DOCUMENT RESUME

ED 137 387 TM 006 193

AUTHOR Davis, Richard W.

TITLE Development of an Appropriate Method for Applying the Semantic Differential to the Measurement of Affective Outcomes of Instruction.

PUB DATE [Apr 77]

NOTE 20p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New York, New York, April 1977)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.

DESCRIPTORS *Affective Behavior; *Affective Tests; Course Evaluation; *Factor Analysis; Factor Structure; Higher Education; Mathematical Models; Reliability; *Semantic Differential; Statistical Analysis; Tests of Significance; Validity

IDENTIFIERS Canonical Analysis

ABSTRACT

A method for statistical analysis of semantic differential data in educational evaluation is discussed. Estimated scores for unobserved affective variables are obtained using the canonical factor regression method. This method overcomes previous prolems of bias and inefficiency in computing composite affective indices. In an application of the technique, a two-week program of instruction was shown to produce significant changes in affective response to two of the eight concepts tested. (Author)

Development of an Appropriate Method for Applying the Semantic Differential to the Measurement of Affective Outcomes of Instruction

Richard W. Davis

University of Minnesota

Paper presented at the Annual Meeting of the National Council on Measurement in Education: New York, April, 1977

Introduction

The current wide-spread use of the semantic differential (SD) for indexing outcomes of instruction is not surprising. The SD is easy to administer, has a large body of literature to support its use, and is relatively easy to analyze. Still, few researchers or evaluators are aware of the breadth of use of the SD. Hecht (1971) reviewed twenty-some studies using the SD in the evaluation of instructional outcomes in science education alone. Bi-polar adjective scales have been used to evaluate conferences (Frisen & Bumbarger, 1972), college freshman classes (Hoover, Baumann, & Shafer, 1970), instructional programs (Carver & Phipps, 1969), and even individual lessons (Evans, 1970).

But many studies using SD instruments to evaluate outcomes of instruction appear to have basic methodological or theoretical difficulties. In general the three most important difficulties are these:

A. The studies are undertaken without consideration of the assumptions or limitations of traditional SD theory.

B. The studies fail to precisely consider what it is that the SD indexes.

C. Studies fail to employ appropriate analytical methods.

The purpose of this paper is to set out an alternative approach to evaluation using the SD, based on a restricted model for SD application which meets the practical and methodological requirements of educational evaluators. While this restricted model cannot be applied in all situations, it does appear to have important advantages in the measurement of affective outcomes of instruction in those situations where one is particularly interested in indexing a general evaluative response to the instruction, where the learner population is of moderate size, and where one is evaluating responses to easily defined concepts.

The Restricted Model

The major requirement of any SD application is that the researcher respect the structural or geometric properties of the SD. In traditional SD analysis this means dealing with the conventional semantic space solution of the analysis of a set of bi-polar adjective scales. Osgood, Suci, & Tannenbaum (1957) set forward a model space made up of three dimensions: Evaluative, Potency, and Activity (EPA). They arrived at this structure by the factor analysis of many scales, and argue that it is stable across concepts, people, and even across cultures. The restricted model of affective evaluation pursued here focuses on the single dimension of the evaluative factor. While there are several reasons for making this restriction, the central argument is simply that the affective response information contained in SD responses appears in that dimension.

In one of his early studies, Osgood (1952) found that many of the bi-polar scales applied to a concept loaded heavily on a single factor. That factor, he found, accounted for a large proportion of the total variance in the scales, often as much as 70% or more. In addition, several individual scales loaded very highly on the first factor. The restricted model here represents a throw-back to that earlier model of the evaluative factor for the purpose of using SD response data in the evaluation of affective outcomes.

The advantage of the single factor model of the SD in evaluation being set out here is that it helps the investigator solve three basic problems encountered in SD evaluation studies. These are:

1. The problem of what it is that the SD indexes.
2. The problem of concept-scale interactions.
3. The problem of appropriate methodology.

Each of these will be considered in turn.

What does the SD index? An investigator following traditional SD theory who has read The Measurement of Meaning several times will invariably respond to the question of what the SD indexes with the knee-jerk reflex: The SD measures connotative meaning. But does it? Or, even more important, is connotative meaning what the evaluator wishes to measure when looking at affective responses to instruction? Probably not. The question has been debated almost constantly for twenty years by critics from Carroll (1959) to Miron (1969) who asks: "What is it that is being differentiated by the semantic differential?" Perhaps a better question for evaluators is: What is it that we want the semantic differential to differentiate? One answer which is of importance for many evaluators is: Affective responses to concepts.

If one examines the sorts of scales which are commonly found to load heavily on the first unrotated factor, there are many which are easily identified with affective evaluation. Among these are fair--unfair, kind--cruel, valuable--worthless, and honest--dishonest. Considered alone, the first factor (or E-factor) may be characterized as representing general affect--a general disposition toward the concept being indexed by the SD. It is sort of an evaluative G-factor corresponding in evaluation to the general intelligence factor found in intelligence testing. Considered by itself, it is of great practical interest to the educational investigator. It provides a single index of the general affective response of a set of people toward a set of concepts. In certain applications, at least, the SD's first factor indexes affect.

Concept-Scale Interactions. A difficulty for many users of SD methodology is the problem of concept-scale interactions. Essentially the problem is that in many studies researchers have failed to find the same pattern of scale loadings on factors which Osgood found. The importance of this finding is that it suggests that, contradictory to Osgood's position, that EPA structure is not stable. In fact, it is quite common to find scales traditionally thought of as definitive for the Potency dimensions (active--passive) loading at high levels on the first factor or E-factor. Levy (1969) and Heise (1969) summarize the concept-scale interaction literature well. Osgood and Suci (1955) suggest that, in fact, there is not a necessary correspondence between the factors obtained in an analysis of SD scales and the underlying dimensions of semantic space.

But the concept-scale interaction problem has practical consequences for the investigator wishing to use bi-polar adjective scales to index evaluative responses to concepts. Heise (1969) indicates that one cannot simply assume that the scale-concept relationships found in the analysis of one set of concepts will hold for another set of concepts as well. Therefore, the pattern of loadings must be re-determined through factor analysis or multi-dimensional scaling for each new study. This raises the complexity of SD studies greatly--frequently putting them out of the reach of many investigators.

The use of the restricted model of the SD advanced here helps overcome concept-scale interaction problems in two principal ways. First, since one is concerned with only a single factor or single underlying dimension, the question of interaction is limited to a question of which scales load on the first factor rather than which scales load on which of several factors. Thus the problem is simplified. Second, the analysis which needs to be performed to answer this question is also simplified. One does not have to use a full factor analysis (although factor analytic techniques will be used) to find out which scales contribute in what way to the underlying evaluative dimension.

While one cannot argue that the evaluative factor is stable with respect to the pattern of scale loadings, one can at least argue that the problem is much less critical in the restricted model than it is in the traditional model.

<u>Appropriate methodology</u>. If the methods the investigator employs in examining relationships in the SD data are to be appropriate, they must be consistent with the notions set out in the previous discussion. Specifically, they must:

1. Consider adequately the problem of structure.

2. Yield an appropriate index of the evaluative dimension of the structure.

Ideally the methodology should also be kept relatively simple.

The problem of structure is not complex, but a remarkable number of studies fail to deal with the problem adequately. The key issue is that since one is arguing that the SD is a multiple-scale index of some underlying dimension or factor, one must test relationships on the basis of that underlying dimension, and not on the scales themselves. This point deserves expansion.

Consider the question of testing the hypothesis that an instructional treatment brought about a change in affective response to the concept BANANA in a social studies unit on Central America. One hopes that the SD can be used to determine if a change took place. But if one simply performs a set of <u>t</u>-tests on the individual scales used in the analysis, there are major methodological problems. First, one does not know which scales index the general evaluative response. One guesses that good-bad does, but what about the others? Then suppose that six out of ten tests were significant. Does that indicate six independent significant results, or is the investigator simply measuring the same underlying relationship six times? A further problem is that since the tests are not independent (scales are clearly correlated) the likelihood of a type I statistical error becomes very large, even with a small number of scales. In short, the methodological approach of performing repeated tests of significance on individual SD scales is bankrupt and should not be employed under any circumstances.

Given, then, that the investigator must find some way of calculating a combined index of the underlying relationships in the data, how should he or she approach the problem? The concept--scale interaction problem strongly suggests that one cannot assume relationships between scales and concepts in a study based on the analysis of concepts in another study. So some analytic technique, such as MDS, factor analysis, or cluster analysis must be employed on the data obtained in the evaluation study itself. The two approaches which will be outlined here are both based on factor analysis, although neither is, strictly speaking, a factor analytic method.

The CFRM method. Allen (1974) suggests that the canonical factor, regression method (CFRM) is a sound general method of computing composite indexes of unobserved variables. Allen provides an extensive development of the argument for using this particular approach, but the features of particular interest in this application are that the technique provides maximally reliable and valid estimates of the unobserved variable, and that the obtained factor is uncorrelated with the residual variance in the data. In addition the approach has practical merit since the routines used to calculate the index are available through standard statistical programs such as SPSS.

The procedure followed in calculating the scores for the unobserved evaluative index using the CFRM are outlined in Figure 1. Essentially one performs a factor analysis of the pooled SD data across the scales using Rao's Canonical Method of factor analysis (Rao, 1955). Using the factor score coefficients obtained by the regression method (Harris, 1967), one calculates factor scores for the pre- and post-instruction responses obtained for each concept in the analysis. Generally several concepts are used in a given analysis. An appropriate test of significant differences between means (usually a t-test) is then used to determine if the values for the general evaluative response index differs across the instructional treatment.

Insert Figure 1 about here

The advantage of the CFRM method is that the index employed has several desirable properties. It is uncorrelated with the residual variance in the analysis. It can be used to determine what proportion of the total variance the index represents, and it includes the contributions to the score of each scale in the analysis. Its major disadvantage is that it is computationally complicated and is perhaps more detailed than most evaluators need.

Unit weight method. For those situations in which the precision and statistical niceties of the CFRM method are not required, a simpler method is available. After obtaining the loadings of the scales on the first unrotated factor (and Roa's Canonical Method is again recommended), one identifies those scales which load highest on the factor. In a set of 16 scales it is common to find over half with absolute loadings of .500 or higher. Summing over those scales with loadings above some pre-set level (such as .600) will produce a composite index. Even though this index does not have ideal statistical properties, one has a strong argument for suggesting that the index does measure general evaluative response. In fact, the CFRM index and the unit weight index should intercorrelate highly.

An Empirical Demonstration

The approach set forward above was employed in the evaluation of affective outcomes of a college course in Instructional Systems. As part of a formal program of course design and development evaluation questions regarding the affective outcomes of a unit on the management of instruction were raised. Specifically the course evaluators asked if the unit brought about changes in the affective response by learners to key concepts of the instruction.

Concepts. Four concepts were identified for analysis. These were LEARNING RESOURCES, A-V MATERIALS, MANAGEMENT OF INSTRUCTION, and AMERICAN EDUCATION. The last concept was chosen as a result of a student comment in an earlier course evaluation. The others were selected on the basis of the course content outline.

In addition to the four experimental concepts employed in the study, four "placebo" concepts were also investigated. The purpose of using these concepts was to show that the evaluative response to concepts not central to the instruction would not change. The four concepts were chosen for their high familiarity. They were: KITCHEN, ARMY, MINNESOTA, and MYSELF AS A STUDENT. In a normal evaluation study using the SD, such dummy concepts would not be necessary. They were included here for the purposes of the empirical demonstration.

SD instrument. The SD instrument employed in the analysis consisted of sixteen scales of bi-polar adjectives. Each learner responded to each set of

Insert Figure 2 about here

scales for each concept investigated. One set of synonymous scales were included to provide an internal check for erratic response patterns. Scales were typically chosen because of their meaningfulness to the evaluation question (Carter, Ruggels, & Chaffee, 1968). The concept for each set of scales was printed by hand at the top of the form.

The same form was used pre- and post-instruction. Simple instructions were given prior to each administration. The inter-test interval was approximately two weeks. Responses were hand punched onto computer cards, although the use of op-scan or sense-mark cards would facilitate the data entry process in typical educational applications. Finally, all data was analyzed using the SPSS package. A simple, original FORTRAN program for calculating a reliability coefficient was used.

The Analysis. The canonical factor analysis of the pooled scale data yielded a large first factor. Unrotated it loaded highly on several scales

Insert Figure 3 about here

including bad--good and useful--useless at r = -.835 and r = .911 respectively. These two extremely high loadings clearly mark the dimension which the first factor indexes as an evaluative and/or utility dimension. Four of the scales, reflecting difficulty and speed, had low loadings and low communality estimates associated with them. In effect, then, twelve of the sixteen scales loaded heavily on the factor. Somewhat surprisingly, scales such as active--passive, which are normally thought of as not evaluative, loaded very heavily on the first factor. This finding lends support within the study for the concept/scale interaction criticism of the SD.

Factor score coefficients were calculated using the least-squares regression method (Lawley & Maxwell, 1971). Using these weights, a combined index was calculated for the unobserved evaluative variable. One variable was calculated for each of the concepts.

Reliability and validity of the index. Prior to using the composite index for testing the hypothesis that shifts in evaluative response had occurred, the reliability and validity of the composite index was determined for the pooled response data. The approach taken was to determine the Omega reliability coefficient due to Heise & Bohrnsted (1970). A discussion of the formula and the simple program used for calculating the coefficient are given in Figure 4.

Insert Figure 4 about here

Very briefly, the Omega coefficient is an index of the proportion of variance in the composite measure which is common variance. As such it is an estimate of the correlation between the composite index and the true value of the unobserved variable.

For the pooled response data, the reliability of the composite index is calculated to be $\Omega = .949$. This suggests that more than 94% of the variance observed in the composite index was true score variance of the unobserved variable. The validity coefficient corresponding to the Omega reliability coefficient is Rho. It is simply the square-root of the correlation coefficient and represents the estimated correlation between the composite index and the unobserved variable. That is calculated to be $\rho = .974$.

Tests for treatment effects. To test the hypothesis that the instructional treatment produced changes in the student's evaluative responses to the concepts, a paired t-test was performed on the pre- and post-instruction composite indexes

Insert Figure 5 about here

for each concept. Among the eight tests performed, only two were significant. The responses to the concepts A-V Material and American Education showed significant changes in mean value across the instructional treatment. Within the context of the evaluation study in which the tests were conducted, this finding was interpreted to mean that attitude shifts did occur, but that the shifts were not general to all concepts of the instruction.

A broader interpretation given to the observations at the time was that students had formed negative values regarding the way in which American Educators used audio-visual materials in general.

The simpler approach to analysis of the response data, involving pooling of response data from scales which appear to load heavily on the first factor, was not pursued in this study. In general, if the CFRM method is performed it is the approach of first choice and obviates the necessity of using the less accurate but simpler method. It is not likely that any significant shifts in evaluative response would have been observed using that method.

Notes for further development. The restricted approach to using the SD in educational evaluation, then, is presented as a more sound method than is characteristically employed in representative studies in the literature. In addition to the overall simplicity of the approach one achieves the important advantages of knowing what is being measured--evaluative response--and of relating to the essentially geometrical properties of the SD. In the restricted model one avoids the pitfalls on either extreme. Performing tests on individual scales is simply inappropriate; a full-blown semantic space model is too complex and imposes too many restrictions on the analysis to be useful to most evaluators.

Certainly there are other models. Maguire (1973), Evans (1970), and Stiggins (1972) suggest what some of them might be. But for the immediate purposes of determining changes in values associated with concepts due to instructional treatment, the restricted model set out here appears to meet the needs of a broad range of evaluation applications.

Figure 1.

Visualization of the Canonical Factor-Regression Method Analysis Procedure Employed in Testing Hypothesis One.

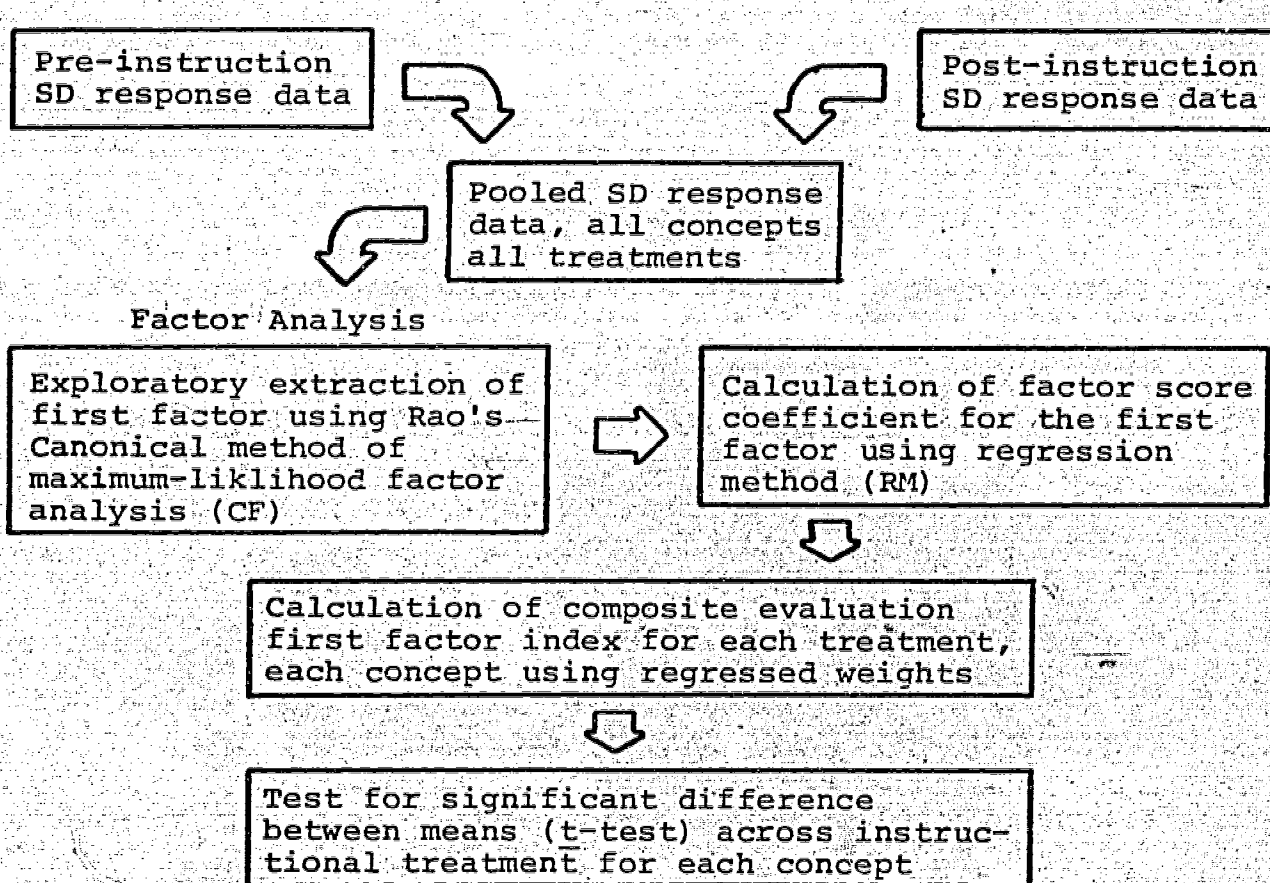

Figure 2.

Semantic Differential Instrument Used in Evaluation of Affective Outcomes

CONCEPT: ____________________

worthless :___:___:___:___:___:___:___: valuable

knowable :___:___:___:___:___:___:___: unknowable

easy :___:___:___:___:___:___:___: difficult

bad :___:___:___:___:___:___:___: good

fast :___:___:___:___:___:___:___: slow

small :___:___:___:___:___:___:___: large

exciting :___:___:___:___:___:___:___: boring

orderly :___:___:___:___:___:___:___: clutered

active :___:___:___:___:___:___:___; passive

interesting :___:___:___:___:___:___:___: dull

confusing :___:___:___:___:___:___:___: understandable

simple :___:___:___:___:___:___:___: difficult

strong :___:___:___:___:___:___:___: weak

useful :___:___:___:___:___:___:___: useless

effortless :___:___:___:___:___:___:___: demanding

important :___:___:___:___:___:___:___: unimportant

Figure 3.

Generation of a composite index of an unobserved evaluative factor based on maximum-likelihood factor analysis of sixteen scales of the semantic differential.

Scale	Loading	h^2	U^2	Weight
worthless--val.	-.825	.656	.334	-.122
knowable--unkn.	.637	.443	.557	.058
easy--difficult	.127	.019	.981	.007
bad--good	-.835	.656	.344	-.123
fast--slow	.497	.272	.728	.035
small--large	-.728	.547	.453	-.082
exciting--boring	.784	.613	.387	.103
orderly--clutt.	.660	.469	.538	.063
active--passive	.769	.592	.408	.096
interesting--d.	.769	.596	.404	.096
confusing--und.	-.705	.524	.476	-.075
simple--difficult	.085	.009	.991	.004
strong--weak	.692	.501	.499	.070
useful--useless	.912	.735	.265	.175
effortless--deman.	-.364	.134	.866	-.021
important--unim.	.816	.636	.365	.114

Figure 4.

Calculation of Factor Weights, Reliability, and Validity

Weights:

$$\omega_i = \frac{f_i / U_i^2}{1 + \Sigma(f_i^2 / U_i^2)}$$

where ω_i is the linear weight for scale i,

f_i is the factor loading between scale i and the factor ,

U_i^2 is the uniqueness associated with scale i ($1-h^2$).

Reliability:

$$\Omega = \Sigma(f_i \omega_i)$$

where Ω is the Omega reliability coeeficient.

Validity:

$$\rho = \sqrt{\Omega}$$

where ρ is the Rho validity coefficient.

FORTRAN Program for Computing Above Values

```
      PROGRAM RELY (INPUT, OUTPUT, TAPE8)
      DIMENSION F(50), H(50), U2(50), WE(50)
      READ (8,101) NUMSAL
101   FORMAT (I2)
      SIGMA = 0
      DO 10 I = 1,NUMSAL
      READ (8,103) F(I),H(I)
103   FORMAT (2F6.0)
      U2(I) = 1 - H(I)
10    SIGMA = SIGMA + F(I)**2/U2(I)
      DENOM = 1 + SIGMA
      DO 20 I = 1,NUMSAL
20    WE(I) = (F(I) / U2(I))/ DENOM
      RELY = SIGMA/DENOM
      VALID = SQRT(RELY)
      PRINT 105
105   FORMAT 1H1
      DO 30 I = 1,NUMSAL
      PRINT 107, F(I), H(I), U2(I), WE(I)
107   FORMAT (2X,4F10.3)
30    CONTINUE
      PRINT 109, RELY, VALID
109   FORMAT (//2X,14HRELIABILITY      ,
     +F6.3/2X,14HVALIDITY          ,
     +F6.3)
      CALL EXIT
      END
-EOR-
*card with number of scales in col 1-2*
*data cards, one per scale, with
 scale weight punched in col 1-6 and
 the uniqueness punched in 7-12*
```

Figure 5.

Comparison by *t*-test for paired observations of composite factor scores for an evaluative index between pre- and post-instructional administrations of 16 semantic differential scales on eight concepts. (N = 37 and df = 36 for all cases)

Concepts	Mean	SD_m	SE_m	Delta(S)	SD	SE	*t*	p(2-tail)
1. KITCHEN	-.307	.132	.022	.208	.094	.016	1.17	.251
	(-.325)	(.132)	(.022)					
2. LEARNING	-.360	.111	.018	.008	.067	.011	.70	.490
RESOURCE	(-.368)	(.116)	(.019)					
3. ARMY	-.152	.200	.033	.006	.083	.014	.44	.660
	(-.158)	(.215)	(.035)					
4. A-V	-.359	.103	.017	.020	.051	.008	2.32	.026 *
MATERIAL	(-.379)	.111	.018					
5. MINNE-	-.327	.163	.027	.004	.066	.011	.33	.744
SOTA	(-.331)	(.151)	(.025)					
6. MANAGE-	-.310	.121	.020	.008	.064	.011	.78	.440
MENT								
7. MYSELF	-.282	.184	.030	-.007	.096	.016	-.43	.671
AS A...	(-.276)	(.178)	(.029)					
8. AMERICAN	-.246	.190	.031	1.390	.861	.141	9.82	.001 ***
EDUCAT.	(-1.636)	(1.02)	(.168)					

NOTE: Numbers in parentheses refer to post-instruction data, numbers above refer to pre-instruction data.

*Result significant with p .05.

***Results significant with p .001.

Bibliography

Allen, M. P. Construction of composite measures by the canonical-Factor-regression method. In H. Costner (Ed.), Sociological Methodology 1973-1974. San Francisco, Jossey-Bass, 1974.

Carroll, J. Reviews: The measurement of meaning. Language, 1959, 35, 58-77.

Carter, R. Ruggels, W., & Chaffee, S. The semantic differential in opinion measurement. Public Opinion Quarterly, 1968, 38, 666-679.

Carver, F. & Phipps, L. The semantic differentia and instructional evaluation. Illinois School Research, 1969, 5, 42-45.

Evans, G. Use of the semantic differential technique to study attitudes during classroom lessons. Interchange, 1970, 1, 96-106.

Freisen, D. & Bumbarger, C. An experiment in attitude change. Alberta Journal of Educational Research, 1970, 16, 211-216.

Harris, C. W. On factors and factor scores. Psychometrica, 1967, 32, 363-379.

Hecht, A. A semantic differential evaluation of attitudinal outcomes of introductory physical science (Doctoral dissertation, University of Illinois, 1970). Dissertation Abstracts International, 1971, 31, 4576A. (University Microfilms No. 71-5,123).

Heise, David R. Some methodological issues in semantic differential research. Psychological Bulletin, 1969, 72, No. 6, 406-422.

Heise, D. R. & Bohrnstedt, H. W. Validity, Invalidity, and reliability. In E. F. Borgatta and G. W. B-hrnstedt (Eds.). Sociological Methodology, 1970. San Francisco: Jossey-Bass, 1970.

Hoover, K. J., Baumann, V. H. & Shafer, S. M. The influence of class-size variation on cognitive and affective learning of college freshmen. The Journal of Experimental Education, 38(3), 1970, 40-43.

Lawley, D. & Maxwell, A. Factor analysis as a statistical method. New York: American Elsevier, 1971.

Levy, Phillip. Concept-Scale Interaction in Semantic Differential Research: Solutions In Search Of A Problem. Br. J. Psychol., 1972, 63, 235-236.

Maguire, R. O. Semantic differential methodology for the structuring of attitudes. American Education Research Journal, 1973, 10, 295-306.

Miron, M. S. What is it that is being differentiated by the semantic differential? Journal of Personality and Social Psychology, 1969, 12, 189-193.

Osgood, C. E. The nature and measurement of meaning. Psychological Bulletin, 1952, 49, 197-237.

Osgood, C. E. & Suci, G. J. Factor analysis of meaning. Journal of Experimental Psychology, 1955, 50, 325-338.

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

Rao, C. R. Estimation and tests of significance in factor analysis. Psychometrica, 1955, 20, 93-111.

Stiggins, R. Stability of semantic meaning space and change in concept meaning during teacher training (Doctoral dissertation, Michigan State University, 1972). Dissertation Abstracts International, 1973, 33, 6188A. (University Microfilms No. 73-12,829.)