An Examination of Methodological Issues Relevant to the Use and Interpretation of the Semantic Differential

Irene T. Mann
Michigan State University

James L. Phillips
Oklahoma State University

Eileen G. Thompson
Michigan State University

A number of methodological issues have been raised regarding the semantic differential technique. This study re-examined several key problems, particularly the assumed bipolarity of scales, instructions regarding use of the midpoint, and concept-scale interaction, all of which may contribute to a lack of precision in the technique. In addition, this study utilized an analysis of variance model to partition variance in semantic differential ratings. Forty subjects responded to one of four instruments on two occasions. Instruments differed in terms of polarity type (bipolar or unipolar) and presence or absence of an irrelevance option. Twenty-four concepts uniformly distributed throughout semantic space were judged on either 15 or 30 scales. Results indicated that the Evaluation-Potency-Activity (EPA) structure was both robust and reliable. However, several features of the data argued for caution in the use of the semantic differential technique. Both the Concept × Scale interaction and the Scale × Concept × Person interaction accounted for substantial proportions of variance in semantic differential ratings. Suggestions were offered to minimize such effects.

Since its introduction by Osgood, Suci, and Tannenbaum (1957), the semantic differential has been widely used by psychological investigators. At the same time, use of the instrument has generated many methodological questions concerning its validity. Recent reviews (Heise, 1969; Osgood, May, & Miron, 1975) have isolated and addressed several key issues. These issues provide the basis for the present study.

One issue which has been dealt with extensively in the literature is the assumption that semantic space is bipolar. The bipolarity assumption requires every scale to be end-anchored by a pair of adjectives that are antonyms or that *function* as antonyms in the context of the rating task. To the extent that this is not true, interpretation of the instrument is subject to some question.

Related to the bipolarity assumption is the practice of assigning multiple meanings to the middle response category. The subject who chooses this category may do so, according to the standard instructions, to indicate neutrality, ambivalence, or irrelevance. There are two issues here. Does the confounding of meanings for the midpoint have some disconcerting, confusing, or distorting effect on the person performing the rating and hence on the rating itself? Does this confounding have a distorting effect on the resulting factor structure, independent of its effects on the rater?

Another problem has been the extent to which a given scale undergoes changes in meaning from one concept to another. This context effect has been labeled Concept × Scale interaction. Although it is acknowledged as problematic (cf.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 3, No. 2 Spring 1979 pp. 213–229

Osgood, May, & Miron, 1975), there have been few attempts to quantify it or to assess the degree to which it may bias results. Further, although differential meanings of concepts from person to person are to be expected and in no way present a problem for the technique, the same is not true for the differential meaning of scales across persons. The possibility of a Scale × Person interaction is thus as troublesome as a Concept × Scale interaction.

Bipolarity Assumption

Osgood, Suci, and Tannenbaum (1957) assumed that a scale corresponds to a line in the space defined by Evaluative, Potency, and Activity dimensions passing through the origin of that space. This line is presumably anchored by polar terms which are opposite in meaning and equidistant from the origin. A number of studies have examined this assumption and have supported it (Taylor & Kumata, cited in Osgood et al., 1957; Vidali, 1973; Vidali & Holeway, 1975). These studies showed that combinations of unipolar judgments corresponded to bipolar judgments.

Green and Goldfried (1965), however, examined correlations between pairs of adjectives assumed to be opposites. They found little support for a generalized bipolar model of semantic space. The correlation between unipolar ratings for a given scale varied from concept to concept.

Other researchers have asked subjects to rate the scale anchors themselves on other semantic differential scales. The symmetry of the scale profile of the assumed opposites around the midpoint of these other scales serves as a bipolarity index. Using this procedure, Mordkoff (1963, 1965) found that only 14 of the 28 pairs of terms were indeed symmetric. Osgood, May, and Miron (1975), however, reported very small deviations from the midpoint for the scale ratings of assumed opposites. Other research that examined bipolarity using methodologies other than the semantic differential (Carter, Ruggels, & Chaffee, 1968; Ross & Levy, 1960; Terwilliger, 1962) has not supported the bipolarity assumption.

Thus, the evidence for the bipolarity assumption has been mixed. If it is accepted that the bipolarity assumption holds, at best, only approximately, what are the practical consequences? That is, to what extent are the patterns of ratings in the data distorted by the invalidity of the assumption?

The Midpoint

Conventional instructions inform the subject that the midpoint is to be checked if the concept is neutral on the scale, if the scale is irrelevant or unrelated to the concept, or if both sides are equally associated with the concept. In terms of the conception of a three-dimensional semantic space, neither the irrelevance response nor the equal association response should be made by subjects. All dimensions and all scales should be relevant for any concept. Also, when both sides of the scale are seen as appropriate, an "averaged-out" or ambivalent response is implied, and the scale cannot operate in a linear fashion.

There is evidence that subjects use the scale midpoint to indicate ambivalence or irrelevance. Carter, Ruggels, and Chaffee (1968) and Oetting (1967) have shown that subjects given the option of indicating irrelevance do so. Oetting also found that when subjects repeated the task with no irrelevance option, the majority used the midpoint for those ratings which previously had been marked as irrelevant. Davis (1972) found that subjects gave low appropriateness ratings to certain concept-scale combinations which they had rated as irrelevant.

Only one study, Forthman (1973), has examined the effect of permitting subjects to differentiate among the meanings of the midpoint on the resulting factor structure. This study suggests that such a procedure produces results discrepant from the Evaluation-Potency-Activity (EPA) structure. Clearly, this warrants further study.

Concept × Scale Interaction

According to Osgood et al. (1957), "the meanings of scales and their relations to other

scales vary considerably with the concept being judged" (p. 187). Heise (1969) has also noted the existence of Concept × Scale interaction. One empirical indication of Concept × Scale interaction occurs when scales do not have their usual loadings on the EPA dimensions for particular sets of concepts.

Several studies using particular classes of concepts have failed to find the typical EPA structure and may be interpreted as evidence for Concept × Scale interaction (Osgood, Ware, & Morris, 1961; Komorita & Bass, 1967). A number of researchers have demonstrated Concept × Scale interaction by showing that within a given class of concepts, scale loadings or factor structures differ (Bynner & Romney, 1972; Heskin, Bolton, & Smith, 1973; Kubiniec & Farr, 1971; Presley, 1969; Rosenbaum, Rosenbaum, & McGinnies, 1971). A recent study by Mayerberg and Bean (1978) demonstrated Concept × Scale interaction by showing that the pattern of scale combinations to form various factors differs across different concept domains.

Researchers have also attempted to identify classes of concepts and classes of scales which are likely to produce Concept × Scale interaction. Smith and Nichols (1973) and Nichols and Smith (1973) found that when scales requiring value judgments were applied to concrete objects or when a scale which referred to an objective characteristic was applied to a nonphysical construct, midpoint responding and factor instability were likely to occur.

Concept × Scale interaction also has implications for the computation of factor scores. Bynner and Romney (1972), Clark and Kerrick (1967), Heaps (1972), and Presley (1969) agree that Concept × Scale interaction makes factor loadings from an overall analysis inappropriate for the computation of factor scores; individual concept analyses are necessary. This has implications for applications such as attitude assessment. Quantification of the magnitude of this effect is important, given the wide use of the semantic differential for this type of assessment.

Concept selection and scale selection are troublesome in regard to Concept × Scale interaction. Concept × Scale interaction can either be viewed as an unwelcome artifact (to be minimized) or as something that validates the technique (see Osgood, May, & Miron, 1975, p. 349) but which may require special attention. Much depends on the researcher's aims.

Individual Differences in Semantic Differential Responses

Research has shown that there are differences among persons in the ways in which they use scales and that different scales have different degrees of relevance. The existence of individual differences has been shown by investigations of differences between groups (Krieger, 1963; Lloyd & Innes, 1969; Warr, Schroder, & Blackman, 1969). Crockett and Nidorf (1967) performed separate factor analyses for each subject; they found that the typical three-dimensional semantic space did not adequately characterize the judgments of all the subjects.

Studies using three-mode factor analysis have also provided evidence for individual differences in scale use (Shikiar, Fishbein, & Wiggins, 1974; Snyder & Wiggins, 1970; Wiggins & Fishbein, 1969). The presence of more than one subject factor indicates the existence of individual differences with respect to the number of dimensions, interpretation of these dimensions, and the relative importance placed on them.

Individual differences may be reflected in Scale × Person, Concept × Person and Scale × Concept × Person interactions. Does one treat these individual differences as error variance and collapse across persons when analyzing semantic differential data or does one attempt to analyze and to explain these differences? The decision would seem to depend on how much of the variability in responses is accounted for by these interactions. If this is low relative to the variance accounted for by concepts and by scales, then the interactions can be reasonably ignored. Information to answer this question has not been available. What is required is an analytic approach such as the one discussed in the next section.

Components of Variance in Semantic Differential Responses

The methodological issues addressed above suggest potential artifacts or contaminants of the EPA structure. It is of interest to quantify the extent of these effects by examining the proportion of variance accounted for by various factors in the semantic differential design.

Maguire (1973) has proposed that semantic differential data be examined using an ANOVA model in which the factors are concepts, scales, and persons. A similar analysis, but with different factors, was performed by Kahneman (1963). Such an analysis has the advantage of isolating such characteristics of the data as Concept × Scale interaction and Scale × Person interaction.

The Concept × Scale interaction in Maguire's model does not correspond conceptually to the effect that has been labeled Concept × Scale interaction. If a set of concepts are ordered in one way by a particular scale but in a quite different way by another scale, this would contribute to the magnitude of a Concept × Scale interaction effect in the ANOVA model. But if the scales were dissimilar, e.g., one an Evaluative scale and one a Potency scale, this differential ordering is just what would be expected from the three-dimensional EPA factor structure. Alternatively, if the ANOVA model were to be applied separately for Evaluative, Potency, and Activity scales, then the magnitude of the Concept × Scale interaction effect would accord with the usual meaning of Concept × Scale interaction and would be a measure of its magnitude.

The ANOVA also allows for statistical tests of a number of effects. The error term for such effects in the model as given by Maguire was the Scale × Concept × Person interaction. However, as noted above, this three-way interaction is in itself of theoretical interest. The Maguire model can be extended to include a fourth factor: occasions. While such an occasions factor is of no importance, it functions as a replication factor in the design. Through its interactions with other factors, it provides the bases for appropriate error terms and allows for an assessment of the test-retest reliability of the instrument.

Method

Subjects

The subjects were 40 Michigan State University undergraduates enrolled in elementary psychology classes who participated for course credit. Students were randomly assigned to experimental conditions. A time for each student's second session was randomly selected from that student's free times during the two weeks following the first session. Students participated individually on the task.

Design

Students were presented with one of four semantic differential instruments, created by crossing presence or absence of an irrelevance option with the use of unipolar or bipolar scales. The students repeated the semantic differential ratings on a second occasion. The overall design was thus a 2 (presence or absence of the irrelevance option) × 2 (unipolar or bipolar) × 24 (concepts) × 3 (dimensions) × 5 (scales within a dimension) × 10 (persons) × 2 (occasions) design.

Apparatus

The instructions and the instruments were presented on a Hewlett-Packard 2640 video display terminal connected to a Hewlett-Packard 2000 mini-computer. Students responded by striking the appropriate key on the terminal keyboard. The presentation of instructions and items as well as the recording of the responses was accomplished using the program PROCTR (Price, 1977; Thompson, 1977).

Procedure

Upon arriving at the laboratory, students were told that the task involved judgments of the meaning of various words. The operation of the terminal was briefly explained to them. Each student was presented with instructions appropriate for the assigned condition. In the condition with bipolar scales and no irrelevant re-

sponse option, the instructions were identical to the standard semantic differential instructions (Osgood, Suci, & Tannenbaum, 1957). In those conditions which incorporated the irrelevant response option, the instructions were modified to explain that students were to use the numeral "0" to indicate that a scale was not relevant to a particular concept. Although seated at some distance from the student, the experimenter was present on every occasion to answer any questions.

Scales were presented in a constant order for all conditions. Concepts were presented in one of five random orders. In both of the bipolar conditions, each student rated each of 24 concepts on 15 scales. In the unipolar scale conditions, each student rated all 24 concepts on 30 scales. That is, the bipolar scale tough-fragile became two unipolar scales, tough-the opposite of tough, and fragile-the opposite of fragile. In these conditions, the example presented in the instructions was in the unipolar form. The position of the two anchors was randomly determined.

The 15 scales were randomly selected from lists of scales previously used in semantic differential research, with the constraint that 5 be primarily Evaluative scales, 5 be primarily Potency scales, and 5 be primarily Activity scales. Concepts were randomly chosen from the list of 366 concepts previously used by Jenkins, Russell, and Suci (1958), with the constraint that they be uniformly distributed over the Evaluative, Potency, and Activity dimensions. The concepts used were *ocean, dream, chair, boulder, rage, dreariness, fine, discomfort, kitchen, dawn, criminal, snail, comfort, trees, truth, sex, stench, coal, music, moon, puppies, sickness, candy,* and *youngster.*

Results

Factor Structure

Factor analysis of the data for all instrument types and for both occasions provided strong support for an EPA structure. For these analyses, the ratings for the two opposite unipolar scales were collapsed by averaging the rating on one scale with the reflected rating on the other. For the irrelevance-option instruments, 0's were converted to 4's (the scale midpoint). The factor analyses used principal components analyses[1] and were performed using, in turn, the three procedures which have been employed most frequently for such data. These were (1) the stringing-out procedures, treating each Concept × Person combination as a unique observation; (2) the summation procedures, in which all students' ratings of a given concept on a given scale were summed and averaged; and (3) the average correlation method, in which the interscale correlations were computed separately for each student and a matrix of average interscale correlations was obtained by averaging all students' correlations, using r to z transformations.

Despite the fact that this latter procedure induced a non-Gramian matrix, with the factor analysis containing negative eigenvalues, the rotated factor structures for all three procedures were comparable. Rotation of three factors in each case resulted in a highly interpretable EPA structure. Correlations of factor loadings for the appropriate factors across the solutions for the three procedures were quite high, ranging from .78 to .97 over all instrument types and occasions. Future reference to aspects of the factor structure in this paper refer to the results obtained with the stringing-out procedure.

For every combination of instrument type and occasion, three factors accounted for more than 58% of the variance. Varimax rotation of these three factors uniformly resulted in an EPA structure, with all scales loading on the expected factor.

Factor analyses of individuals' ratings used the 15 × 24 scale × concept matrix, with ratings on the two opposite scales collapsed for the unipolar instruments. The first three factors accounted for an average of about 65% of the variance for the individuals. Rotation of three

[1]Use of the principal factors methods involving estimation of communalities, resulted in a virtually identical factor structure for these data.

factors resulted in a well-defined EPA structure for 85% of the students receiving bipolar scales with the irrelevance option, 70% of the students receiving unipolar scales with standard instruction, and 65% of the students receiving unipolar scales with the irrelevance option.

Bipolarity of Scales

The most straightforward analysis of bipolarity consists of correlating the ratings on one scale in the unipolar instrument with those on its presumed semantic opposite, across subjects and concepts. These correlations for both the standard instructions and the irrelevance option instructions, for both first and second occasions, are presented in Table 1. An inspection of these results would support the assertion that these scales are but only approximately bipolar. This would be the mandatory conclusion were the criterion of a perfect negative correlation between the appropriate unipolar scales to be adopted as demonstrating bipolarity. Such a criterion would be appropriate, however, only if both unipolar scales were perfectly reliable.

Since each scale was administered on two separate occasions, it was possible to compute test-retest reliabilities for each scale. A measure of the maximum expected correlation between a pair of unipolar scales is the square root of the product of the two reliabilities. To obtain a single measure of bipolarity for this pair of scales, the average correlation across occasions was obtained using an r to z transformation, and this correlation was used to form the ratio to the maximum expected correlation. If the hypothesis of bipolarity were true, then this ratio would be uniformly close to negative one. The obtained results supported the bipolarity hypothesis quite strongly. For the instrument with the irrelevance option, the mean value of this ratio over 15 bipolar scales was $-.787$ when 0's were coded as 4's; and $-.845$, when 0's were treated as missing data. The median values of this ratio were $-.966$ and $-.975$, respectively. Similar results were obtained with standard instructions concerning the meaning of the midpoint. Here the mean was $-.878$ and the median was $-.929$.

This strong evidence for bipolarity was illustrated using a different analytic procedure which did not incorporate information about scale reliability. Factor analyses were performed on the unipolar scale instruments using the stringing-out procedure. For the instrument

Table 1
Correlations Between Opposite Unipolar Scales for Unipolar Instruments

	Instrument with Irrelevance option				Instrument with Standard option	
Scale	Occasion 1 0=4	Occasion 2 0=4	Occasion 1 0 out	Occasion 2 0 out	Occasion 1	Occasion 2
good-bad	-.74	-.83	-.76	-.84	-.68	-.68
valuable-worthless	-.77	-.77	-.79	-.78	-.69	-.57
beautiful-ugly	-.71	-.73	-.74	-.76	-.56	-.69
wise-foolish	-.47	-.51	-.50	-.61	-.43	-.28
kind-cruel	-.57	-.65	-.64	-.75	-.56	-.46
hard-soft	-.44	-.64	-.46	-.67	-.46	-.36
powerful-powerless	-.75	-.76	-.78	-.77	-.41	-.47
strong-weak	-.73	-.64	-.74	-.69	-.60	-.46
rugged-delicate	-.46	-.26	-.53	-.30	-.40	-.16
tough-fragile	-.38	-.27	-.43	-.32	-.35	-.27
active-passive	-.57	-.51	-.62	-.57	-.59	-.57
fast-slow	-.65	-.73	-.67	-.75	-.39	-.33
tense-relaxed	-.55	-.61	-.59	-.67	-.50	-.34
excitable-calm	-.43	-.55	-.48	-.62	-.36	-.09
hot-cold	-.32	-.48	-.36	-.52	-.20	-.28

with the irrelevance option, each scale loaded on the same factor as its designated opposite on both occasions. Under standard instructions, the "tense" and "relaxed" scales loaded on different factors on both occasions, and the "powerful" and "powerless" scales loaded on different factors on the first occasion only. Thus, appropriate scales acted as opposites in terms of the factor solutions for the unipolar scale instruments.

A further issue concerns the effect of the use of unipolar scales on the factor structure. When each unipolar scale rating was averaged with its opposite, the factor solution on the grouped data was similar to that obtained using the bipolar scales. For individual students, as noted above, the EPA structure emerged less clearly when the unipolar instruments were used.

The Midpoint-Irrelevance and Neutrality

As noted above, permitting students to indicate scale irrelevance with a 0 resulted in stronger bipolarity. Overall, however, the irrelevance option had little effect on the factor structures.

When the irrelevance option was available, it was used more often than the neutral (4) response. In addition, students receiving the irrelevance option instruments used either a 0 or a 4 much more often ($\bar{X} = 7.26$, bipolar; $\bar{X} = 9.32$, unipolar)[2] than standard instruction students used the multiple-meaning 4 response ($\bar{X} = 5.47$, bipolar; $\bar{X} = 5.75$, unipolar).

There was a uniform tendency to respond to a given scale with neutrality or irrelevance across type of instructions. The number of neutral (4) responses were computed for each scale in each instrument which did not provide for the irrelevance response, both for Occasion 1 and Occasion 2. Similarly, the numbers of neutral/irrelevant (0 and 4) responses were computed for each scale in each instrument which did provide for the irrelevance response. Pairs of instruments were selected and compared; the members of each pair differed only with respect to the presence or absence of the irrelevant response option. For each such pair, the correlation between neutral responses and neutral/irrelevant responses was computed over scales. There were four such correlations and they ranged from .79 to .91.

The number of 0 and 4 responses for each bipolar scale averaged across the two occasions is presented in Table 2. The scales showing the highest number of 4 responses under standard instructions were wise-foolish, kind-cruel, hard-soft, fast-slow, and hot-cold. These same scales together with the tough-fragile, excitable-calm, and tense-relaxed scales showed high levels of 0 responding in the irrelevance option conditions. The scales receiving the largest number of neutral responses under the irrelevance option instructions were hot-cold and hard-soft. Regarding the unipolar instruments, the scales which were least bipolar were those for which the level of irrelevant responding was highest.

Examination of the use of the irrelevant response in the Concept × Scale matrix indicated that several concepts received a large number of 0's. *Chair* and *boulder, kitchen, stench, coal,* and *candy* were in this group. This was particularly the case for the wise-foolish, tense-relaxed, and excitable-calm scales. It appeared that the combination of person-referent attributes and concrete objects presented problems for students' judgments. Applying physical attributes, such as hot-cold and fast-slow, to such abstract concepts as *truth* also resulted in irrelevant responses, though this pattern was less uniform.

The Analysis of Variance

The assessment of judgments at two points in time permitted the use of analysis of variance to examine the effects of scales, concepts, persons,

[2]These were the mean number of neutral/irrelevant responses, per scale per occasion, averaged across subjects and across concepts.

Table 2
Frequencies of Irrelevant (0) and Midpoint (4) Responses, Averaged across the two occasions

	Bipolar Instruments		
	With Irrelevance Option		With Standard Instructions
	0's	4's	4's
good-bad	9	25	31
valuable-worthless	10	17.5	17
beautiful-ugly	19	31	31.5
wise-foolish	83.5	22.5	105
kind-cruel	60	32	69
hard-soft	51	39	74.5
powerful-powerless	34.5	16.5	33
strong-weak	39	25.5	38.5
rugged-delicate	49	30	50.5
tough-fragile	61.5	30	54.5
active-passive	25	8.5	28
fast-slow	58	35	78.5
tense-relaxed	57.5	30	51
excitable-calm	49	33	55
hot-cold	51.5	56.5	104

occasions, and the interactions among these factors. The ANOVA design was a completely crossed four-factor Scale (15) × Concept (24) × Person (10) × Occasion (2) design, with one observation per cell. Each factor was treated as a random factor in this design. The overall ANOVA was performed separately for each instrument type. For the irrelevance option instrument, 0's were converted to 4's for this analysis. Judgments on the unipolar instruments were collapsed by averaging the ratings, appropriately reflected, for the two corresponding unipolar scales.

Because this was a completely randomized design, there were problems in testing the significance of a number of the factors. The four-way interaction of scale, concept, person, and occasion provided an appropriate error term for testing the three-way interactions. There was, however, no appropriate error term for testing the two-way interactions. If any of the three-way interactions were significant, the device of pooling error terms was clearly inappropriate, and quasi-*F* ratios provided the only means of testing other effects in the design. However, the concern of this study was primarily with the percent of variance accounted for by the factors in the design.

It has been noted that occasions had little effect on the results. The reliability of the semantic differential procedure may be examined by assessing the effect of occasions and the interactions containing occasions upon results. The occasions factor was significant only for the unipolar instrument with the irrelevance option. In all cases, the occasions factor accounted for .08% of the variance or less. All of the interactions involving occasions together accounted for only 11.2% to 18.6% of the variance, with the four-way interaction providing nearly all of this. It is clear, therefore, that in this study the semantic differential proved to be highly reliable for all instrument types.

As noted previously, two effects which are of particular interest cannot be examined directly with the overall ANOVA. These effects—the

Concept × Scale and the Scale × Concept × Person interactions—will be examined in the following section. With respect to other effects in the overall ANOVA, both the main effect for concepts and the main effect for scales[3] were significant and accounted for a substantial proportion of the variance.

The Concept × Scale and Scale × Concept × Person Interactions

In order to examine the effects of the Concept × Scale and the Scale × Concept × Person interactions, the ANOVA described above was applied separately for each of the three scale sets—Evaluation, Potency, and Activity—for each instrument type. The results of these ANOVAs were very similar across instrument type. To avoid redundancy, only the ANOVAs for Evaluation, Potency, and Activity scales for bipolar scales and standard instructions were included. These are summarized in Table 3. The concept effect was significant in every case, representing the fact that concepts were sampled from throughout semantic space. The main effect for scales was significant for all dimensions for all bipolar instruments and in all the unipolar evaluation analyses. For the Activity dimension and for Potency with standard instructions, the scales main effect was not significant with unipolar scales. Overall, the scales effect was relatively small in comparison to the concept main effect.

Among the two-way interactions, the Concept × Person, Scale × Person, and Concept × Scale interactions were of interest. The Concept × Person effect was significant in all cases, accounting for 11.5% to 20.2% of the variance of Evaluative scales, 11.0% to 16.5% of the variance of Activity scales, and 14.9% to 18.3% of the variance of Potency scales.

The Scale × Person interaction was significant in 11 of the 12 cases. For the unipolar irrelevance-option instrument and the Activity dimension, this effect was marginally significant. This interaction, while significant, accounted for only a small proportion of the variance, ranging from 1.3% to 4.4%.

The Concept × Scale interaction was significant for all three dimensions for all four instrument types. This interaction accounted for 5.5% to 9.8% of the variance for Evaluative scales, 11.4% to 15% of the variance for Activity scales, and 8.8% to 14.1% of the variance for Potency scales (see Table 4). It is thus clear that the Concept × Scale interaction is a relatively substantial factor in these results and is more marked for the Activity and Potency dimensions.

The Concept × Scale × Person interaction was highly significant in all cases. The percent of variance accounted for by this effect ranged from 10.0% to 13.6% for Evaluative scales, 18.4% to 24.2% for Activity scales, and 16.3% to 24.4% for Potency scales (see Table 5). The percent of variance accounted for by the Concept × Scale × Person interaction was greater than that accounted for by the Concept × Scale interaction in every case.

The Effect of Instrument Type

Analysis of variance procedures were also used to assess the significance and percentage of variance accounted for by the presence or absence of the irrelevance option and by the use of bipolar or unipolar scales (polarity type). These analyses were carried out separately for each EPA dimension for each occasion.[4] The result-

[3]The overall ANOVA was done four times, once for each of the four types of instruments. The main effect for scales was sigificant at $p < .05$ in three of the four ANOVAs and was marginally significant, $p < .08$, in the fourth.

[4]All of the ANOVAs reported in this study were performed using software available on the CDC6500 at Michigan State University. An omnibus ANOVA which incorporated the two between-subjects instrument factors and which treated scales as nested-within-scale type was appropriate and was planned. Unfortunately, software to perform such an analysis was not available due to core storage limitations. Thus, it was necessary to do the piecing together described.

Table 3

Analysis of Variance of Ratings For Instrument With Standard Instructions and Bipolar Evaluative Potency and Activity Scales

Source	df num	df den	MS	F
Evaluative Scales				
Concept	23	155	187.5478	15.2639*
Scale	4	93	58.7933	5.3742*
Person	9	87	19.7970	2.50212*
Occasion	1	1	4.1667	5.59329
CxS	92	180	8.7579	7.7928*
CxP	207	255	5.1035	3.8226*
CxO	23	101	.3649	.3829
SxP	36	82	3.4547	2.6406*
SxO	4	37	.6438	.6949
PxO	9	56	1.4768	1.2982
CxSxP	828		.9859	2.1155*
CxSxO	92		.6040	1.2959*
CxPxO	207		.8152	1.7492*
SxPxO	36		.7884	1.6918*
CxSxPxO	828		.4660	
Potency Scales				
Concept	23	144	106.4193	7.3799*
Scale	4	80	48.7569	3.7877*
Person	9	73	14.7285	2.1592
Occasion	1	5	.2017	.0975
CxS	92	218	10.8358	8.7668*
CxP	207	314	5.1501	3.0664*
CxO	23	75	.7382	.8139
SxP	36	89	2.6342	1.7639*
SxO	4	19	1.5215	2.1110
PxO	9	47	1.5989	1.3733
CxSxP	828	828	1.3969	1.7791*
CxSxO	92	828	.6243	.7951
CxPxO	207	828	1.0678	1.3599*
SxPxO	36	828	.8816	1.1228
CxSxPxO	828		.7852	
Activity Scales				
Concept	23	137	122.9345	7.1069*
Scale	4	93	48.3079	2.9009*
Person	9	69	7.3739	.7339
Occasion	1	5	.4817	.3480
CxS	92	300	13.3407	6.7500*
CxP	207	443	5.2692	2.6387*
CxO	23	72	1.5060	1.6885*
SxP	36	312	6.0781	3.0165*
SxO	4	55	.0754	.0829
PxO	9	79	1.5622	1.6790
CxSxP	828	828	1.9319	2.4691*
CxSxO	92	828	.8269	1.0569
CxPxO	207	828	.8474	1.0831
SxPxO	36	828	.8655	1.1061
CxSxPxO	828		.7824	

*p .05

Table 4
Percentage of Variance Accounted for CxS Interaction for Four Instrument Types

	Standard Instructions	With Irrelevance Option
Evaluative Scales		
Bi-polar	9.83	7.82
Uni-polar	5.51	9.06
Potency Scales		
Bi-polar	14.06	12.10
Uni-polar	11.38	12.27
Activity Scales		
Bi-polar	14.96	14.17
Uni-polar	11.38	12.27

ing design, then, for each case, was a 5 (scales) × 24 (concepts) × 2 (presence or absence of the irrelevance option) × 2 (polarity type) design, with repeated measures on scales and concepts. Because scales and concepts were random factors, some significance tests were based upon quasi-*F* ratios.

In terms of the proportion of variance accounted for, the effects of the irrelevance option and type of scale were minor. Each of the main effects and interactions involving these factors accounted for less than 1% of the total variance, with only two exceptions. The Concept × Scale × Polarity Type interaction did account for approximately 1% of the variance for Potency and Activity scales on both occasions.

Tests of significance showed a significant main effect for polarity type for Potency, Occasion 1, $F(1,23) = 7.06$ and Occasion 2, $F(1,27) = 6.64$; Activity, Occasion 1, $F(1, 5) = 7.57$; and Evaluation, Occasion 2, $F(1,33) = 4.66$. This effect was marginally significant for Activity, Occasion 2, $F(1,6) = 4.71$; and Evaluation, Occasion 1, $F(1,23) = 7.06$, and Occasion 2, $F(1,27)$ higher scores in the unipolar case for all dimensions and all conditions. The main effect for presence or absence of the irrelevance option was not significant in any analysis.

Table 5
Percentage of Variance Accounted for by CxPxS Interaction for Four Instrument Types

	Standard Instructions	With Irrelevance Option
Evaluative Scales		
Bi-polar	9.83	13.60
Uni-polar	12.77	11.49
Potency Scales		
Bi-polar	16.32	21.13
Uni-polar	24.37	17.09
Activity Scales		
Bi-polar	19.49	24.18
Uni-polar	22.29	18.41

The interactions in these analyses which are of theoretical interest are the Concept × Scale × Polarity Type and Concept × Scale × Irrelevance Option effects. These effects were all nonsignificant, with the exception of the Concept × Scale × Polarity Type interaction for the Activity dimension, Occasion 2, $F(92,3312) = 1.35$, and the Concept × Scale × Irrelevance Option interaction for Potency, Occasion 2, $F(92,3312) = 1.42$.

Discussion

Stability of E, P, and A

One clear conclusion to be drawn from the present study is that given an appropriate choice of scales and concepts, the typical EPA structure of semantic differential data is highly robust. Variations in instructions and in instrument format did not alter, in any important way, the usual three-dimensional structure. In addition, the EPA structure was highly reliable over time. Factor structures varied little from Occasion 1 and Occasion 2, and the occasions factor and its interactions were not significant in the ANOVA analyses.

It should be noted that the present study did not address the issue of the generality of the EPA structure. The scales used were selected from those found to have high loadings on E, P, or A in prior research, and concepts were selected based on their distribution in this space. Thus, the research did not answer the question whether these dimensions are in fact the basic dimensions of meaning which underlie all concept judgments.

Scale Considerations

Despite the fact that this study replicated once again an EPA factor structure, deviations from standard instructions provided some interesting insights into the process of semantic differentiation and clarified controversial issues pertaining to that structure. Foremost among the latter is the issue of the bipolarity of semantic space. Factor analyses of unipolar scales not only revealed the same EPA structure obtained with bipolar scales, but also a correction of the correlation of each scale and its presumed opposite for unreliability in each scale showed that the approximate bipolarity revealed by Table 1 was indeed bipolarity of a very substantial sort.

The recommendation that standard semantic differential instructions be revised to allow for an irrelevance response separate from the scale midpoint seems appropriate. Overall, examination of the factor structures and of the ANOVA results indicated that use of the irrelevance option tended not to change the pattern of results significantly. However, if the inclusion of the irrelevant response option has an effect on factor structures, it would be useful to detect combinations of scales and concepts that result in a high level of irrelevance responding. Such information would increase the precision of the semantic differential by providing a more careful matching of concepts with scales than has been the case. Factors which contribute to the judgment of irrelevance have already been given and seem related to those suggested previously by Smith and Nichols (1973).

In general, deviations from standard procedures had little effect on the factor structure or on the effects tested in the ANOVA. The exception to this was the main effect for polarity type, reflecting significantly higher (less favorable, less potent, and less active) unipolar ratings than bipolar ratings. However, this effect accounted for less than 1% of the variance in all cases and was difficult to interpret.

Concept and Scale Interactions

The application of ANOVA to semantic differential data departed from the usual methodology in this area. It permitted the comparison of the effects of the various factors in the methodology and within the procedure itself. Significant effects accounting for a substantial proportion of the variance were observed for concepts and for scales in the overall ANOVAs. The significant concept effect validated the attempt to select concepts uniformly distributed

throughout semantic space. The concept effect continued to be substantial when the ANOVA was performed for each of the three dimensions, thus illustrating a satisfactory distribution of concepts over each dimension.

It was the procedure of performing ANOVAs separately for the three types of scales that provided a new look at problematic features of semantic differential data. It has been argued that with this analysis, the magnitude of the Concept × Scale effect corresponds to and is a measure of what has been called Concept × Scale interaction. The data from the present study indicated quite clearly that such an interaction exists and is a serious problem for research using the semantic differential. It occurred to a greater extent with Potency and Activity scales than with Evaluative scales, but Concept × Scale interaction accounted for over 10% of the variance, even for Evaluative scales.

This finding has implications for use of the semantic differential in obtaining dimensionally "pure" ratings of concepts. Certain concepts are likely to be distorted in terms of their position in semantic space. Those positions are calculated using scales which load on one factor but which have varying relationships to each other depending on the concept judged.

This interpretation of the Concept × Scale interaction is demonstrated graphically in Figure 1. In these graphs, concepts are ordered along the horizontal axis according to their overall rating on the relevant dimension. The rating of each concept on each of the five scales making up this dimension are plotted on the vertical axis. If no Concept × Scale interaction were present, the function for each scale should be monotonically increasing and all curves should be parallel. This is clearly not the case.

A substantive interpretation of this interaction can be noted in the graph for Evaluative scales. Concept 21 was *puppies*. *Puppies* were judged to be relatively beautiful, good, valuable, and kind but were seen as very foolish, as opposed to wise. This is the essence of Concept × Scale interaction: In the context of *puppies*, "foolish" is simply not a negative evaluative term. Similarly with respect to the potency graph, *sex* (Concept 16) is rated as strong and powerful, but also as delicate, fragile, and soft.[5]

This difficulty calls into question research involving attitude measurement which has used the semantic differential to assess the differences between concepts on particular dimensions but in which the Concept × Scale interaction has not been assessed. What is needed in such studies is a demonstration that Concept × Scale interaction did not exist (or was minimal) for the particular stimuli used.

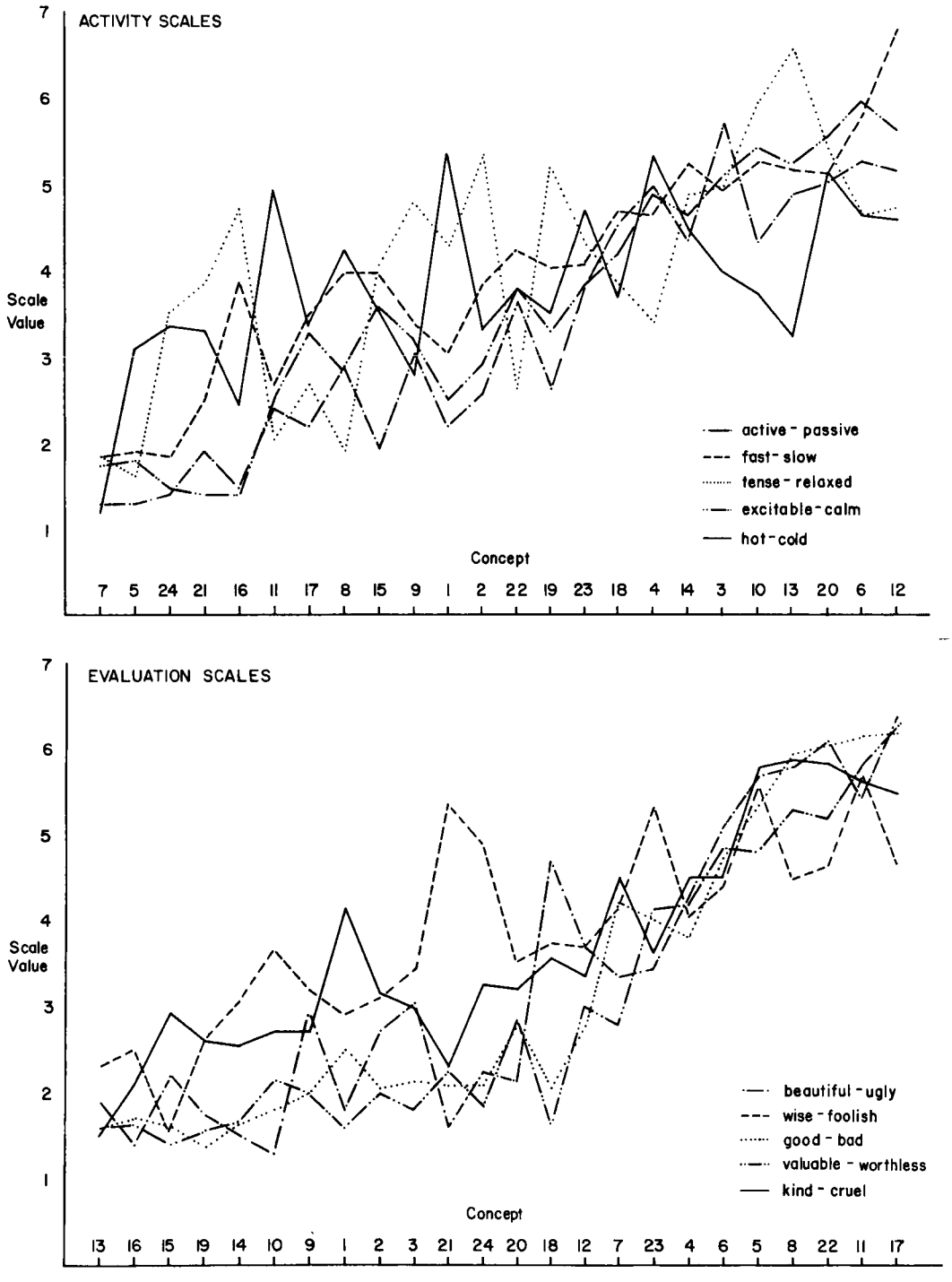
There are two further points relevant to the Concept × Scale interaction issue. This study demonstrated that Concept × Scale interaction was more than just the result of an inappropriate choice of scales and concepts. The methods used in stimulus selection, the clear and unequivocal presence of the EPA factor structure, as well as the proportion of variance accounted for by the concept effect, suggest that the various positions in semantic space were well sampled. Nevertheless, the Concept × Scale effect was substantial. The second point is just the obverse of the first. Despite the presence of a substantial Concept × Scale interaction, the traditional EPA structure emerged. Thus, it would seem that however profound the interaction effect, the factorial structure that sorts scales into the Evaluative, Potency, and Activity categories was sufficiently powerful to overcome and to obscure it.

The presence of Concept × Scale interaction will not come as a surprise to researchers in this area. The presence of this effect has been widely noted. What may be surprising is the magnitude of the Scale × Concept × Person effect. In every case, this three-way interaction accounted for a larger proportion of the variance than did the Concept × Scale interaction. Thus, overall, the magnitude of individual differences was high, and these differences were manifested

[5]As noted previously, a case can be made for partitioning the Potency dimension into the two factors strong-powerful and hard-rugged-tough. In those cases in which four factors were rotated, it was precisely this partition which resulted.

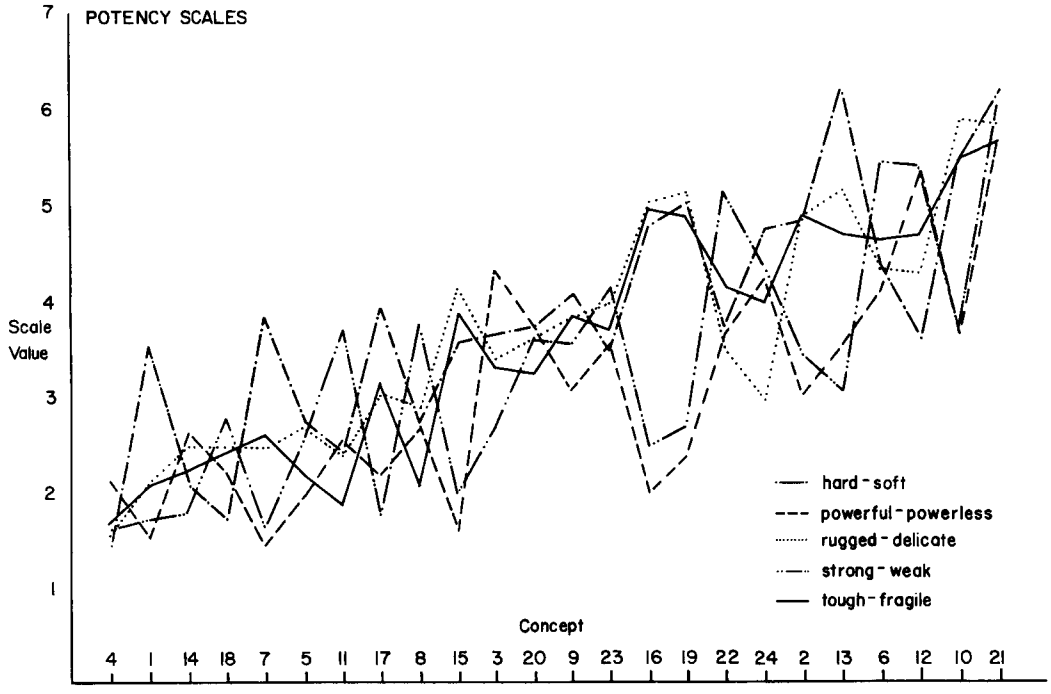
Figure 1

Semantic differential ratings of 24 concepts on each of
5 Evaluative scales, 5 Potency Scales, and 5 Activity scales



(continued on next page)

(Figure 1 continued)



most strongly in the three-way interaction. While the Scale × Person effect was significant in most cases, the proportion of variance accounted for by this interaction was low. The Concept × Person interaction would be expected to be high, representing individual differences in the affective meanings of various concepts. The effect was strongest for Evaluative scales. It is this interaction which justifies use of the semantic differential to assess between-group differences in attitudes.

The substantial Scale × Concept × Person interaction, however, indicated that individuals' use of scales within a dimension differed according to the concept being judged. In three of the four analyses for Evaluation, the three-way interaction accounted for a lower proportion of the variance than the expected Concept × Person interaction. In all other cases, however, the reverse was true.

The finding that the interaction of scale, concept, and person was larger than the Scale × Concept effect is damaging to interpretations based on three-dimensional semantic space. It suggests that alternative representations of semantic space may be required if such space is to be meaningful at the level of the individual person's cognitive structure. The present data certainly do not support extreme claims that EPA is nothing more than a statistical artifact produced by averaging over subjects. The results which have been presented based on factor analyses of individual persons' data clearly refute that. But those analyses did show that EPA is not equally descriptive of everyone. Moreover, even if each person were to display the identical

factor structure, the substantial context distortions shown by the Scale × Concept × Person effects must be taken into account. These effects appear to have little impact on the factor structure itself, but they do reflect a much greater idiosyncracy in semantic differential judgments than has previously been shown.

References

Bynner, J., & Romney, D. A method of overcoming the problem of concept-scale interaction in semantic differential research. *British Journal of Psychology*, 1972, *63*, 229–234.

Carter, R. F., Ruggels, W. L., & Chaffee, S. H. The semantic differential in opinion measurement. *Public Opinion Quarterly*, 1968, *32*, 666–674.

Clark, V. A., & Kerrick, J. S. A method of obtaining summary scores from semantic differential data. *Journal of Psychology*, 1967, *73*, 211–218.

Crockett, W. H., & Nidorf, L. J. Individual differences in responses to the semantic differential. *Journal of Social Psychology*, 1967, *73*, 211–218.

Davis, K. L. *S*'s "appropriateness" ratings of semantic differential concept-scale combinations: Generality with respect to synonymous concepts, relationship to rating polarity and ability to differentiate *S*'s labelling of the mid-scale rating position (Doctoral dissertation, University of South Carolina, 1972). *Dissertation Abstracts International*, 1973, *34*, 388B–389B. (University Microfilms No. 73–16, 300)

Forthman, J. H. The effects of a zero interval on semantic differential rotated factor loadings. *Journal of Psychology*, 1973, *84*, 23–32.

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

Heaps, R. A. Use of the semantic differential technique in research: Some precautions. *Journal of Psychology*, 1972, *80*, 121–125.

Heise, D. R. Some methodological issues in semantic differential research. *Psychological Bulletin*, 1969, *72*, 406–422.

Heskin, K. J., Bolton, N., & Smith, F. V. Measuring the attitudes of prisoners by the semantic differential. *British Journal of Social and Clinical Psychology*, 1973, *12*, 73–77.

Jenkins, J. J., Russell, W. A., & Suci, G. J. An atlas of semantic profiles for 360 words. *American Journal of Psychology*, 1958, *71*, 688–699.

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

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.

Krieger, M. H. A control for social desirability in a semantic differential. *British Journal of Social and Clinical Psychology*, 1963, *2*, 94–103.

Kubiniec, C. M., & Farr, S. D. Concept-scale and concept-component in interaction in the semantic differential. *Psychological Reports*, 1971, *28*, 531–541.

Lloyd, B. B., & Innes, J. M. Influence of past experience on meaningfulness of concepts on a semantic differential. *Psychological Reports*, 1969, *24*, 269–270.

Maguire, T. O. Semantic differential methodology for the structuring of attitudes. *American Educational Research Journal*, 1973, *10*, 295–306.

Mayerberg, C. K., & Bean, A. G. Two types of factors in the analysis of semantic differential attitude data. *Applied Psychological Measurement*, 1978, *2*, 469–480.

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 of semantic differential scales. *Psychological Reports*, 1965, *16*, 691–692.

Nichols, H., & Smith, R. G. Perception of intensional and extensional meaning domains in a semantic differential application. *Speech Monographs*, 1973, *40*, 322–325.

Oetting, E. R. The effect of forcing response on the semantic differential. *Educational and Psychological Measurement*, 1967, *27*, 699–702.

Osgood, C. E., May, W. H., & Miron, M. S. *Cross cultural universals of affective meaning*. Urbana, IL: University of Illinois Press, 1975.

Osgood, C. E., Suci, G. J., & Tannebaum, P. H. *The measurement of meaning*. Chicago, IL: University of Illinois Press, 1957.

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. *Journal of Abnormal and Social Psychology*, 1961, *62*, 62–73.

Presley, A. A. Concept-scale interaction in the semantic differential and its implications for factor scores. *British Journal of Psychology*, 1969, *60*, 109–113.

Price, L. A. *Automated administration of psychological instruments: Pedagogical and practical considerations*. Paper presented at the Conference on Computers in the Undergraduate Curricula, Michigan State University, 1977.

Rosenbaum, L. L., Rosenbaum, W. B., & McGinnies, E. Semantic differential factor structure stability across subject, concept, and time differences. *Multivariate Behavioral Research*, 1971, *6*, 451–469.

Ross, B. M., & Levy, N. A comparison of adjectival antonyms by simple card pattern formation. *Journal of Psychology*, 1960, *49*, 133–137.

Shikiar, R., Fishbein, M., Wiggins, N. Individual differences in semantic space: A replication and extension. *Multivariate Behavioral Research*, 1974, *9*, 201–209.

Smith, R. G., & Nichols, H. J. Semantic differential stability as a function of meaning domain. *Journal of Communication*, 1973, *23*, 64–73.

Snyder, F. W., & Wiggins, N. Affective meaning systems: A multivariate approach. *Multivariate Behavioral Research*, 1970, *5*, 453–468.

Terwilliger, R. F. Free association patterns as a factor relating to semantic differential responses. *Journal of Abnormal and Social Psychology*, 1962, *65*, 87–94.

Thompson, E. G. *The role of computer-supported instruction in a course on the psychology of oneself.* Paper presented at the Conference on Computers in the Undergraduate Curricula, Michigan State University, 1977.

Vidali, J. J. Single anchor Stapel scales versus double anchor semantic differential scales. *Psychological Reports*, 1973, *33*, 373–374.

Vidali, J. J., & Holeway, R. E. Stapel scales versus semantic differential scales: Further evidence. *Psychological Reports*, 1975, *36*, 165–166.

Warr, P. B., Schroder, H. M., & Blackman, S. A comparison of two techniques for the measurement of international judgment. *International Journal of Psychology*, 1969, *4*, 135–140.

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, IL: Aldine, 1969.

Acknowledgments

The research reported in this paper was conducted under the auspices of the Computer Institute for Social Science Research, Michigan State University. The senior author was supported by an NIMH Traineeship under Grant No. MH10779 to the Department of Psychology, Michigan State University. Funds to support the final revisions of the manuscript were provided by the College of Education and the Department of Psychology at Oklahoma State University. We gratefully acknowledge the assistance of Mr. Richard Clements with the data analysis.

Author's Address

James L. Phillips, Department of Psychology, Oklahoma State University, Stillwater, OK 74074.