scales measuring that factor are included in analyses.

3. Green and Goldfried (1965) employed unipolar rather than bipolar adjective scales, and more than three factors were found to characterize their data.

4. When adjective ratings are used to assess persons, one frequently finds about five important factors appearing (e.g., Borgatta, 1964; Norman, 1963).

5. In a study by Komorita and Bass (1967), it was found that when ratings of a single concept on a set of Evaluation scales are factored, the Evaluation dimension appears to split into three subdimensions—functional value, hedonic value, and ethical value. Other studies in which correlations were calculated over individual ratings (e.g., Smith, 1961) also have found Evaluation breaking into subdimensions.

6. Wiggins and Fishbein (1969) had subjects rate the similarity of 15 SD scales and subjected the data to several dimensional analyses. Averaging ratings across all subjects, they found the usual EPA structure under-

lying the similarity judgments. However, additional analyses revealed individual differences in the structure of the semantic space. Some subjects operated in terms of a simplified, two-dimensional space in which Activity scales either aligned with Evaluation or Potency or else failed to achieve any communality at all. Other subjects operated with the usual EPA structure and still other subjects operated with elaborated, four-dimensional structures. The four-dimensional structures developed as a result of splintering in either the Evaluation or Activity factors. Wiggins and Fishbein found that the individual differences in semantic space have small but significant correlations with some personality variables. These results do not suggest the invalidity of the basic EPA structure, nor do they suggest that numerous additional factors will be found. However, the results do indicate that the dimensionality of the semantic space can vary as a function of the individuals who are employed as subjects.

On the other hand, there are counter-arguments and some indications that the EPA dimensions are the only *major* dimensions of average response.

1. In the program of cross-cultural research being conducted by Charles Osgood and his colleagues, it has been found that the EPA dimensions generally are the most significant sources of variance, and these are the only factors that replicate across cultures.[8]

2. Studies dealing with more than 20 concepts and employing a fair number of scales have not resulted in an expansion in the number of factors. Indeed, some of the minor factors found in the thesaurus study appear to drop out as significant sources of common variance when more concepts are considered. For example, Wright (1958) applied 30 scales to 40 concepts and DiVesta (1966) used 27 scales with 100 concepts, and in both cases three or four factors accounted for most of the common variance. In the program of cross-cultural research mentioned above, 100 concept are rated on 50 SD scales within each culture, and the EPA structure dominates the factor analyses. This latter finding is particularly impressive when it is remembered that the scales for each language are not derived through translation, but through a completely indigenous procedure that emphasizes the basic contrasts within a given culture. So, at this point, a great variety of scales have been employed and the EPA dimensions have been repeatedly verified as the most significant sources of variation.

3. A study of 1,000 concepts by Heise (1965) included scales meant to measure Stability (the dimension which sometimes appears as a fourth factor in SD data). In computing correlations over the mean ratings for all 1,000 concepts, the Stability dimension failed to appear and the Stability scales realigned on the standard EPA structure.

4. In addition, the results of studies in which more than three dimensions are found frequently can be interpreted in terms of methodological variations in the studies as indicated below.

The Green and Goldfried (1965) study dealt with only 10 concepts, and many of the analyses were made over individual ratings, taking one concept at a time; these points in themselves set the study aside from most other SD work. More important, however, is the fact that unipolar ratings may have a markedly different nature from bipolar ratings. When only one adjective is presented, its denotative meaning may have more impact on ratings than when it is used with another adjective, and it may be easier to rate peripheral or fleeting aspects of the stimulus. For example, someone asked if dogs are bad may say "some dogs are bad" and asked separately if dogs are good may say "most dogs are good," whereas if asked whether dogs are good or bad, he would have to make a single summary statement. The Green and Goldfried approach probably is less comparable to ordinary SD procedures than it is to techniques for studying denotative meaning through associative structure (Deese, 1965).

The frequent appearance of five or more dimensions in impressions of persons when rated by other individuals could indicate that there are more meaningful dimensions of response to persons than there are for concepts

[8] C. Osgood, personal communication January, 1969.

in general or it may reflect a split of basic dimensions into subdimensions that occurs because both rater differences and stimuli differences are entered into analyses. If the latter is true, then we should find the usual EPA structure were we to factor mean ratings for person stimuli rather than the ratings of persons by individual others. Burke and Bennis (1961) had persons in training groups rate each other, and the mean ratings for each person in a group were then calculated and used as the basis for a factor analysis over 84 person stimuli. In this analysis only three basic factors appeared which were named Friendliness, Dominance, and Participation. Thus, it does appear that when rater variance is averaged out, the three EPA dimensions are the most salient aspects of person impression.

In general, the findings and arguments reviewed above lend strong support to the proposition that E, P, and A are the most significant factors underlying *averaged* SD ratings. As more comprehensive studies are done, other general factors may appear (something like the Stability dimension is an especially prominent possibility). However, these additional dimensions evidently would be minor in the sense that they contribute to the variance of mean ratings on most scales only to a small degree, and certainly no great proliferation of dimensions beyond the basic three seems likely at this time.

While the literature supports the validity and significance of the EPA dimensions as the basic structure underlying averaged SD ratings, a number of studies also provide strong evidence that the situation is more complicated when one deals with the structure of individual judgments rather than group means. (Though it may seem surprising at first that analyses of group means and of individual ratings could lead to different results, the extensive social science literature on "ecological correlation"—for example, Dogan and Rokkan, 1969—substantiates the possibility.) In particular, whenever factor analyses of adjective ratings are carried out across individual ratings rather than over group means, more than the three EPA factors are found (Green & Goldfried, 1965; Komorita & Bass, 1967; Norman, 1963; Wiggins & Fishbein, 1969). The key studies are those by Komorita and Bass and by Wiggins and Fishbein.

The study by Komorita and Bass (1967) found that the Evaluative dimension splintered into three subdimensions when analyses were carried out over individual ratings rather than group means. Thus, this study suggests that there are multiple modes of evaluating and that the EPA structure may be an oversimplification when applied to individuals. Unfortunately, though, the study was not elaborate enough either to understand the mechanisms involved or to discard alternative explanations of the results as being due to concept-scale interaction or subject scale relevance (see the discussion of these matters in the section on concept-scale interaction).

The study by Wiggins and Fishbein (1969) was considerably more detailed. It indicates that the simple EPA structure is not a completely accurate description of all individual affective reactions and also suggests some of the mechanisms underlying individual differences. In brief, Wiggins and Fishbein found that there are different types of subjects, some employing a two-dimensional structure (EP), others a three-dimensional structure (EPA), and still others a four-dimensional structure (EPA with either the Evaluation or the Activity dimension splintering into two factors). The EPA structure is clearly the common denominator for all types, but it is subject to some collapse or elaboration at the individual level. The Wiggins and Fishbein results stimulate the speculation that when all types are combined in a single sample and factor analyses are carried out over individual ratings, one would find a five-factor structure: E_1, E_2, P, A_1 and A_2 with E_1 and P being the largest factors since they are common to the most subjects.

Considerably more work needs to be done on individual differences in semantic spaces. At present it is possible to conclude only that the affective responses of individuals do vary primarily along dimensions of evaluation, potency, and activity, but that some persons engage in more affective differentiation and

some persons less than the simple three-factor structure indicates.

Concept-Scale Interaction

Every year studies are done in which an investigator obtains ratings using a set of SD scales, factors the data, and makes the discovery that some standard scales do not have their usual alignment on the EPA structure. For example, in a particular analysis it may turn out that what appeared to be an Evaluation scale now loads on Potency. Such findings usually occur when ratings for a single concept or a single class of concepts are analyzed, and so the phenomenon has come to be called "concept-scale interaction." It should be noted that findings of this sort are not novel. For example, in *MM* (p. 178) it was reported that the scales pleasurable-unpleasurable and masculine-feminine correlated positively for the concept *Adlai Stevenson* but negatively for the concept *My Mother*, and in many studies since then the structure of ratings has been found to be a function of the stimulus concepts.

The issue posed by concept-scale interaction is not whether it exists empirically—it does, but how such results are to be interpreted. Osgood et al. (1957, p. 187) found the phenomenon to be so ubiquitous and striking that they developed a psychological hypothesis to explain it. Other researchers seem inclined to interpret such results as undermining the whole tradition of SD research. Below, two other orientations are taken toward the phenomenon. First, arguments are presented that suggest that many instances of concept-scale interaction represent nothing more than methodological artifacts and thus have no substantive significance. Second, it is proposed that some true instances of concept-scale interaction do occur and are of considerable importance because they indicate stimuli which instigate semantic shifts in adjectives and thereby suggest content areas in which it may be necessary to develop special SD instruments.

Before considering the variety of ways in which concept-scale interactions can arise, it should be noted that analyses revealing this phenomenon usually are carried out over individual ratings rather than group means. This raises the possibility that individual differences in scale usage, as reported by Wiggins and Fishbein (1969), could be confused with concept-scale interactions when comparisons are made across different populations of subjects. The implication is that if one wants to study concept-scale interactions, one should carry out all analyses and comparisons within a single subject population.

Concept selection. Carroll (1959a; 1959b) has warned that the SD factor structure can be affected by the choice of concepts employed in a study, and his observation can be extended to account for some instances of concept-scale interaction. Suppose that one were dealing with only political concepts. These tend to vary primarily along a dimension of "benevolent dynamism" versus "malevolent insipidness" which is a composite of Evaluation, Potency, and Activity (*MM*, pp. 120–124). Thus, if the mean ratings for a number of political concepts were factored, the first and probably the most significant factor would be one in which Evaluation, Potency, and Activity scales all clustered together. While these results would seem to indicate that the scales take on special meanings when applied to political concepts, the actual fact is that the results are merely an artifact produced by choosing concepts all lying along a single line in the SD space.

Unique factor structures associated with a particular choice of concepts do not constitute evidence for scale-concept interaction. That is, they do not indicate that the factorial composition of scales has changed as an effect of the concepts rated. However, there is no way to discriminate between this artifact and real scale-concept interaction on a post hoc basis. The misinterpretation can be avoided only by avoiding the artifact, and this is done by proper selection of concepts for analysis. The matter of selecting concepts or stimuli is discussed further in the section on ad hoc factor analyses.

Biased rating errors. Kahneman (1963) has interpreted many instances of concept-scale interaction as artifacts arising because of correlations between true scores of concepts and individuals' rating errors. Kahneman begins by defining the true score of a concept on a given scale as the mean rating in the

given population of subjects; there is some variation among individual ratings on either side of the true score. Kahneman next hypothesizes that some persons are exaggerators —they tend to give ratings more polarized than the true score, and others are attenuators—they tend to give ratings less polarized than the true score. He also hypothesizes that the degree of exaggeration or attenuation is a function of the polarization of true scores—that is, as the true score is more intense, exaggerators exaggerate more (or more certainly) and attenuators attenuate more. Now suppose that a concept's true score is quite good and quite powerful. Exaggerators will rate it extremely good and extremely powerful, attenuators will rate it slightly good and slightly powerful, and normals will rate it quite good and quite powerful. Thus correlating over individual ratings for this concept, one will find that the good-bad and powerful-powerless scales are positively related. Suppose that a true score of another concept is quite good and quite powerless. Exaggerators will rate this one extremely good and extremely powerless, attenuators slightly good and slightly powerless, and normals quite good and quite powerless. Thus, in this case, it is found that the good-bad and powerful-powerless scales are negatively related. Kahneman shows empirically that his explanation can be used with considerable efficiency to predict the signs of correlations between scales when correlations are calculated across individual ratings for a single concept.

Unique factor structures deriving from nothing more than exaggeration errors are unlikely to contribute anything to the SD literature other than confusion, and one generally would prefer to avoid this kind of artifact. The simplest way to avoid the problem is to calculate correlations between scales using mean ratings for concepts rather than the individual ratings. In other words, use concepts as the units of analysis rather than individuals. When this is impractical, individual ratings can be used if ratings for several different concepts, carefully chosen to be balanced in the semantic space, are all pooled so that the rating biases for different concepts tend to cancel one another out. If it is necessary to analyze ratings for just a single concept, the artifact can be avoided only by employing a homogeneous sample of subjects, that is, all exaggerators, all normals, or all attenuators.

Relevance. Even aside from the problem of exaggeration and attenuation, a within-concepts analysis may produce an atypical factor structure because different scales have different degrees of relevance for different subjects, and nonrelevant scales yield nonmeaningful data (Mitsos, 1961). To illustrate the possibility, consider again the study by Komorita and Bass (1967) in which within-concepts analyses of Evaluation scales resulted in three different factors—functional evaluation, hedonistic evaluation, and ethical evaluation. It is suggested that the three different clusters of scales may have arose because of differences between subjects in the kinds of Evaluation judgments which are most meaningful (or, in other words, in the kinds of scales which are most relevant). Some subjects may be functionalists and evaluate primarily in terms of valuable, beneficial, etc.; others, the hedonists, evaluate in terms of pleasant, attractive, etc.; and others, the moralists, evaluate in terms of clean, honest, sincere, etc. Now suppose a concept like *woman* is being rated. The functionalists with low evaluation would rate *woman* slightly valuable and beneficial and neutral on all else since other scales are irrelevant to functionalists; functionalists with high evaluation of *woman* would rate the concept extremely valuable and beneficial, and again, neutral on other scales. Thus, valuable and beneficial would correlate with each other but not with the other scales. The same thing would hold true for the scales favored by hedonists and moralists. If the relevance principle operated in the exaggerated fashion indicated here, we would get out three independent dimensions; if relevance of scales were not quite so clear-cut, we still would get three factors, but they would be somewhat correlated with one another. Furthermore, in calculating correlations over means for concepts rather than individual ratings, these three factors would tend to merge if functionalists, hedonists, and moralists generally agree in their net evaluation of things, that is, if all think *woman* is nice, but in dif-

ferent ways. The mean rating of *woman* would be moderate on all the evaluation scales if the ratings by functionalists, hedonists, and moralists were averaged together. Similarly, mean ratings on a concept like *enemy* probably would be low on all the evaluation scales. Therefore, the evaluation scales in the cross-concept analyses would tend to correlate with one another.

The assumption here is that there are differences between subjects in the types of evaluations which are meaningful (one might call these differences in value orientations), and they may be reflected in within-concepts analyses. Such value-orientation factors reflect basic differences between subjects, but they do not constitute concept-scale interaction since they are dependent on subjects rather than on concepts. If one wanted to do within-concepts analyses without having this matter enter in, one would need to divide the subject sample into homogeneous subgroups on the basis of value orientations and run analyses within subgroups.

The evidence for this type of effect is mostly indirect and the comments here are speculative. Furthermore, it is not known whether Potency and Activity scales also might fragmentize in this manner.

True concept-scale interaction. Three conditions now have been identified that could produce unique factor structures when analyzing certain concepts or classes of concepts and none of these constitute "true" concept-scale interaction. There are two conditions that could give rise to real concept-scale interaction.

Concept-scale interaction can arise because a scale has different degrees of relevance for different concepts. For example, sweet-sour may be highly relevant in evaluating food, moderately relevant in evaluating people, and of low relevance in evaluating abstract ideas. The amount of meaningful variation in ratings is proportional to relevance and, in practice, therefore, there would be little meaningful variation in sweet-sour ratings of abstract ideas. Thus, in rating this class of concepts, the sweet-sour scale would show little relation to any other scale and could not have its customary high loading on Evaluation. Relevance thus produces concept-scale interaction in the following sense. If a scale is irrelevant to a concept or to a class of concepts, ratings on it may have low communality with other scale ratings so the scale drops out of its usual factor location—it measures nothing.

Concept-scale interaction also can arise due to semantic shifts in the scale adjectives which develop because of the environment provided by a concept. Osgood, et al. (1957) noted that

> *sharp* as applied to concepts like ME and AMERICA has a dynamic favorable meaning (the slang usage) and correlates highly with scales like successful, intentional, and progressive; on the other hand, *sharp* as applied to concepts like BOULDER and KNIFE has its ordinary denotative meaning and correlates with scales like angular and rough [pp. 178–179].

This seems to be a real case of concept-scale interaction—the meanings of the scale words change depending on the environment provided by the concept, and since the meanings are different, the scale's factorial composition may be also. Another dramatic instance is provided by a study in which colors, forms, and words were rated on the same SD scales by samples of Americans and Japanese. Correlations between scales were calculated within each stimulus class and with languages; then, the tables of correlations were themselves correlated across stimulus class and across languages. The relations between scales were more similar across language holding constant the class of stimuli than across stimuli classes holding language constant. This indicates that the scales were used quite differently in rating the three classes of stimuli (reviewed in Miron & Osgood, 1966).

The existence or possible existence of concept-scale interaction, whether it is a function of relevance or stimulus environment, means that an SD ideally should be validated and adjusted for every new stimulus class with which it is used. A generalized SD, using standard scales like good-bad, powerful-powerless, and fast-slow, certainly is useful for rough and ready measurements, but more precise measurements will be attained only by tailoring instruments to each content domain so as to control for true concept-scale interactions.

Ad Hoc Factor Analyses

There are several reasons for carrying out new factor analyses of SD data. First, the validity of the basic EPA structure for averaged ratings always is subject to renewed test using novel scales and stimuli in different populations and cultures. Second, with the recent cumulation of evidence favoring individual differences in semantic space, there are compelling reasons for exploring the structure of individuals' SD ratings. It appears that at the individual level, studies will help to define the maximum dimensionality of the semantic space and when carried out ambitiously, such studies also will serve to define different varieties of cognitive-affective structures and the relationship between these structures and other personality variables. Third, the existence of true concept-scale interaction means that a variety of SD instruments must be available in order to obtain precise measurements in different content or stimuli domains, and factor analyses are required to develop instruments containing specially tailored scales.

The following discussion deals with some of the requirements that must be met in studies aimed at extending the SD to new content domains. This application of factor analysis to SD data is of general interest and is economically feasible for most investigators. Specific requirements for factor analyses intended as validity studies or those aimed at exploring individual structures are not considered here. However, it is to be noted that such studies often call for considerable resources and a high level of methodological sophistication. Basic methodological issues in validity studies are discussed by Osgood (1964) and Jakobovits (1966), and Kelly (1955) provides a rich store of information on the problems of investigating individual structures.

The general procedure for extending the SD to a new content domain is to have a sample of subjects use selected scales to rate concepts from the content area; then the data are factor analyzed to determine the underlying dimensions and the factor loadings of each scale on each dimension. However, to get valid and meaningful results, one must employ a rather structured study design. Only by taking care in the procedures can one avoid the various artifactual results discussed under concept-scale interaction.

One should follow the usual rule for test-development studies of drawing subjects from the same population as will be of interest in later work. The existence of individual differences in semantic space suggests that populations might differ, so factor structure based on one population may not be generalizable to other populations.

An unfortunate choice of concepts can lead to completely misleading factor-analytic results, and to avoid the problem, one must have concepts which represent as nearly as possible the entire SD space. This means that, at least, one should have concepts representing the eight combinations of the basic EPA dimensions: E + P + A +, E + P + A −, E + P − A +, E + P − A −, E − P + A +, E − P + A+, E − P −A +, and E − P − A−. The total number of concepts to be considered depends on the study design chosen. Here there are two major alternatives.

Design 1. The approach which avoids most artifacts and which yields the dimensions of variation among average ratings is that in which the mean ratings for concepts are the units of observation for correlation and factor analyses. In this case, a rather large number of concepts must be considered to get stable results. Although analyses have been conducted with fewer concepts, 40 appears to be a reasonable lower bound; this would allow five concepts from each octant of the basic SD space. Since the emphasis is on group means and these are known to develop stability with rather small sample sizes, as few as 15 or 20 subjects rating each concept might be feasible.

Design 2. There are cases when it would be very difficult to find 40 concepts in a particular content area which distribute around the SD space. In such a case, another approach might be taken. One could use a smaller number of concepts (still well distributed in the space) and a larger number of raters and calculate correlations over the individual ratings of the concepts, pooling ratings for all concepts. If the concepts are chosen carefully, it is possible to analyze the variance due to concepts and the variance

due to subjects together (Kahneman, 1963). ("Carefully" means that the concepts are balanced so that exaggeration and attenuation errors tend to cancel one another out.) In this approach, the set of ratings for one concept given by one subject is treated as one observation, and correlations are calculated over a total of $m \times n$ observations where m is the number of concepts and n is the number of subjects. It should be noted that this approach is very likely to lead to fragmentations of the basic EPA dimensions corresponding to individual variations in semantic space; thus, there ordinarily will be more than three significant factors. However, it can be anticipated that the smaller factors have low relevance for many subjects, so it might be reasonable to ignore them if one is interested only in basic EPA dimensions having high commonality.

With either design, one gathers ratings of the concepts on the scales of interest plus a number of reference scales chosen for their purity of loading on the EPA dimensions as indicated in previous studies. The data then are factored and the factors rotated to an orthogonal solution with the axes passing as nearly as possible through the reference scales. This procedure gives the factorial composition of the new scales in terms of the standard EPA structure.

As an illustration, suppose that one wanted to develop an SD specifically for the study of social roles and identities. The first step would be to turn to the literature and to the subject population of interest to determine what kinds of contrasts appear to be meaningful in this content domain. Presumably a large number of such contrasts could be developed. One then would prune these down, retaining scales which are clearest in meaning, most relevant, and which appear to represent the full EPA structure plus other dimensions that are supposed to be of interest. One might reasonably end up with 30 new scales for the study. Added to these would be some basic reference scales, say 3 for each EPA dimension giving a total set of 39 scales.

The next step would be choosing a balanced set of concepts; it is assumed here that Design 2 is used so relatively few concepts are required. To obtain concepts which are distributed throughout the SD space, one can employ either of two tactics. One could choose a large number of concepts which subjectively seem to represent various regions of the space and verify one's opinions by a small pilot study using standard scales and few subjects. Alternatively, one could use available SD dictionaries and other literature in which the ratings of concepts have been reported and try to draw from these sources a set of concepts which are balanced and well distributed. Taking the latter approach (using the dictionary of Snider and Osgood, 1969, Appendix), one gets the following set of role concepts which roughly represent the eight combinations of EPA dimensions: soldier (+++), grandfather (++−), child (+−+), widow (+−−), thief (−++), devil (−+−), homosexual (−−+), and beggar (−−−). This gives only one concept per octant of the SD space, but in Design 2, it is assumed that individual variations in the perceptions of these concepts will help distribute observations more evenly through the space. Step 3 is to have subjects rate the eight concepts on the 39 scales, and for this phase of the program, one would require a minimum of 50 different subjects working for somewhat less than 1 hour. The results of the study at this stage are an $8 \times 39 \times 50$ cube of data (concepts by scales by subjects) or a total of 15,600 measurements to be entered into factor analyses. The analyses themselves are conducted over the individual ratings of concepts so that each correlation between scales would be based on 8 times 50 or 400 observations.

When it comes to ad hoc factor analyses, it is clear that even the absolute minimal study is a fairly formidable undertaking. However, studies which proceed with a less than adequate design frequently may as well not be conducted at all since the results are liable to be distorted by methodological artifacts and thus misleading.

Conclusions

John Carroll (1959a) concluded his critical review of *The Measurement of Meaning* with the sentence: "Nevertheless, the reviewer is

inclined to characterize the book by asserting: it is *good*, it is *active*, it is *potent* [p. 77]." The "successful" profile for the SD still seems warranted after more than 10 years of additional studies and applications. The SD has become a standard and useful tool for social psychological research.

There is probably no social psychological principle that has received such resounding cross-group and cross-cultural verification as the EPA structure of SD ratings. Furthermore, few traditions of research are associated with comparable productivity or with the richness of findings that has developed in SD applications. On the side of frustration, we have seen that the metric of bipolar scales is understudied; scale-checking styles exist making puzzles of some findings; the exact number of dimensions is still an unsettled issue; some individual variations do exist in EPA structure; and concept-scale interaction still stands as a mysterious specter, often less important than it seems, but too important to ignore.

One of the basic points that emerged in this review was that factor analyses of ratings for only one or a few concepts are highly wrought with hazards. Tailoring the SD to a new content area requires a rigorous research design, and inexpensive substitutes can yield instruments which are distorted and, therefore, worse than worthless.

REFERENCES

ARTHUR, A. Z. Clinical use of the semantic differential. *Journal of Clinical Psychology*, 1965, 21, 337–338.

ARTHUR, A. Z. Response bias in the semantic differential. *British Journal of Social and Clinical Psychology*, 1966, 5, 103–107.

BEITNER, M. S. Word meaning and sexual identification in paranoid schizophrenics and anxiety neurotics. *Journal of Abnormal and Social Psychology*, 1961, 63, 289–293.

BORGATTA, E. F. The structure of personality characteristics. *Behavioral Science*, 1964, 9, 8–17.

BURKE, R. L., & BENNIS, W. G. Changes in perception of self and others during human relations training. *Human Relations*, 1961, 14, 165–182.

CARROLL, J. B. Review of *The measurement of meaning*. *Language*, 1959, 35, 58–77. (a)

CARROLL, J. B. Some cautionary notes on the semantic differential. Paper presented at the meeting of the American Psychological Association, Cincinnati, September 1959 (b)

CLIFF, N. Adverbs as multipliers. *Psychological Review*, 1959, 66, 27–44.

CRONBACH, L. J. Coefficient alpha and the internal structure of tests. *Psychometrika*, 1951, 16, 297–334.

DEESE, J. The associative structure of some common English adjectives. *Journal of Verbal Learning and Verbal Behavior*, 1964, 3, 347–357.

DEESE, J. *The structure of associations in language and thought*. Baltimore: Johns Hopkins Press, 1965.

DIVESTA, F. J. A developmental study of the semantic structures of children. *Journal of Verbal Learning and Verbal Behavior*, 1966, 5, 249–259.

DIVESTA, F. J., & DICK, W. The test-retest reliability of children's ratings on the semantic differential. *Educational and Psychological Measurement*, 1966, 26, 605–616.

DIXON, T. R., & DIXON, J. F. The impression value of verbs. *Journal of Verbal Learning and Verbal Behavior*, 1964, 3, 161–165.

DOGAN, M., & ROKKAN, S. *Quantitative ecological analysis in the social sciences*. Cambridge: M.I.T. Press, 1969.

FORD, L. H., JR., & MEISELS, M. Social desirability and the semantic differential. *Educational and Psychological Measurement*, 1965, 25, 465–475.

GOLDFRIED, M. R., & KISSEL, S. Age as a variable in the connotative perceptions of some animal symbols. *Journal of Projective Techniques and Personality Assessment*, 1963, 27, 171–180.

GOLLOB, H. F. Impression formation and word combination in sentences. *Journal of Personality and Social Psychology*, 1968, 10, 341–353.

GREEN, R. F., & GOLDFRIED, M. R. On the bipolarity of semantic space. *Psychological Monographs*, 1965, 79(6, Whole No. 599).

HARRIS, C. W. (Ed.) *Problems in measuring change*. Madison: University of Wisconsin Press, 1963.

HEISE, D. R. Semantic differential profiles for 1,000 most frequent English words. *Psychological Monographs*, 1965, 79(8, Whole No. 601).

HEISE, D. R. Affectual dynamics in simple sentences. *Journal of Personality and Social Psychology*, 1969, 11, 204–213. (a)

HEISE, D. R. Separating reliability and stability in test-retest correlations. *American Sociological Review*, 1969, 34, 93–101. (b)

HEISE, D. R. The semantic differential and attitude research. In G. Summers (Ed.), *Attitude measurement*. Chicago: Rand McNally, 1970.

HOWE, E. S. Probabilistic adverbial qualifications of adjectives. *Journal of Verbal Learning and Verbal Behavior*, 1962, 1, 225–242.

HOWE, E. S. Associative structure of quantifiers. *Journal of Verbal Learning and Verbal Behavior*, 1966, 5, 156–162. (a)

HOWE, E. S. Verb tense, negatives, and other determinants of the intensity of evaluative meaning. *Journal of Verbal Learning and Verbal Behavior*, 1966, 5, 147–155. (b)

JAKOBOVITS, L. A. Comparative psycholinguistics in the study of cultures. *International Journal of Psychology*, 1966, 1, 15–37.

KAHNEMAN, D. The semantic differential and the structure of inferences among attributes. *American Journal of Psychology*, 1963, 76, 554–567.

KELLY, G. A. *The psychology of personal constructs*. New York: Norton, 1955.

KOMORITA, S. S., & BASS, A. R. Attitude differentiation and evaluative scales of the semantic differential. *Journal of Personality and Social Psychology*, 1967, 6, 241–244.

LURIA, Z. A semantic analysis of a normal and a neurotic therapy group. *Journal of Abnormal and Social Psychology*, 1959, 58, 216–220.

MALTZ, H. E. Ontogenetic change in the meaning of concepts as measured by the semantic differential. *Child Development*, 1963, 34, 667–674.

MESSER, S. JAKOBOVITS, L. A., KANUNGO, R., & LAMBERT, W. E. Semantic satiation of words and numbers. *British Journal of Psychology*, 1964, 55, 155–163.

MESSICK, S. J. Metric properties of the semantic differential. *Educational and Psychological Measurement*, 1957, 17, 200–206.

MILLER, G. A. The magical number seven, plus or minus two. *Psychological Review*, 1956, 63, 81–97.

MIRON, M. S. A cross-linguistic investigation of phonetic symbolism. *Journal of Abnormal and Social Psychology*, 1961, 62, 623–630.

MIRON, M. S., & OSGOOD, C. E. Language behavior: The multivariate structure of qualification. In R. B. Cattell (Ed.), *Handbook of multivariate experimental psychology*. Chicago: Rand McNally, 1966.

MITSOS, S. B. Personal constructs and the semantic differential. *Journal of Abnormal and Social Psychology*, 1961, 62, 433–434.

MOGAR, R. E. Three versions of the F scale and performance on the semantic differential. *Journal of Abnormal and Social Psychology*, 1960, 60, 262–265.

MORDKOFF, A. M. An empirical test of the functional antonymy of semantic differential scales. *Journal of Verbal Learning and Verbal Behavior*, 1963, 2, 504–508.

MORDKOFF, A. M. Functional vs. nominal antonymy in semantic differential scales. *Psychological Reports*, 1965, 16, 691–692.

NEURINGER, C. Effect of intellectual level and neuropsychiatric status on the diversity of intensity of semantic differential ratings. *Journal of Consulting Psychology*, 1963, 27, 280.

NICKOLS, S. A., & SHAW, M. E. Saliency and two measures of attitude. *Psychological Reports*, 1964, 14, 273–274.

NORMAN, W. T. Stability-characteristics of the semantic differential. *American Journal of Psychology*, 1959, 72, 581–584.

NORMAN, W. T. Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings. *Journal of Abnormal and Social Psychology*, 1963, 66, 574–583.

OSGOOD, C. E. Studies on the generality of affective meaning systems. *American Psychologist*, 1962, 17, 10–28.

OSGOOD, C. E. Semantic differential technique in the comparative study of cultures. *American Anthropologist*, 1964, 66(3, Pt. 2), 171–200.

OSGOOD, C. E., SUCI, G. J., & TANNENBAUM, P. H. *The measurement of meaning*. Urbana: University of Illinois Press, 1957.

PEABODY, D. Two components in bi-polar scales: Direction and extremeness. *Psychological Review*, 1962, 69, 65–73.

SMITH, R. G. A semantic differential for theatre concepts. *Speech Monographs*, 1961, 28, 1–8.

SNIDER, J. G., & OSGOOD, C. E. (Eds.) *Semantic differential technique: A sourcebook*. Chicago: Aldine, 1969.

STRICKER, G., & ZAX, M. Intelligence and semantic differential discriminability. *Psychological Reports*, 1966, 18, 775–778.

SUCI, G. J. A comparison of semantic structures in American Southwest culture groups. *Journal of Abnormal and Social Psychology*, 1960, 61, 25–30.

WELLS, W. D., & SMITH, G. Four semantic rating scales compared. *Journal of Applied Psychology*, 1960, 44, 393–397.

WIGGINS, N., & FISHBEIN, M. Dimensions of semantic space: A problem of individual differences. In J. G. Snider & C. E. Osgood (Eds.), *Semantic differential technique: A sourcebook*. Chicago: Aldine, 1969.

WRIGHT, B. A semantic differential and how to use it. Chicago: Social Research, Inc., 1958. (Mimeo)

ZAX, M., GARDINER, D. H., & LOWY, D. G. Extreme response tendency as a function of emotional adjustment. *Journal of Abnormal and Social Psychology*, 1964, 69, 654–657.

(Received February 7, 1969)