DOCUMENT RESUME

ED 041 262 24 AL 002 463

AUTHOR	Kasschau, Richard A.
TITLE	Quantitative Measures of Word Meaning and Effects of Variations in Meanings on Ease of Learning.
INSTITUTION	South Carolina Univ., Columbia. Dept. of Psychology.
SPONS AGENCY	Institute of International Studies (DHEW/OE), Washington, D.C.
BUREAU NO	BR-9-D-002
PUB DATE	Jun 70
GRANT	OEG-4-9-500002-019-057
NOTE	48p.
EDRS PRICE	EDRS Price MF-$0.25 HC-$2.50
DESCRIPTORS	*Adjectives, *Association (Psychological), Cognitive Processes, Experimental Psychology, Measurement Instruments, *Paired Associate Learning, Patterned Responses, *Psychological Tests, Semantics, Testing, Verbal Learning
IDENTIFIERS	*Semantic Differential

ABSTRACT

The research reported here was directed at two distinct but related problems: (1) the assumption of bipolarity underlying standard semantic differential scales, and (2) the demonstration of the similarities between D-4 (the square root of the sum of squares of the difference between each word's mean rating and 4.00 on a number of scales) as a measure of average intensity of meaningfulness and the average number of associations elicited by a word in a predetermined length of time. Three experiments (or groups of experiments) were carried out and are reported here: (1) "Unipolar vs bipolar semantic differential rating scales," (2) "Semantic satiation as a function of initial meaning intensity and unipolar vs bipolar rating scales," and (3) "The effects of average degree of polarization on paired-associate and serial learning and the von Restorff effect." Each of the three sections has been written with the support of, but independent of, the other two sets of data in order to minimize duplication. (JD)

BR 9-D-002
PA 24

FINAL REPORT
Project No. 9-D-002
Grant No. OEG-4-9-500002-0019-057

QUANTITATIVE MEASURES OF WORD MEANING AND EFFECTS OF VARIATIONS IN MEANINGS ON EASE OF LEARNING

Richard A. Kasschau
Department of Psychology
University of South Carolina

June, 1970

U. S. DEPARTMENT OF
HEALTH, EDUCATION, AND WELFARE

Office of Education
Bureau of Research

FINAL REPORT
Project No. 9-D-002
Grant No. OEG-4-9-500002-0019-057

QUANTITATIVE MEASURES OF WORD MEANING AND EFFECTS OF VARIATIONS IN MEANINGS ON EASE OF LEARNING

Richard A. Kasschau

Department of Psychology
University of South Carolina

Columbia, South Carolina 29208

June 30, 1970

The research reported herein was performed persuant to a grant with the Office of Education, U. S. Department of Health, Education, and Welfare. Contractors undertaking such projects under Government sponsorship are encouraged to express freely their professional judgment in the conduct of the project. Points of view or opinions stated do not, therefore, necessarily represent official Office of Education position or policy.

U. S. DEPARTMENT OF
HEALTH, EDUCATION, AND WELFARE

Office of Education
Bureau of Research

Contents

Page

Tables:

Experiment I

Experiment II

Experiments III-V (I-III)*

Figures:

Experiments III-V (I-III)*

*Note: Experiments III, IV, and V of the original research proposal have been renumbered Experiments I, II, and III respectively in the third paper of this final report.

Preface

The research herein reported has been directed at two distinct, but related, problems. The first problem concerned the assumption of bipolarity underlying standard semantic differential (SD) scales. The first experiment (Pp. 1-15) specifically investigated possible use of unipolar SD scales in preference to the typical bipolar format, as well as what consistencies existed among the ratings of a set of words on each bipolar and the complementary unipolar scales.

The second experiment (Pp. 16-26) was also related to the first problem, but in this effort semantic satiation, or more generally, change in rated meaningfulness, was used to specify the effect(s) of the presence of a second anchoring adjective (as exists on bipolar scales) as compared with such changes on scales having only one adjectival anchor.

The second problem had to do with demonstrating the similarities between D_4 (the square root of the sum of squares of the difference between each word's mean rating and 4.00 on a number of scales) as a measure of average intensity of meaningfulness and the average number of associations (m) elicited by a word in a predetermined length of time. Previous studies had demonstrated that variations in m produced wide differences in ease of learning. The last group of experiments (Pp. 27-41) attempted to demonstrate effects similar to those obtained using m, but instead holding m constant and varying D_4 in a variety of learning situations.

The Publication Manual (1967) of the American Psychological Association was used for purposes of organizing the separate reports. Each of the three reports which compose this Final Report have been written with the support of, but independent of, the other two sets of data. Of necessity there may be slight overlap in the content of these reports, but every effort has been made to minimize this duplication, consistent with the obvious goal of clarity of communication.

UNIPOLAR VS BIPOLAR SEMANTIC DIFFERENTIAL RATING SCALES

Richard A. Kasschau

University of South Carolina

Abstract

One hundred fifty Ss (N = 30/group) rated each of 96 nouns and adjectives on 10 bipolar, 7-position (B7) semantic differential (SD) scales or on one of two sets of unipolar, 4-position or unipolar, 7-position (U7) SD scales generated from the B7 scales. Summary tables containing the mean and standard deviation of each rating were obtained, as well as a table of the distance of each mean rating from neutrality. Numerous tests based on the data in these tables were generally consistent in demonstrating that mean ratings were most spread out on the B7 scale and had the smallest standard deviation. Additionally, the average distance from neutrality of the B7 rating of each word was correlated +.78 with the disagreement manifested by Ss rating the same word on the complementary U7 scales generated from the reference B7 scale. It was concluded that the B7 scale was the best measure of semantic meaningfulness and that w. d-scale interactions may have interferred with previous attempted demonstrations of the bipolarity of semantic space.

Of the many articles concerning the methodological problems associated with the semantic differencial (SD), very few have been addressed to the assumption concerning the bipolarity of the rating scales. According to this assumption, the anchoring adjectives at each end of the SD adjective rating scales are of equal and opposite semantic intensity. Demonstrating the varacity of this assumption has long been recognized to be a difficult problem (cf., Osgood, Suci, & Tannenbaum, 1957; and more recently, Heise, 1969).

Ross and Levy (1960) asked Ss to use playing cards to produce patterns of three rows and three columns which might represent various adjectives used as stimuli. Whereas the most beautiful, the simplest, and the commonest patterns were fairly well agreed upon across Ss, the opposite adjectives resulted in a large number of distinct patterns. The authors concluded it was probably not correct to consider the antonyms to be of equal and opposite semantic force in the description of simple patterns. Although the evidence was contrary to the assumption of bipolarity, it should be pointed out that the reproduction of patterns from playing cards is not analogous to what occurs when an individual must rate a word on an adjective scale.

More recently, Green and Goldfried (1965) have reported the only relevant study directly concerned with the assumption of bipolarity. They asked Ss to rate the degree and type of relationship existing between words and single adjectives on unipolar, seven-position (U7) scales of the following form:

good

______:______:______:______:______:______:______

It was expected that to the extent a word was rated to the left (negative relation) side of the good scale, the same word should be rated to the right (positive relation) of the bad scale. However, several correlational and factor analyses failed to confirm numerous predictions generated by Green et al. from the assumption of bipolarity.

A very important aspect of a word's location on any standard bipolar, seven-position (B7) scale was ignored and might influence the rating of a word on U7 scales. The mean rating of a word on a B7 scale would seem to be crucial to any prediction of how the word would be rated on the related U7 scales. For example, a S would probably experience difficulty in locating HOSPITAL at only one point on the B7 good-bad scale because a HOSPITAL as a service institution is good, but the concept of finding oneself in a HOSPITAL is bad. HOSPITAL, in other words, involves both good and bad aspects. This difficulty is exemplified by the problem of interpreting any rating at the middle position of a B7 scale.

A rating in this position could indicate that (1) the scale is irrelevant or antithetic to the concept being rated, (2) the concept is neutral on the particular scale, or (3) both adjectives of the scale are equally positively related to the concept being rated. Any prediction as to the location of such a word on the relevant U7 scales would obviously depend primarily on whether the neutral B7 rating resulted from either of the first two situations or the latter one.

Another difficulty regarding the B7 scales occurs in creating unipolar scales. The use of a U7 scale, as was done by Green et al. (1965), is based on the implicit assumption that on the B7 scale a *S* is rating the relative importance of both anchoring adjectives. Under this assumption a rating of 4 would typically be interpreted to indicate that both adjectives are equally positively (or negatively) related to the word. On the other hand, if it is assumed that as *S*'s rating deviates from 4, it indicates only an increasingly positive relation between the rated word and one of the polar adjectives, then a 4-position, unipolar adjective scale (U4) such as the following would be more appropriate:

good

_____:_____:_____:_____

where the left end indicates complete neutrality (or irrelevance), and the right end indicates an extremely positive relation between the rated word and the anchoring adjective.

In the present study the emphasis was on the words being rated and how the mean ratings of those words on B7 scales was related to the U4 and U7 ratings of the same words. A total of 96 words were chosen with emphasis on obtaining a substantial range of rated meaningfulness on the B7 scales. Ten B7 scales were chosen including all scales used by Zippel (1967) and Kasschau (1969) representing a balanced selection of scales with 3 loading primarily on the Evaluative factor, 2 each on Potency and Activity, and 1 each on Tautness, Novelty, and Receptivity. The *a priori* hypotheses being tested have been listed in conjunction with the appropriate statistical tests in the results section.

Method

Design. Five groups of *Ss* (N = 33 each) rated each of 96 words on 10 SD scales in one of the following arrangements: (1) B7 scales (*good-bad*, *slow-fast*, etc.), (2) one set of U7 scales (*good*, *slow*, etc.), (3) the complementary U7 scales (*bad*, *fast*, etc.), (4) one set of U4 scales (*good*, *ugly*, etc.), or (5) the complementary U4 scales (*bad*, *beautiful*, etc.) for a total of 960 word-scale ratings per *S*. The ratings were summarized in five separate 10 scale by 96 word tables in which were reported the average rating of each word on each scale. Five tables were also prepared summarizing the standard deviations of these means, and five additional tables reported the average Polarity Score (PS) of each mean, where PS is the distance of a mean from the most neutral scalar rating position.

Subjects. A total of 165 undergraduates enrolled in the introductory psychology course at the University of South Carolina participated in the experiment in five separate groups of 33 in partial fulfillment of a course requirement. In the final groups there were 38 females (mean age: 19.1) and 112 males (mean age: 19.9) assigned at random to the groups in the approximate ratio of 1 female to 3 males.

Apparatus and Materials. The 96 words to be rated were chosen from the lists of Jenkins, Russell, and Suci (1958), Noble (1952), Jenkins (1960), and Gerow and Bryant (1968) in accord with the following criteria in decreasing order of importance. The words were to be (1) three syllables or less, (2) primarily used as nouns or adjectives, (3) of relatively low or high D_4, and (4) of average associative meaningfulness.

The B7 scales were chosen so as to represent a balanced selection of B7 scales with loadings on each of the major, most widely confirmed factors of semantic space. The 10 scales selected are listed in Table 1. For factors represented by more than one scale the B7 scales were balanced so that, for example, if fast was on one side, active was on the other. The U4 and U7 sets were composed by using the adjectives on the left end of the B7 scales as one set; those on the right as the other. Regardless of the bipolar or unipolar format, these ten scales were printed on 8½ by 11-inch mimeograph paper using a different order for each word. However, across groups the order of scales for any word was identical. Centered above the top scale on each sheet was one of the 96 words to be rated.

The ratings were collected from each S in two booklets. The first booklet contained two pages of instructions modified as necessary from Green et al. (1965). This was followed by two sheets containing practice words and 10 scales of the appropriate format. For all 165 Ss the first practice word was HAM, the second was VIET NAM. Next was a sheet instructing Ss to wait until told to proceed, and suggesting that after finishing the first book S should take a short break before completing the second book. Following this sheet were 50 sheets containing one word and 10 scales each. The second booklet had a cover sheet and the remaining 46 words. The 96 words were randomly ordered with 5 Ss sharing an identical order.

Procedure. The data were collected on three successive days from groups of Ss. Assignment of Ss to groups was random. The first day 33 Ss did the ratings on the B7 scales; the second day 66 Ss rated the words on the U7 scales (two different sets of rating scales being used by 33 Ss each), and the third day the final 66 Ss rated the words on the U4 scales. Except for necessary differences in instructions all Ss were treated identically.

After obtaining specific personal information, Ss were encouraged to read along silently while the instructions were read aloud by the experimenter in conjunction with a large semantic differential scale on the blackboard at the front of the air-conditioned room. The positions of this sample scale were specifically labelled neutral or slightly-, quite-, or extremely-positive (or negative, if appropriate) and the scale was left on the board throughout the experiment. Following an opportunity to ask any questions, Ss were allowed 3 minutes to rate the two sample words. Following an additional opportunity to ask questions, Ss were then allowed to work through the booklet at their own speed. The average S required approximately one hour and 20 minutes to complete the ratings, including the break between booklets which the majority of Ss availed themselves of.

Treatment of Data. As necessary 3 Ss/group were deleted (1) for failure to follow instructions, (2) for having other than English as a native language, or (3) at random if an otherwise satisfactory set of ratings resulted. Thus the summary statistics were based on an N of 30/group. The data were put onto Digitek sheets to be converted to IBM cards. The print-out from these cards was double-checked against the original booklets. Following correction of errors of transcription, the summary tables and statistical tests were computed on an IBM 7040 computer.

Results and Discussion

The basic comparisons among the means, standard deviations, and PSs of the B7, U4, and U7 data are reported and discussed first. Following this the U4 and U7 data were rearranged (as will be described) for purposes of testing specific hypotheses regarding some relations between B7 ratings and the appropriate U4 and U7 ratings. The PS value of each mean was calculated by determining the absolute value of each mean rating minus 4.00 for the B7 and U7 groups and the mean ratings minus 1.00 for the U4 groups. Thus the PS value indicates distance of each mean rating from the most neutral point of each rating scale, i.e., it is a measure of rated intensity of meaningfulness.

In terms of initial analyses of the data direct comparisons of 4- and 7-position scale ratings are not legitimate except in the instance of the PS measures, where values from any S can vary only between 0 and 3. Collapsing across scales and groups (for the U4 and U7 subgroups), the average PS for words rated on the B7 scales was .98, that for the U7 scales was .83, and that for the U4 scales was 1.01. A simple randomized design analysis of variance performed on the average PS per word as a function of B7, U7, and U4 scales indicated these means were significantly different, F (2,285) = 9.57, $p < .001$. Duncan's multiple-range test indicated the average U7 PS differs ($p < .001$) from the average U4 and B7 PS which did not differ significantly from each other.

Similarly, collapsing across scales and groups the average standard deviation for the ratings of words on the B7 scales was 1.35, that for ratings on the U7 scales was 1.46, and that for ratings on the U4 scales was 0.97. Of particular interest was a comparison of the standard deviations of the B7 and U7 ratings. The average difference in the variability of ratings for the 96 words on B7 as opposed to U7 scales was significant, $\underline{t}$ ($\underline{df}$ = 190) = 6.25, $\underline{p} < .001$. This confirmed a suggestion by Heise (1969) that U7 ratings might have more sources of variance than ratings on scales which were doubly anchored.

With this set of 96 words, use of U7 scales resulted in ratings which showed a greater degree of variability in ratings and did not deviate as far from the most neutral position relative to the comparable B7 rating. Likewise use of the U4 scales resulted in no significant difference in the average PS per word as compared to B7 scale ratings, and only slight improvement in the variability of the ratings.

In a related program of research, Luria (1959) reported that summing across scales, words rated closer to 4 on B7 scales are less reliable and more likely to shift in an immediate retest. He reported a -.81 rho between the average test-retest shift and extremity of score on the initial test. In the present results, summing across scales, a Pearson product-moment correlation of -.51 ($\underline{df}$ = 94, $\underline{p} < .001$) was obtained between the average B7 PS of each word and the average standard deviations of the mean ratings which generated that PS. In other words, as the intensity of rated meaningfulness increased, the agreement across $\underline{S}$s as to where the word should be rated also increased.

Interestingly, the same relation was nonsignificant for the U7 ratings where a correlation of -.06 was obtained between the average U7 PS of each word and the average standard deviations of the mean ratings which generated the PS. However for U4 ratings, where PS was simply the distance of a rating from the no-relation or neutral rating of 1, the same correlation was -.30 ($\underline{df}$ = 94, $\underline{p} < .005$).

It was also expected, given that the unipolar scales were derived from B7 scales used in this study, that the B7 PS values for each word should likewise be negatively correlated with the U7 standard deviations. Collapsing across scales, the Pearson product-moment correlation between the B7 PS of each word and the average deviation of the mean U7 ratings was nonsignificant, $\underline{r}$ = .02 ($\underline{df}$ = 94). However, the same correlation between B7 PS values and the U4 standard deviations of the mean ratings was significant, $\underline{r}$ = -.51 ($\underline{df}$ = 94, $\underline{p} < .001$), indicating that as the B7 PS of a word increased, $\underline{S}$s rating the word on U4 scales showed increasing agreement as to its rating.

At this point relative to B7 ratings, words rated on U7 scales showed less rated intensity of meaningfulness and greater disagreement across $\underline{S}$s as to the precise U7 rating of each. In addition, neither B7 PS nor U7 PS demonstrated any consistent relationship to the agreement $\underline{S}$s showed in making U7 ratings. Furthermore, also relative to B7 ratings, words rated on U4 scales showed no significant increase in rated intensity of meaning. This increase had been expected as a result of removing the rating positions which exist on the opposite side of the neutral position in both the unipolar and bipolar seven position scales. Likewise, it should be noted that despite significant correlations between both U4 and B7 PS values and the average standard deviations of the U4 mean ratings, the B7 standard deviation (1.35) was derived from a seven position scale whereas the U4 standard deviation (0.97) was deprived from a four position scale.

One factor that might have contributed to the magnitude of the U4 standard deviations was the fact that, for example, all the ratings which would have been to the left of the neutral position on the good-bad B7 scale might end up in position number 1 on the bad scale. That is, all ratings that are neutral on a given U4 scale as well as those ratings of words with meaning directly contrary to the rating scale adjective in a U4 scale might end up in position 1. To check this possibility a random sample o f approximately 10% of each S's ratings of each word on each scale was drawn. For the B7 data, approximately 18% of all ratings fell in the neutral, most meaningless positions; for the U7 data, 26%, and for the U4 data, 44%. Based on these comparisons the B7 scale yielding high average PS and relatively good inter-S agreement seems preferable as a measure of semantic meaningfulness.

This latter analysis raised a problem with the unipolar ratings which was both correctable and offered opportunities for using the unipolar rating scales to ascertain what Ss were indicating when they rated a word near the middle of a B7 scale. The tables of means, standard deviations, and PSs of the two groups of U7 and U4 ratings were rearranged in a manner best explained using an example. The mean B7 rating of each word on each scale was used to define one U7 and U4 scale as the more relevant (MR) scale and the complementary U7 or U4 scale as the less relevant (LR). The mean rating of ABORTION on the good-bad scale was 5.00. Thus, for ABORTION the bad U7- and U4-scale was the MR scale and the good U7- and U4-scale was the LR scale. By contrast, the mean rating of BABY on the good-bad scale was 1.80, so the good scale was MR and the bad scale was LR. Proceeding in this manner, using the B7 mean rating, the 10 by 96 tabled ratings of the two U7 groups (one having rated the words on good, slow, etc., the other having rated the same words on bad, fast, etc.) were rearranged into two new tables, one containing MR unipolar ratings, the other containing LR unipolar ragings. Each of the ten columns in each of the newly created tables thus contained ratings from the MR good or bad scale, the MR ugly or beautiful scale, etc. Except as noted, all subsequent analyses have been performed on these reconstituted tables of means, standard deviations, and PSs.

Hypothesis 1. Because of the presumably stronger relationship between the MR scale and the meaning of the words being rated, the average PS of each word rated on an MR scale should be greater than the average PS of that word on an LR scale for both U4 and U7 scales. The data on which the test of this hypothesis was based is reported in Table 1 which lists both the average PS and the average standard

Insert Table 1 about here

deviation of the mean ratings of all words on each U4 and U7 scale as a function of the 10 B7 reference scales. Each of the analyses of variance performed on the data with reference to Hypotheses 1 and 2 was a two-factor analysis including two levels of relevancy and 10 different B7 reference scales. Although the 10 scales obviously composed a within-Ss effect, reconstitution of the original tables of PS, standard deviations and means confounded this aspect of the design. The two-factor analysis is thus slightly conservative in its ability to detect differences in this experiment, but the magnitude of the effects of primary interest in Table 1, those defined by Hypotheses 1 and 2, are such that use of this analysis is unlikely to lead to spurious conclusions.

For the U4 scales, the average PS per word across all MR scales for each S was 1.35; that for the LR scales 0.66, and the difference was highly significant, F (1,1900) = 1120.47, p < .001. Similarly for the U7 MR scales the average PS was 1.01, that for the LR scales was 0.65, and the difference was again highly significant, F (1,1900) = 172.36, p < .001. In both of these analyses the average PS differed significantly as a function of the B7 reference scale; for U4, F (9,1900) = 18.11, p < .001, and for U7, F (9,1900 = 27.01, p < .001. Likewise, the interaction of scale relevancy by scale was also significant in both analyses; for U4, F (9,1900) = 13.80, p < .001, and for U7, F (9,1900) = 6.10, p < .001.

Quite obviously regardless of which unipolar measurement form was used the MR scales consistently yielded larger average PSs. Furthermore, the significant interaction indicated the difference in average PS differed across B7 reference scales, but the consistency and magnitude of the MR-LR difference was a clear confirmation of Hypothesis 1.

Hypothesis 2. The ratings of each word should tend to cluster closer to the most neutral, meaningless rating position on both the U4 and U7 LR scales. As a result, the average standard deviation of the mean rating of each word on the LR scales should be less than that for the ratings on the MR scale for both U4 and U7 scales.

The average standard deviation for all words rated on the MR U4 scales was 0.98; that for the LR U4 scales was 0.85, and the difference was highly significant, F (1,1900) = 273.25, p < .001. By contrast, for the U7 scales, the average standard deviation for all words on the MR scales was 1.55; for the LR scales it was 1.38, and the difference was highly significant, F (1,1900) = 186.01, p < .001. As was true for the PS analyses, in both of these analyses the magnitude of the average standard deviation differed significantly as a function of the reference B7 scale; for U4, F (9,1900) = 7.92, p < .001, and for U7, F (9,1900) = 2.39, p < .025. Similarly, the interaction of scale relevancy by scale was also significant in both analyses; for U4, $\underline{F}$ (9,1900) = 9.08, $\underline{p} < .001$, and for U7, $\underline{F}$ (9,1900) = 9.05, $\underline{p} < .001$.

The significant interaction indicated the amount of disagreement across Ss (as reflected in the average standard deviations of the unipolar ratings) differed for different B7 reference scales. Again, however, the consistency and magnitude of the MR-LR difference should be noted. The apparent reversal, in contrast to Hypothesis 2, manifested by the U7 data seemed to result from Ss' considerable confusion when trying to rate words as to the degree of negative relation to U7 scales which subsequent analysis demonstrated to be LR scales. This apparently produced the larger average standard deviation for the LR scale ratings.

The final analyses attempt to corroborate the precise relation between the B7 PS ratings and various measures on the two related U7 scales. Particular emphasis was placed on demonstrating that words with a low B7 PS relate differently to the U7 scales than do words with a high B7 PS.

Hypothesis 3. Again referring to the original tables of mean ratings and PS values, the theoretical position developed by Green and Goldfried (1965) regarding the U7 scales would indicate that as the rating of any word deviated from the mid-point on one U7 scale it should have deviated in the opposite direction on the complementary U7 scale if the assumption of bipolarity is valid. As a test of this, a Pearson product-moment correlation coefficient was calculated involving all of the 960 pairs of ratings of each word on the pairs of complementary U7 scales. The resulting correlation was -.57 which was highly significant ($\underline{df}$ = 958, $\underline{p} < .001$), indicating that across all words and scales as the rating increased on one U7 scale it decreased to a significant degree on the complementary scale.

Hypothesis 4. Given the significance of the above correlation it should also have been true that as the B7 PS increased, the MR U7 mean should have increased and the LR U7 mean should have decreased. Pearson product-moment correlations were calculated based on each of the 960 word-scale B7 PS values and the MR and LR U7 means. Both of these correlations were highly significant in the expected direction. For the B7 PS-MR U7 mean pairs the correlation was +.68 ($\underline{df}$ = 958, $\underline{p} < .001$) indicating that as the B7 PS increased so did the average U7 mean rating on the MR scales. For the B7 PS-MR U7 mean pairs the correlation was -.59 ($\underline{df}$ = 958, $\underline{p} < .001$) indicating that as the B7 PS increased the average U7 mean rating on the LR scales decreased.

Hypothesis 5. Referring again to the original tables of means, standard deviations, and PS values, it was desired, in light of the preceding three correlations, to determine the extent to which increases in B7 PS were related to increases in divergence between the mean ratings on each pair of complementary U7 scales. Thus a correlation was performed between the B7 PS value of each word-scale rating combination and a Disagreement

Measure designed to increase with increasing disparity of the mean ratings of any word on the complementary U7 scales.

The Disagreement Measure was defined for each word as the absolute value of the square of one U7 mean subtracted from the square of the complementary U7 mean (e.g., $|U7^2_{good} - U7^2_{bad}|$). This measure could vary from 0 (if any word had identical mean ratings on both complementary U7 scales) to 48 (if any word had a mean of 7.00 on one U7 scale and 1.00 on the complementary U7 scale). In general, as the difference between the mean ratings of a word on two U7 scales increased so would the magnitide of the Disagreement Measure. The obtained correlation between each word's B7 PS and the Disagreement Measure was +.78 (df = 958, $p < .001$).

The correlations reported up to this point have considerably clarified the relation between B7 PS and the U7 rating scale values, but the primary question remaining was best answered by means of a frequency count. The final analysis was directed toward determining what proportion of the words alleged to be difficult to locate in one position on a B7 scale (e.g., HOSPITAL on the good-bad scale) will produce ratings indicating a positive relation of the word to both of the complementary U7 scales.

Hypothesis 6. The preceding correlations were all consistent with the expectation that the ratings of any word on the MR U7 scale should be greater than 4.00 and those of the same word on the LR U7 scale should be less than 4.00. There were, however, three other situations ("errors") which might occur. First, as in the instance of HOSPITAL on the good and bad scales, both U7 ratings might be greater than 4.00 indicating a positive relation of the rated word to each scalar adjective. This would seem most likely when words had a B7 mean rating near 4.00 because both adjectival anchor words were equally positively related. Second, both the MR and LR U7 mean ratings might be less than 4.00 as would be expected in those instances where a word was rated near 4.00 on the B7 scale because the scale was irrelevant. Finally, the MR U7 rating might be less than 4.00 and the LR rating might be greater than 4.00, in which case both U7 ratings were "incorrect" with reference to the B7 ratings. This would seem most likely in those instances where the B7 scale was meaningless for purposes of rating a word.

The sixth hypothesis, then, was that the majority of LR and MR U7 ratings should have been "correct" in terms of the mean B7 rating and that the percentage which were correct should have increased as the B7 PS increased. Further, of the three types of "errors" which might occur, that in which both the MR and LR U7 ratings deviate to the positive-relation side of 4.00 should have been the most likely to occur, and finally, the likelihood of any type of error occurring should have decreased as the B7 PS increased.

To test the sixth hypotheses the 960 word-scale B7 PS values were rank-ordered from lowest (least intensity of rated meaningfulness) to highest and then divided into quartiles. Within each quartile the pairs of U7 ratings associated with each B7 PS value were tallied as "correct" if the MR rating was $\geq$ 4.00 and the LR rating was $\leq$ 4.00, "Both +" if both the MR and LR U7 ratings were $\geq$ 4.00, "Both -" if both the MR and LR U7 ratings were $\leq$ 4.00, and "incorrect" if the MR U7 rating was $\leq$ 4.00 and the LR U7 rating was $\geq$ 4.00. The results of the tally are reported in Table 2. A X^2 test of independence was

----

Insert Table 2 about here

----

performed on this data and yielded a value of 191.01 which with 9 df was significant at far beyond the .001 level.

All major aspects of the sixth hypothesis were confirmed. Roughly 60% of the U7 deviations from 4.00 could be predicted knowing the B7 mean rating. Further, the most common "error", which occurred in approximately 27% of all pairs, was that in which the word was rated, however slightly, as positively related to both of the anchoring adjectives used in the B7 scale. With one minor reversal (in the "Both +" column comparing the lowest and second lowest quartiles) the likelihood of making any error decreased consistently as the B7 PS increased.

General Discussion

Two relatively distinct topics were involved in the present research: (1) a comparison of U4, U7, and B7 scales with reference to the type of information yielded by each scale type, and (2) use of unipolar rating scales to detail certain aspects of ratings typically collected on the standard B7 scales.

With reference to the first topic numerous findings in the present report argue in favor of retention of the bipolar scale form. First, the average rated intensity of semantic meaningfulness was significantly greater on the B7 scales than on the U7 scales. Second, as compared to the U7 scales, ratings on the B7 scales indicated a higher agreement across Ss as to the rating of words, consistent with a suggestion by Heise (1969) that U7 scales might be subject to more sources of variance. Similarly, as compared with B7 scales, use of U4 scales produced no reduction in the standard deviation of each mean rating over and above what the shift from a seven- to a four-position scale would have been expected to produce. Third, the U4 scales did not yield significantly more polarized average ratings despite the absence of anchoring adjectives of opposite meanings. Thus the U4 scales, while involving only the degree of positive relation between a word and one scalar adjective, did not reveal any information not also detailed by B7 ratings. Further, it was indicated that of all the scale types, the U4 scales yielded the highest percentage of ratings in the most neutral, meaningless position. These rating on all scale types convey the least positive information, or are subject to the most alternative explanations.

Fourth, the intra-scalar relationship between rated intensity of meaning and agreement across Ss as to the rated position was highest for the B7 scale. Finally, it was noted independently by all four experimenters in attendance while the B7, U4, and U7 ratings were collected that Ss asked the least questions and evidenced the least difficulty in understanding use of the B7 rating scales. All these findings argue in favor of retaining the

B7 rating scale format as opposed to either the U4 or U7 format.

With reference to the second topic, despite Green and Goldfried's (1965) difficulties in demonstrating what they considered consistent performance by Ss on the U7 scales, some very substantial correlations (none less than -.57) were obtained in the present study which collectively demonstrated that the position of a word on the B7 scale was very strongly related to its ratings on the U7 scales. In particular, the higher the B7 PDS of a word, the greater the likelihood the U7 ratings would deviate in opposite directions from 4 on the U7 scales generated from that B7 scale (with the MR U7 scale demonstrating a positive relation between the rated word and the anchoring adjective) and the less likelihood any "error" (as previously defined) would be committed. It should be noted also that these correlations in which B7 PDS values have been related to MR and IR U7 ratings were based on three independent groups of Ss, not on within-Ss comparisons.

The major difficulty with the assumption of bipolarity results from the three disparate sets of circumstances which can result in a word being rated at or near 4 on the B7 scale, viz., equal applicability of both adjectives to the word being rated, equal inapplicability (irrelevance) of both adjectives, or genuine neutrality of the word being rated on the B7 scale of concern. The "Both +" and "Both -" errors tallied in Table 2 cause difficulty when attempting to use B7 ratings to predict the U7 ratings. However, both these "errors" result from the positive or negative relation of a word to both anchoring adjectives of a scale—"errors" which are not contrary to the basic tenet of the assumption of bipolarity. In fact, the only "error" totally contrary to expectations generated from the mean B7 ratings were the U7 ratings in which the MR mean was less than 4 and the IR mean was greater—an error accounting for only about 5% of all 960 word-scale combinations.

More work is obviously needed to isolate and identify the causes of some words being rated near 4 on B7 scales, but in terms of which type of scale, B7, U4, or U7 is best for obtaining measures of rated semantic meaningfulness, the present results clearly indicate the standard bipolar, seven-position semantic differential scale.

References

Gerow, J. R. III, & Bryant, R. P. Word association, frequency of occurrence, and semantic differential norms for 116 stimulus words. University of Colorado, Technical Report No. 2, 1968.

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

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

Jenkins, J. J. Degree of polarization and scores on the principal factors for concepts in the semantic atlas study. American Journal of Psychology, 1960, 73, 274-279.

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.

Kasschau, R. A. Semantic satiation as a function of duration of repetition and initial meaning intensity. Journal of Verbal Behavior, 1969, 8, 36-42.

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

Noble, C. E. The role of stimulus meaning (m) in serial verbal learning. Journal of Experimental Psychology, 1952, 43, 437-446; 44, 465.

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

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

Zippel, B. Semantic differential measures of meaningfulness and agreement of meaning. Journal of Verbal Learning and Verbal Behavior, 1967, 6, 222-225.

Table 1

Average Polarity Score and Standard Deviations of Ratings on Each More- and Less-relevant U7 and U4 Scale

B7	U7 Scales				U4 Scales			
Reference	Polarity Score		Standard Deviation		Polarity Score		Standard Deviation	
Scale	MR	LR	MR	LR	MR	LR	MR	LR
good-bad	1.44	.84	1.25	1.67	1.70	.64	.92	.81
true-false	.91	.57	1.31	1.54	1.24	.54	1.08	.78
ugly-beautiful	1.20	.99	1.48	1.59	1.57	.54	.93	.79
masculine-feminine	1.19	.61	1.33	1.57	1.44	.68	.95	.86
soft-hard	1.16	.72	1.36	1.54	1.45	.65	.97	.82
active-passive	1.20	.70	1.34	1.59	1.48	.71	.95	.86
slow-fast	.85	.54	1.39	1.59	1.25	.73	.98	.88
angular-rounded	.59	.33	1.44	1.46	1.03	.68	.99	.90
new-old	1.08	.57	1.36	1.50	1.38	.90	1.00	.96
savory-tasteless	.46	.58	1.50	1.47	.94	.52	1.06	.83
Average	1.01	.65	1.38	1.55	1.35	.66	.98	.85

TABLE 2

Distribution of agreement of complementary U7 ratings as a function of average intensity of B7 ratings

Quartile of B7 Polarity Scores	Agreement of U7 Ratings*				
	"Correct"	Both +	Both -	"Incorrect"	Total
Low	93	86	34	31	244
2nd	107	95	28	11	241
3rd	155	56	16	3	230
High	219	20	5	1	245
Total	574	257	83	46	960
Percent	59.8	26.8	8.6	4.8	

* Note: See text for explanation

SEMANTIC SATIATION AS A FUNCTION OF INITIAL MEANING INTENSITY AND UNIPOLAR VS BIPOLAR RATING SCALES

Richard A. Kasschau

University of South Carolina

Abstract

Loss of meaningfulness resulting from repetition was measured on bipolar, 7-position semantic differential (SD) rating scales and compared with satiation as measured on unipolar, 4- and 7-position scales as a function of three levels of meaning intensity for each of the three major factors. Sixty-six *S*s rated 36 words on a number of SD scales of a particular type, then repeated the words orally for 30 sec. each, and rerated them on 1 of 6 SD scales of the same type. Both mean difference and polarity difference scores were used to assess satiation. Although not all effects were significant, the trends across both measures were consistent in showing that the greatest satiation effect was detected by the Activity scales, words of higher initial meaning intensity showed a greater satiation effect, and the unipolar, 7-position scale indicated the largest shifts in meaningfulness. Theoretical implications of the results were discussed in terms of the two techniques of difference score analyses and the assumption of the bipolarity of SD scales.

Most recent investigations of semantic satiation have used the semantic differential (SD) technique to measure changes in meaning, and satiation has been defined as a postrepetition decrease in rated meaning intensity, i.e., the postrepetition ratings move closer to 4. Heise (1969) has indicated how crucial it is to this method of measuring satiation that the point defining meaninglessness be at the mid-point of the SD rating scales in order that movement toward the mid-point be interpreted as loss of meaningfulness. Definition of this mid-point as the point of meaningfulness is an obvious extension of the assumption of bipolarity according to which "...the polar terms...are true psychological opposites, i.e., fall at equal distances from the origin of the semantic space and in opposite directions along a single straight line passing through the origin" (Osgood, Suci, & Tannenbaum, 1957, p. 327). This assumption is, in a sense, built in to Ss using the SD by instructions which specifically label the middle scale position as neutral.

A difficulty associated with the interpretation of a rating of 4 on the standard bipolar, seven-position (B7) SD scales has recently been raised with reference to semantic satiation by Schulz (1967) and more fully discussed by Kasschau (1969). The particular difficulty is caused by the fact that on a B7 scale as a word moves toward neutrality from one extreme it does reflect loss of meaningfulness relative to the anchoring adjective at that end of the scale. However, at the same time that movement may also reflect gain in meaningfulness relative to the adjective at the opposite end of the SD scale. If so, then removing the second, competing adjective by using unipolar scales of the form suggested by Green and Goldfried (1965) ought to result in an increment in the amount of satiation which can be demonstrated.

Kasschau (1970) has indicated that use of the unipolar, seven-position (U7) rating scales, which range from an extremely negative relation through neutrality to an extremely positive relation between the word being rated and the anchoring adjective, is predicated on the (implicit) assumption that on the B7 scale a S is rating the relative importance of both anchoring adjectives. Just as likely is the possibility that as a B7 rating deviates from 4 it indicates only an increasingly positive relation between the rated word and one of the polar adjectives. If so, then a unipolar, four-position (U4) scale ranging from neutrality to an extremely positive relation would be more appropiate.

The second independent variable was the initial meaning intensity of the words to be repeated as indexed by an independent collection of norms (Kasschau, 1970) in which 96 words were rated on 10 B7 scales and 20 complimentaryU4 and U7 scales generated from the B7 scales. This variable originally suggested by Amster (1964) was included with particular interest in replicating the findings of Kasschau (1969) that initial meaning intensity influenced the magnitude of satiation obtained only when the Mean Difference Score (MDS) technique of analysis (Yelen & Schulz, 1963) was used but not when the Polarity Difference Score (PDS; Lambert & Jakobovits, 1960) was used with B7 data. It was of further interest to determine whether this finding could be extended to data collected using unipolar scales.

Two different B7 scales with primary loadings on each of the three major factors of semantic space were included as the third independent variable. This was done primarily to assure measurement of meaning change on each dimension, but also to extend the findings of Kasschau (1969) to the effect that words measured on B7 scales which loaded on the Activity factor showed the greatest amount of satiation. It was of interest to see whether this effect would extend to the unipolar scales.

The experimental design, then, involved two within-Ss variables (three levels of prerepetition meaning intensity: 0.5, 1.5, and 2.5 units distant from neutrality; and three dimensions of semantic space: Evaluation, Potency, and Activity, represented by two SD scales each) and one between-Ss variable (type of scale on which the ratings were made: B7, U4, and U7).

At the same time an investigation of how aspects of the rating scale format may influence satiation is appropriate, an investigation is also appropriate into aspects of the ratings of the words themselves. Schulz (1967) noted the surprising fact that the likelihood that a rating is changed, for whatever reason, was apparently unrelated to the intensity of meaning it reflected. Another aspect of the ratings of words which may bear on the likelihood of that word changing its rated meaning is the disagreement (as indexed by the standard deviation of the prerepetition ratings of a word) manifested by Ss as to where a word should be rated. Possibly, a small prerepetition standard deviation may indicate words of which Ss are more sure of the meaning, or of which the meaning is more firmly held. Without taking repeated prerepetition measures, use of the standard deviation as an index of disagreement necessitates the assumption that a between-Ss measure is indicative of a within-Ss process (viz., uncertainty as to where to rate a word). However, this assumption is not without precedent and support in tangentially related work concerning another measure of meaningfulness (Laffal & Feldman, 1962).

If the disagreement between Ss as to the prerepetition rating cf a word does influence the amount of satiation demonstrated by each word, then a positive correlation should obtain between these two measures. On the assumption that words with small standard deviations indicate words of which Ss are more sure of the meaning it was expected that words with small standard deviations would show a lesser tendency to satiate than words with larger standard deviations.

Method

Although the details vary considerably, the apparatus, materials and procedure of the present study are similar to those reported by Kasschau (1969). The details are repeated here for purposes of clarity.

Subjects. A total of 66 undergraduates enrolled in the introductory psychology course at the University of South Carolina participated in the experiment in three groups of 22 in partial fulfillment of a course requirement. The final groups were composed of 17 females (mean age: 20.2) and 49 males (mean age: 18.9).

Apparatus and Materials. As an aid to the subsequent choice of words to be satiated, the six B7 scales (Evaluation: good-bad and ugly-beautiful; Potency: soft-hard and masculine-feminine, and Activity: active-passive and slow-fast) were chosen from among those included in the report of Kasschau (1970) in which 96 words were rated on B7, U4, and U7 scales. The B7 scales chosen were used to generate 12 U4 and U7 scales. Thirty-six words were selected such that on the B7 scale they had (a) a similar rating on both scales having a common, underlying factor, and (b) a minimum average meaning intensity on each of the other four scales. Using the Kasschau (1970) norms six words were selected to be rated on each of the six critical B7 scales such that the subset of six words covered the entire range of a particular B7 scale. The 36 words selected are listed in Table 1, paired with the B7 scale on which they were measured. Whereas the original selection of 36 words was based on previously collected data, all subsequent reference to the intensity of meaning of these words is with reference to the prerepetition ratings by the 66 Ss of the present study except as noted.

The More Relevant (MR) unipolar scale for each word was defined as the unipolar scale anchored by the adjective on the same side of 4 as the average B7 rating of the word. Thus, for SUNLIGHT, the mean prerepetition rating of which was 1.36 on the good-bad B7 scale, the MR U4 and U7 scales were anchored by good. By contrast, for RANCID, the mean rating of which was 6.05 on the good-bad B7 scale, the MR U4 and U7 scales were anchored by bad. Thus, initial selection of the 36 words was based on previously collected norms, but definition of meaning intensity was done after the experiment on the basis of the prerepetition ratings of each of the words on the appropriate scale. For the groups using unipolar scales for their ratings, only the MR scales were used, defined a priori by the B7 ratings of Kasschau (1970).

The S's prerepetition ratings of all words were obtained by use of a 49-page booklet. Each page following the cover sheet contained one word and five SD scales of the appropriate format, including the one scale on which satiation was later to be measured. The first and last two words were drawn from the surplus pool; the remaining eight surplus and 36 critical words were randomly arranged in the intervening pages. For the B7 scales, each critical SD scale appeared eight times and 24 additional scales also appeared eight times for a total of 240 individual word-scale ratings. For the booklets containing unipolar scales each critical MR SD scale appeared four times and the complementary less relevant (LR) adjective scale also appeared four times. Forty-eight additional unipolar scales (generated from the additional B7 scales above) likewise appeared four times for a total of 240 individual word-scale ratings.

Pacing for the repetition task was provided by a Hunter Timer which activated a Selmer electronic metronome which emitted a click twice per second.

The critical words and scales were typed on white cards and mounted in overlapping clear plastic holders which assured presentation of only one word and/or scale at a time. Two columns of such holders, containing the words to be repeated and separated by a third column with the SD scales, were mounted on a 15 X 24 inch board for presentation to S. There were two appropriate sample scales included in the center column along with six B7 scales for the group using the bipolar scales and 12 U4 or U7 scales for the groups using the unipolar scales. The SD scales were identical in form to those in the booklet except that the numbers 1-4 or 1-7 as appropriate had been added below the blanks.

Procedure. Data were gathered from Ss on an individual basis. Upon arrival for the experiment Ss were assigned to the B7, U4, or U7 group with only two restrictions: (1) that the ratio of males to females be approximately constant in all groups, and (2) that N Ss had been assigned to each group before the N + 1st S was assigned to any group. Each of the 36 words was repeated once by each S, and the random order in which the words were repeated was changed after every sixth S.

Prior to opening the rating booklet, Ss were read instructions adapted from Green et al. (1965) in conjunction with a sample rating booklet containing appropriate examples for rating each word on five SD scales.

After S completed the 240 ratings, further instructions were read regarding paced repetition and oral SD ratings. This included exposure of both the first example word (NURSE) and the dull-sharp scale (for B7 Ss) or the sharp scale. After any questions had been answered, S practiced repeating NURSE for 30 seconds, while fixating the word being repeated, and then rated it on the wise-foolish (or wise) scale. This was followed immediately by repetition of the second example word (TABLE) which was then rated on the usual-unusual (or usual) scale. The S was then stopped and given the opportunity to ask any additional questions. Without any further interruptions, the process of expose, repeat, rate, pause, expose, etc. was administered for the 36 words previously rated.

Analysis of Data. Three steps were taken to simplify the analysis. First, the absolute value of initial meaning intensity was used so only three levels per scale remained, defined in terms of the prerepetition rated intensity of meaning for each group of Ss. The neutral point was defined as rating position 4 for the U7 and B7 scales and rating position 1 for the U4 scales.

Second, satiation as a function of the major semantic factors was represented by Evaluation, Potency, and Activity, rather than by six individual scales.

Third, the distinction between the MDS technique of Yelen and Schulz (1963) and the PDS technique of Lambert and Jakobovits (1960) exists only for the data collected on the seven-position scales, since the different techniques of analysis yield identical results from the U4 data. Hence, the resulting design has a 3 X 3 X 2 or 3 mixed factorial with three levels of initial meaning intensity, three levels of SD factors (both within-Ss variables) and two levels of the between-Ss or rating scale variable for the MDS technique and three levels of the between-Ss factor for the PDS technique.

Results

With regard to the data two points deserve mentioning: First, the three subsets of 12 words representing different levels of meaning intentity as chosen from Kasschau (1970) had average ratings 2.02, 1.53, and 0.75 units removed from 4 on the B7 scales. The current prerepetition ratings yielded subsets 2.25, 1.60, and 0.58 units removed from 4 on the B7 scales. The rank-order correlation coefficient between the 1970 and current ratings was + .95 ($p < .001$). The corresponding values from the U7 scale data were 2.91, 2.32, and 0.98; and 2.67, 2.27, and 0.32 with the correlation being + .88 ($p < .001$). Finally, the corresponding values from the U4 data were 2.13, 1.61, and 1.18 and 2.34, 2.08, and 1.62 with the correlation being + .95 ($p < .001$).

Second, the mean prerepetition ratings for each word on each type of scale are listed in Table 1 in the order of increasing magnitude of rating on the relevant B7 scale in the norms collected by Kasschau (1970). Since initial meaning intensity immediately prior to repetition was of primary interest , it was necessary, in some instances, to rearrange the words that had been expected to represent a given intensity of meaning on a particular scale. Thus, for example, in the U4 data collected on the ugly and beautiful scales CRIMINAL and WATER had been expected to yield the highest prerepetition intensities of meaning, whereas GLOOMY and WATER in fact did so. Thus, for the Evaluative dimension for the U4 data SUNLIGHT, RANCID, GLOOMY, and WATER were the high intensity words, RIGHT, FEAR, PUNGENT, and SYMPHONY were the moderate intensity words, and SMALL, LOW CRIMINAL, and FAR were the low intensity words. The low, moderate, and high intensity subsets of ratings were similarly constituted for the other five or eight combinations of dimension and type of scale.

Insert Table 1 about here

The mean changes in meaningfulness for each word were calculated in terms of both the PDS and MDS measures, with each set of measures analyzed separately. Considering first the analyses of the PDS data, which is summarized in Table 2, the factor loading of the scales used to measure satiation was a significant source of variance, F (2,126) = 3.74, $p < .05$, as was

Insert Table 2 about here

the initial meaning intensity of the words satiated, $\underline{F}(2,126) = 3.81$, $\underline{p} < .025$.

By contrast, the different types of SD scales on which satiation was measured were not found to be a significant source of variance, $\underline{F}(2,63) < 1.00$. None of the double-order interactions was significant, but the triple-order interaction was significant, $\underline{F}(8,252) = 2.42$, $\underline{p} < .025$, although no interpretable pattern could be detected in the data.

The MDS data as summarized in Table 2 was also analyzed. As was true for the PDS data, the factor loading of the scales used to measure satiation was found to be a highly significant source of variance, $\underline{F}(2,126) = 8.03$, $\underline{p} < .001$. Contrary to the findings using the PDS analysis, the initial meaning intensity of the words to be satiated was not a significant source of variance for the MDS data, $\underline{F}(2,84) = 2.10$, $\underline{p} < .20$. The main effect of the type of scale on which satiation was measured, however, was a significant source of variance, $\underline{F}(1,42) = 4.70$, $\underline{p} < .05$. None of the double- or triple-order interactions was significant.

The rank-order correlation between the standard deviation of the prerepetition mean rating on the B7 scale and the amount of satiation evidenced by each word as rerated by the B7 group was -.02 which was nonsignificant.

Discussion

There are several trends in the data which are consistent across both the PDS and MDS techniques of analysis, albeit not always significantly so. First, the factor loading of the scales was a significant source of variance with Activity scales detecting the greatest satiation. Second, in terms of the initial meaning intensity of the words, the High intensity words showed approximately twice as much satiation as the Low and Moderate words, although only the PDS analysis yielded significance for this variable. Third, satiation as measured on the U7 scales was approximately twice the magnitude of that measured on either the B7 or U4 scales with the B7-U7 difference significant for the MDS data while the B7-U7 and B7-U4 differences were nonsignificant for the PDS data. Finally, for neither the PDS nor MDS data was the degree of disagreement across Ss as to the prerepetition rating of a word significantly correlated with the subsequent loss of meaningfulness demonstrated by the word. This latter finding, while unexpected, did confirm that this between-Ss agreement had no consistent effect on the average satiation within Ss.

Several comments are in order concerning the results of the principal analyses of variance. Each major independent variable is discussed individually.

Scales. That B7 Activity scales detected the largest satiation effect was consistent with the findings of Kasschau (1969). Of particular interest, however, was that unipolar "Activity" scales likewise detected the largest satiation effects as compared to unipolar "Evaluation" and "Potency" scales, i.e., there was no factor loading by scale type interaction. Thus, for example, the good-bad Evaluation scale was arbitrarily split into two separate unipolar scales. Ratings from these scales were subsequently labeled and analyzed as "Evaluative" ratings without having previously factor analyzed the unipolar ratings of a standard set of words on the good and bad scales as well as a number of other unipolar scales. Nonetheless, the relative magnitude of the satiation detected by U4 and U7 scales was consistent with similar findings on the related B7 scales. These findings offered support for the statement of Osgood (1969) that Evaluation, Potency, and Activity are universal features of human semantic sytems. These factors have operated consistently in spite of marked changes in the scale format most typically used to demonstrate the factors' existence.

Initial meaning intensity. Results generally obtained from variations in initial meaning intensity have been mixed and somewhat inconsistent. Jakobovits and Rice (1967) using PDS data did not find any significant differences in magnitude of satiation with three different levels of initial meaning intensity. Kasschau (1969) found no difference in satiation of Low and High intensity words, both of which demonstrated greater satiation than Moderate words, although a consistent positive relation was demonstrated in the same study between initial meaning intensity and magnitude of satiation using MDS data.

The present MDS data were consistent with the MDS analysis of Kasschau (1969). Although the present trend was nonsignificant, it was indicative of the same positive relationship. At all levels the absolute magnitude of the MDS satiation was comparable in both studies, as was the magnitude of PDS satiation for the Moderate and High intensity words. However, the satiation manifested by the present Low words was lower than that of Kasschau's (1969) Low words--sufficiently so as to yield a significant effect for initial meaning intensity.

The point raised by Yelen et al. (1963) concerning the complex regression hypothesis is not to be denied if the MDS technique is used. Likewise, despite some evidence to the contrary, a similar effect may occur using the PDS technique. However, the possible existence of the regression effect would necessitate only use of meaning intensity as another independent variable or control by satiating only words of a given prerepetition intensity. The lack of interaction between meaning intensity and any other variable, save the uninterpretable triple interaction in the PDS data, indicated the effects of other independent variables in the present study were unaffected by any complex regression effect.

Type of scale. Osgood et al. (1957) suggested that use of unipolar SD scales might result in the neutral or no relation end of each scale assuming the same characteristics or meaning as the opposing end of the relevant B7 scale. The results of the present study include several findings which indicate the U4 and U7 scales may differ in the extent to which the neutral position and/or the negative-relation positions assumed the characteristics of the anchoring adjective most opposite in meaning to that of the MR adjective.

First, the similar magnitude of satiation measured by B7 and U4 scales and analyzed using the PDS indicated similar constraints on change in meaningfulness were encountered by Ss using each type of scale. Second, the U7 scale was the only scale type potentially offering 7 distinct ratings with reference to the positive or negative relation between only the MR adjective and the rated word. These U7 scales detected a larger shift in meaningfulness than either the B7 or U4 scales. For the U4 scales obviously it would be rating position one at the left end which would assume any characteristics in opposition to the anchoring adjective. For the B7 scale it could potentially be either position four or one (seven) which, relative to position seven (one), would assume such characteristics. However, in terms of intensity of rated meaningfulness the PDS analysis functionally forces both the U7 and B7 scales into four position scales. Nonetheless, the U7 scales detected the largest shifts in meaningfulness, indicating that in both the B7 and U4 scales whatever operated in opposition to any greater loss of meaningfulness did so (although nonsignificantly) more than on the U7 scales.

Third, and most convincingly, using the MDS data to compare satiation on B7 and U7 scales, where both were seven-position scale types and the MDS analysis did not, for example, collapse ratings at positions one and seven, the U7 scales detected significantly greatersatiation. Removal of the opposing adjective may not have eliminated all oppositional features of the scale end to the opposite side of four from the prerepetition rating of each word. However, it did result in significantly greater shifts in rated meaningfulness on the unipolar scale where shifts of postrepetition ratings into the opposing side carried greater weight than did such shifts in PDS analyses.

In sum, semantic satiation has been demonstrated using both PDS and MDS scoring techniques on B7, U4, and U7 scales with the latter generally detecting the largest loss of meaningfulness. Scales loading on the Activity dimension detected the greatest satiation effects, with a positive relation between initial meaning intensity and the magnitude of satiation demonstrated.

References

Amster, H. Semantic satiation and generation: Learning? Adaptation? Psychological Bulletin, 1964, 62, 273-286.

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

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

Jakobovits, L. A., & Rice, U. M. Semantic satiation as a function of initial polarity and scale relevance. Paper read at EPA meetings, Boston, 1967.

Kasschau, R. A. Semantic satiation as a function of duration of repetition and initial meaning intensity. Journal of Verbal Learning and Verbal Behavior, 1969, 8, 36-42.

Kasschau, R. A. Unipolar vs bipolar semantic differential rating scales. University of South Carolina, unpublished manuscript, 1970.

Lamber, W. E., & Jakobovits, L. A. Verbal satiation and changes in the intensity of meaning. Journal of Experimental Psychology, 1960, 60, 376-383.

Osgood, C. E. On the whys and wherefores of E, P, and A. Journal of Personailty and Social Psychology, 1969, 12, 194-199.

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

Schulz, R. W. A reply to Jakobovits and Lambert's note on the measurement of semantic satiation. Journal of Verbal Learning and Verbal Behavior, 1967, 6, 958-960.

Yelen, D. R., & Schulz, R. W. Verbal satiation?? Journal of Verbal Learning and Verbal Behavior, 1963, 1, 372-377.

Table 1

Average Ratings of 36 Words* on B7, U7, and U4 Scales
Immediately Prior to Repetition

Semantic factor and scale	Word	Prerepetition Ratings		
		B7	U7	U4
Evaluation:	Sunlight	1.36	6.32	3.73
good-bad	Right	1.50	6.18	3.68
	Small	3.23	4.05	2.36
	Low	4.82	3.41	2.68
	Fear	4.96	4.32	3.05
	Rancid	6.05	5.68	3.18
ugly-beautiful	Criminal	2.59	3.55	2.64
	Gloomy	1.91	5.36	2.95
	Pungent	2.91	5.00	2.82
	Far	4.77	4.50	2.91
	Symphony	5.18	5.68	3.09
	Water	6.14	6.23	3.32
Potency:	Lovely	1.68	6.36	3.59
soft-hard	Round	2.68	5.59	3.14
	Dream	2.55	4.55	2.77
	Winter	5.14	5.41	3.09
	Wagon	5.91	5.23	3.14
	Rugged	6.23	6.45	3.54
masculine-feminine	Sword	1.50	5.91	3.54
	Lift	2.45	5.00	3.04
	Obvious	3.64	4.77	2.59
	Scene	4.23	4.50	2.64
	Wish	4.64	5.23	3.45
	Fragrant	5.55	6.00	3.64
Activity:	Fiery	1.73	6.23	3.68
active-passive	Fervid	2.64	6.05	2.95
	Wet	3.55	4.54	2.45
	Deep	4.41	3.18	2.00
	Plain	5.32	4.77	2.45
	Slow	5.18	4.45	2.68
slow-fast	Somber	2.86	4.27	2.86
	Slack	2.05	4.91	3.00
	Under	3.05	4.18	2.77
	Town	4.68	3.73	2.32
	Hot	5.64	4.95	3.00
	Run	6.82	6.45	3.73

*Note: The words were chosen from Kasschau (1970).

Table 2

Mean Change in Meaningfulness for Both Techniques of Data Analysis as a Function of Each Independent Variable

Independent Variable		Polarity Difference Score	Mean Difference Score
Factor loading of the rating scales	Evaluation	.15	.37
	Potency	.06	.11
	Activity	.20	.45
Prerepetition meaning intensity	Low	.09	.23
	Moderate	.10	.26
	High	.22	.44
Type of rating scale	B7	.11	.18
	U7	.19	.44
	U4	.11	

THE EFFECTS OF AVERAGE DEGREE OF POLARIZATION ON PAIRED-ASSOCIATE AND SERIAL LEARNING AND THE VON RESTORFF EFFECT

Richard A. Kasschau

University of South Carolina

Abstract

In three separate experiments the average degree of polarization (D_4) on the semantic differential of words composing a list was varied, holding associative meaningfulness (m; cf., Noble, 1952) constant. Across all experiments D_4 and number of trials to criterion were significantly positively related. Effects previously demonstrated varying m were essentially replicated holding m constant and varying D_4 with the exceptions that (1) in paired-associate learning variations in stimulus D_4 had a greater effect on performance than similar variations in response D_4, and (2) the von Restorff effect was demonstrated by isolating the 5th item in a 9-item list for both lists of low and high average D_4 in contrast with previous findings (cf., Rosen, Richardson, & Saltz, 1962) where isolation was significant only with low m materials. Implications of D_4 and m as measures of meaningfulness were discussed.

The influence of meaningfulness on the ease of learning verbal material has been explicitly recognized by experimental psychologists since at least 1885 when Ebbinghaus introduced "nonsense" syllables. Having failed in the attempt to create verbal material devoid of meaning, efforts were then made to quantify and measure the relative amounts of meaningfulness possessed by varieties of verbal material. Broadly classified, these many measures may be divided into measures of associative-(cf., Noble, 1952; Deese, 1965, and others) and semantic-meaningfulness (cf., Osgood, Suci, & Tannenbaum, 1957). Certain difficulties have resulted from the absence of a commonly accepted quantitative measure of meaningfulness - an absence at least partly responsible for the large number of different, although often similar, measures which have been developed. Lacking a single acceptable definition of meaningfulness, it would seem advisable to study the similarities and differences in the semantic and associative measures now extant especially to determine the effects which words of varying degrees of meaningfulness (by semantic vis-à-vis associative measures) have on the ease of learning such materials.

The need for such studies is emphasized first by the clear disagreement as to the distinction, or lack of it, between the two types of measures. For example, Deese (1965) after reviewing Bousfield (1961) says, "...Bousfield's results are convincing enough to make it profitable to view the process underlying the semantic differential as an associative one" (p. 72). Osgood (1961), after reviewing the same and related work, says, "...the reactions of subjects to a semantic differential are not predictable from knowing their word-association hierarchies" (p. 105).

Second, Kasschau and Pollio (1967) have demonstrated that with associative or semantic meaningfulness held constant, words varying in the other type of meaningfulness are sufficient to mediate response transfer to a second list with a 41-43% savings. Thus, both semantic and associative measures seem to index aspects of meaningfulness which can influence ease of learning.

Finally, Koen (1962) and Zippel (1967) have reported a Pearson $\underline{r}$ of + .76 between the Noble (1952) association values ($\underline{m}$) of an 18-word sample and the average distance from a neutral rating of 4.00 (D4) of the semantic differential ratings of those words on 20 scales. By contrast, Howe (1965) reported $\underline{r}$ = + .51 and Paivio (1968) reported $\underline{r}$ = + .43 in similar comparisons. These reports indicate that as the association value of the words increased, so did the rated intensity of meaning on the semantic differential, although there has been some disagreement as to the magnitude of the relationship.

In sum, there is evidence to indicate that semantic and associative measures are correlated, but measure distinctive aspects of meaningfulness. To the extent this correlation exists, but falls short of being perfect, there is justification for studying the relation between these measures, especially as they affect the ease of learning material measured by these two techniques.

In each of the following experiments associative meaningfulness has already been shown to be a crucial variable yielding a consistent change in behavior as a function of increments in *m*. The procedures of a previous study were replicated insofar as practicable, but *m* was equated across lists and the materials to be learned were varied in the intensity of rated semantic meaningfulness. Will variations in D_4 with *m* held constant be similar to the effects noted when *m* is varied?

General Method

Subjects. The *S*s were student volunteers from the introductory psychology course at the University of South Carolina. Participation in experiments partially fulfilled a course requirement. Experiment I included three groups composed from 21 females (mean age: 20.8 years) and 42 males (mean age: 19.7). Experiment II included four groups composed from 24 females (mean age: 18.9) and 56 males (mean age: 19.2). Experiment III included four groups composed from 12 females (mean age: 18.7) and 60 males (mean age: 18.6) Within each experiment there was an equal number of *S*s per group assigned at random keeping the ratio of males to females constant.

Apparatus and materials. The words used in Experiment I-III were selected from among 96 nouns and adjectives which a group of 30 undergraduates had previously rated on 10 semantic differential scales and which another 30 undergraduates used as stimulus words in a 30-sec. production test similar to Noble's (1952). The obtained D_4 values had a range 1.16-7.00, and the obtained *m* values had a range 3.80-9.20, deleting only replications of the stimulus word in tallying the responses.

The 96 words were ranked according to increasing magnitude of D_4 with Word 1 having the lowest D_4. To obtain a high-intensity set of words, words were drawn in sequence starting with Word 96 until the requisite number had been obtained. When all lists for an experiment had been selected it was usually necessary to draw additional words so as to exactly equate each list for number of 1- and 2-syllable words, replicated initial letters, and average *m*.

All materials were presented to *S*s on a Stowe Memory Drum (Model 459B) in an air-conditioned, sound-proofed, humidity controlled room.

Experiment I

Noble (1952) demonstrated that the difficulty in reaching a criterion of one perfect recitation of a 12-item serial list was a decreasing curvilinear function of *m*. He also showed that for three pre-experimental ability levels, the difficulty in attaining a criterion of 7 of 12 items correct on those same serial lists was a decreasing curvilinear function of *m*, and that in terms of total trials to criterion there was an interaction between the effects ability and *m*. Holding *m* constant, Experiment I is an attempt to replicate these findings using three serial lists of varying average D_4 dividing *S*s in each group into three ability levels based on practice list performance.

Method. With the following exceptions, the method described by Noble (1952) was replicated exactly. First, neither rest intervals nor color-naming were included in the present procedure. Second, only one practice list was used to establish the pre-experimental ability levels, and each *S* learned exactly the same list. Third, the practice and experimental list were both learned in one experimental session, separated only by a one-min. interval during which *E* engaged *S* in conversation while rearranging the memory drum shutters. Finally, 21 *S*s served in each experimental group, with 7 *S*s at each ability level based on the number of trials to attain a 7/12 criterion on a single 12-item practice list. These levels were: Slow ($>$ 12 trials), Average (8-12), and Fast ($<$ 8).

The experimental lists were obtained as follows. For the low (L) list the 12 words drawn from Words 1-17 had an average *m* = 7.00 and an average D_4 = 1.83. For the medium (M) list, the 12 words drawn from Words 41-59 had an average *m* = 7.04 and an average D_4 = 3.59. For the high (H) list the 12 words drawn from Words 77-96 had an average *m* = 7.34 and an average D_4 = 5.36. The 12 words for the practice list were drawn six each from Words 18-40 and 60-76 and had an average *m* = 7.30 and an average D_4 = 3.39. In all other respects the details of procedure and apparatus are as described by Noble (1952).

Results. The performance of the three experimental groups learning the L, M, and H lists are plotted in Figure 1 showing the number of trials to reach successive criteria.

Insert Figure 1 about here

The difficulty-meaning relationship for successive criteria of correctly anticipated items may be replotted from the data in Figure 1. This relationship was identical in all important respects to that reported by Noble (cf., Figure 2, 1952) and has not been repeated here. The average number of trials to achieve a criterion of one perfect recitation on the experimental list as a function of the three pre-experimental ability levels is reported in Fig. 2. As a check on the equality of variances for the groups having differing D_4 lists and ability levels, Hartley's test for homogeneity of variance was calculated, $\underline{F}_{\max}$ (df = 6, k = 9) = 3.61, and was nonsignificant.

Insert Figure 2 about here

Following this a treatments by levels analysis of variance was performed on the number of trials to achieve a criterion of one perfect recitation of the experimental list. The main effect of average semantic meaningfulness of the experimental list was highly significant, *F* (2,54) = 7.02, $p < .005$. The number of trials required to reach criterion in learning the L list differed significantly ($p < .005$) from that required to learn either the M or H lists which do not differ significantly from one another, according to the Duncan multiple range test.

The interaction of treatments by levels was not significant, F (4,54) < 1.00, indicating that the treatments effect was consistent across all levels of pre-experimental ability.

As a final comparison, the mean number of errors made in achieving a criterion of one perfect recitation as a function of serial position for the L, M, and H lists was calculated. Except for minor reversals at serial positions 6 and 10-12 where the M group made slightly fewer errors than the H group, there was a perfect inverse relationship between errors and average D_4 list value.

Experiment II

Cieutat, Stockwell, and Noble (1958), using variations of H and L associative meaningfulness of the stimulus (S)- and response (R)-terms, demonstrated (1) a greater difference in performance of the H-H and L-L groups than any other comparison, (2) variations in R-term meaningfulness had a significantly greater effect than similar variations in S-term meaningfulness, and (3) both S- and R-term meaningfulness as well as the interaction of these two variables in turn interacted with amount of practice. Substituting H and L semantic meaningfulness, Experiment II was an attempt to replicate these findings, holding m constant.

Method. All essential aspects of the method described by Cieutat, et al. were exactly replicated, with the following exceptions. First, the 10 pairs of words used as the practice list were one- and two-syllable nouns and adjectives drawn from Words 33-62. The mean D_4 of these words was 3.49; the mean m was 7.025.

Second, for the experimental lists 40 one- and two-syllable adjectives and nouns were selected such that for the 20 H words drawn from Words 69-96 the average D_4 was 5.13; the average m was 7.21. Similarly, for the 20 L words drawn from Words 1-30 the average D_4 was 2.08; the average m was 7.02.

All other significant details of apparatus and procedure were replicated. Thus there were four groups each learning an identical practice list for 12 trials and an L-L, H-L, L-H, or H-H experimental list for 12 trials using a correction procedure with both stimulus- and response-terms being pronounced at a standard 2:2 rate with a 4-sec. inter-trial interval.

Results. To assure having obtained four groups of equal ability prior to introduction of the independent variable, a simple randomized design analysis of variance was performed on each S's total number of correct anticipations on the practice list. The experimental groups may be considered of equal initial ability since the obtained F (11,76) = 1.01 was clearly nonsignificant.

Table 1 presents the average performance of the four groups of the experimental lists. The order of increasing difficulty in terms of the cell mean was H-H, H-L, L-H, and L-L. The average difference in performance

----
Insert Table 1 about here
----

due to variations in S-term meaningfulness was greater than that due to variations in response-term meaningfulness.

Figure 3 shows the mean number of correct responses per S as a

---

Insert Figure 3 about here

---

function of practice. A tendency toward negative acceleration is apparent in all four curves. All curves appear to be approaching the same asymptote with the L-L group generally retaining its inferior position and the H-H group generally retaining its superior position.

Because all four curves appear to be approaching the same asymptote and have for all practical purposes overlapped at Trial 6, a 2 x 2 x 5 analysis of variance was performed on the Ss' performance over the first five trials, in accord with a recommendation of Anderson (1963). The analysis yielded a significant effect of stimulus meaningfulness, F (1,76) = 9.78, $p < .001$; a significant effect of response meaningfulness, F (1,76) = 4.33, $p < .05$, and a significant practice effect, F (4,304) = 268.48, $p < .001$. The only other significant effect was the interaction of stimulus-term meaningfulness and practice, F (4,304) = 2.48, $p < .05$.

As a final check Pearson product-moment correlations (r) were calculated between the total correct responses on the practice and experimental lists. The obtained correlations (H-H: .56; L-H: .37; H-L: .61; and L-L: .59) indicated no apparent trend based on ease of learning. The average r (by Fisher's Z transformation) was .54, consistent with the value reported by Cieutat, et al.

Experiment III

Rosen, Richardson, and Saltz (1962) demonstrated that printing in red ink the fifth item of a 9-item serial list of words with the remainder printed in black resulted in that word being learned more rapidly than the same word without benefit of the red ink. This von Restorff isolation effect was obtained only if the average list m was L, not H. In the present experiment m was held constant and the average D_4 was varied across lists. It was predicted that with an L list isolating the fifth item would benefit the ease of learning that item as compared to a nonisolated control, but that isolating the fifth item in an H list would not be of any aid to Ss as compared to a nonisolated control.

Method. With the following exceptions, the method of Rosen et al. (1962) was replicated exactly. First, only 72 Ss were used. Second, to obtain the L list nine words were drawn from among Words 1-14. These words had an average m = 7.37 and an average D_4 = 1.85. The H words were drawn from Words 81-96 and these words had an average m = 7.38 and an average D_4 = 5.50.

Results. For each S Rank 1 was assigned to the list item eliciting the greatest number of correct anticipations, Rank 2 to that eliciting the second greatest number, and so forth down to Rank 9 for the least number. The mean ranks of the isolated and control items in the H and L lists are presented in Table 2. The main isolation effect was

---

Insert Table 2 about here

---

significant, $\underline{F}$ (1,68) = 4.69, p $<$ .05, indicating the mean rank of the isolated item was significantly lower than that of the same item when it was not isolated. However, the interaction of Isolation by Meaningfulness failed to achieve significance, $\underline{F}$ (1,68) = 1.06, p $<$.20, indicating that the effect of isolation on number of correct anticipations was consistent at both levels of meaningfulness.

The mean number of correct anticipations to Item #5 for each group is also presented in Table 2. The same general conclusions can be reached for this data, viz., a difference in mean numoer correct anticipations of isolated as opposed to control items, but no interaction between isolation and list meaningfulness. However, as noted by Rosen et al. (1962) these results are less tenable due to the potentially more rapid learning of the H list.

A two-factor analysis of variance was performed on the total number of correct anticipations made by each $\underline{S}$. As might be expected (cf., Wallace, 1965), the isolation vs. control comparison failed to achieve significance based on this measure, $\underline{F}$ (1,68) $<$ 1.00. However, the H vs. L comparison was significant, $\underline{F}$ (1,68) = 4.27, p $<$.05, indicating a greater number of correct anticipations were elicited by the H list. The interaction effect was nonsignificant, $\underline{F}$ (1,68) $<$ 1.00, indicating the consistency of the H vs. L difference across both isolated and control lists.

Discussion

Variations in average list D_4 produce very apparent differences in the ease of learning such lists in a variety of experimental situations holding $\underline{m}$ constant. In all experiments, high D_4 lists were learned in fewer trials and with fewer errors than lower D_4 lists. In each experiment there were, however, some interesting deviations from results previously demonstrated using $\underline{m}$ as the primary independent variable.

Experiment I. The original study by Noble (1952) was not replicated in twc respects. First, although the overall effect of D_4 variation was significant and the trend of more trials required to learn lower D_4 lists was consistent the M and H lists did not differ significantly in the total average number of trials to reach criterion. It should be noted, however, that the average $\underline{m}$ of the three lists was 7.13 as measured using Noble's (1952) technique with only a 30-sec. interval per word. In terms of the average number of associations it would seem safe to assume that the current words used are of approximately similar meaningfulness to the most highly meaningful words used by Noble (e.g., KITCHEN, $\underline{m}$ = 9.61 using a 1 min. production interval with more extensive editing of "acceptable" responses). Thus the demonstrated effects of D_4 variation were over and above any effects attributable to a substantial degree of associative meaningfulness.

Second, the heterogeneity of variance of the data composing the treatment by levels cell means noted by Noble was not obtained. This presumably resulted from the relatively similar number of trials required by each group to reach criterion. Related to this, the interaction of levels and D_4 also failed to achieve significance in that the number of trials to criterion as a function of D_4 was consistent across all ability levels. The hypothesis advanced by Noble that "slow learners are more sensitive to differences in meaningfulness than are relatively faster learners" would seem to be limited to differences in associative meaningfulness since the interaction of D_4 and levels falls far short of significance.

Noble's general conclusion that difficulty in serial learning is a decreasing exponential function of list *m*-value was also true of list D_4-value. Likewise, there was a relatively perfect inverse relationship between the mean number of errors and list D_4. In sum, the general effectiveness of variations in list D_4 in producing variations in trials to criterion has been well demonstrated in serial learning. In those instances where the results fail to replicate Noble the cause can generally be traced to the high average *m* of the words in the present study.

Experiment II. The finding of Cieutat et al. (1958) that variations in response meaningfulness yielded greater differences in performance than similar variations in stimulus meaningfulness was not confirmed. In fact, there was a tendency for the reverse to be true. As can be noted in Figure 3, with one minor exception the groups can be ordered in terms of their performance with HH $\triangleright$ HL $\triangleright$ LH $\triangleright$ LL. These results were consistent with Paivio's (1968) observation that the superior effect of response *m* has been obtained only with nonsense words as low-*m* items, and that *m* may be more potent on the stimulus side when varied entirely among familiar words. The present results extend this hypothesis to variations of D_4 among familiar words. Although the present words were originally selected without respect to imagery (cf., Kasschau, 1970; Paivio, 1968), after the fact analysis of those words selected which are rated for imagery by Spreen and Schulz (1966) or Paivio, Yuille, and Madigan (1968) indicated both low and high D_4 word lists have average imagery ratings of 5.2-5.7 on a 7-point scale where 7 indicates maximum concreteness. In other words, imagery apparently did not vary significantly in the lists of Experiment II.

Experiment III. The primary difference between the results of the present study and those obtained by Rosen, *et al*., (1962) was that the overall effect of isolation was significant, specifically that the isolation effect was significant at both levels of D_4, rather than only with low D_4 materials. As a result the expected interaction between isolation vs. control lists and H vs. L D_4 lists was not obtained.

Winograd's (1966) work, which demonstrated the apparently greater distinctiveness of H D_4 items, had led to the expectation that the interaction would be significant. Perhaps the explanation for the obtained results was contained in a suggestion made by Kasschau and Pollio (1967) in commenting on the initial superiority of associative relations in mediating response transfer as compared with semantic relations. The suggestion was that *S*s have at least two different types of word relations available, associative and semantic. Where only the semantic similarity type of relation is available, words do not as rapidly gain any benefit from this similarity, relative to associative relations; $\underline{S}$s require some time to avail themselves of these relationships. If this was so, the consistency of the significance of the isolation effect might have resulted from the lack of differentiation of the words composing the high D_4 list especially during the initial learning trials.

In conclusion, it can be said that holding associative meaningfulness constant it has been demonstrated that variations in semantic meaningfulness produce changes in performance consistent with changes produced by similar variations in associative meaningfulness. The D_4 measure does, in fact, appear to measure an aspect of meaningfulness not detected by $\underline{m}$, an aspect which influences performance even at relatively high levels of $\underline{m}$.

References

Anderson, N. H. Comparison of different populations: Resistance to extinction and transfer. *Psychological Review*, 1963, *70*, 162-179.

Bousfield, W. A. The problem of meaning in verbal learning. In C. N. Cofer (Ed.), *Verbal Learning and Verbal Behavior*. New York: McGraw-Hill, 1961, pp. 81-91.

Cieutat, V. J., Stockwell, F. E., & Noble, C. E. The interaction of ability and amount of practice with stimulus and response meaningfulness (*m*, *m'*) in paired-associate learning. *Journal of Experimental Psychology*, 1958, *56*, 193-202.

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

Howe, E. S. Uncertainty and other associative correlates of Osgood's D4. *Journal of Verbal Learning and Verbal Behavior*, 1965, *4*, 498-509.

Kasschau, R. A. Unipolar vs. bipolar semantic differential rating scales. University of South Carolina, unpublished manuscript, 1970.

Kasschau, R. A., & Pollio, H. R. Response transfer mediated by meaningfully similar and associated stimuli using a separate-lists design. *Journal of Experimental Psychology*, 1967, *74*, 146-148.

Koen, F. Polarization, *m*, and emotionality in words. *Journal of Verbal Learning and Verbal Behavior*, 1962, *1*, 183-187.

Noble, C. E. The role of stimulus meaning (*m*) in serial verbal learning. *Journal of Experimental Psychology*, 1952, *43*, 437-446; *44*, 465.

Osgood, C. E. Comments on Professor Bousfield's paper. In C. N. Cofer (Ed.), *Verbal learning and verbal behavior*. New York: McGraw-Hill, 1961, pp. 91-106.

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

Paivio, A. A factor-analytic study of word attributes and verbal learning. *Journal of Verbal Learning and Verbal Behavior*, 1968, *7*, 41-49.

Paivio, A., Yuille, J. C., & Madigan, S. Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of Experimental Psychology*, 1968, *76* (1, Pt. 2).

Rosen, H., Richardson, D. H., & Saltz, E. Supplementary report: Meaningfulness as a differentiation variable in the von Restorff effect. *Journal of Experimental Psychology*, 1962, *64*, 327-328.

Spreen, O., & Schulz, R. W. Parameters of abstraction, meaningfulness and pronunciability for 329 nouns. *Journal of Learning and Verbal Behavior*, 1966, *5*, 459-468.

Wallace, W. P. Review of the historical, empirical, and theoretical status of the von Restorff phenomenon. *Psychological Bulletin*, 1965, *63*, 410-424.

Winograd, E. Recognition memory and recall as a function of degree of polarization on the semantic differential. *Journal of Verbal Learning and Verbal Behavior*, 1966, *5*, 566-571.

Zippel, B. Semantic differential measures of meaningfulness and agreement of meaning. *Journal of Verbal Learning and Verbal Behavior*, 1967, *6*, 222-225.

Table 1

Mean Total Correct Responses
During Trials 2-6 in Experiment II

D4 value of S term	D4 value of R term		Mean	S-Difference
	L (2.08)	H (5.13)		
L (2.08)	27.05	30.85	29.95	4.53
H (5.13)	32.70	36.25	34.48	
Mean	29.88	33.55		
R-difference	3.67			

Table 2

Mean rank within list and mean correct anticipations of the isolated and control terms for 15 learning trials

List		Ranks		Correct Anticipations	
		Mean	Standard Deviation	Mean	Standard Deviation
Low D4	Isolated	5.78	2.40	10.11	3.25
(1.85)	Control	6.36	2.25	9.50	3.96
High D4	Isolated	5.61	1.95	11.06	2.76
(5.50)	Control	7.25	1.82	9.67	2.92

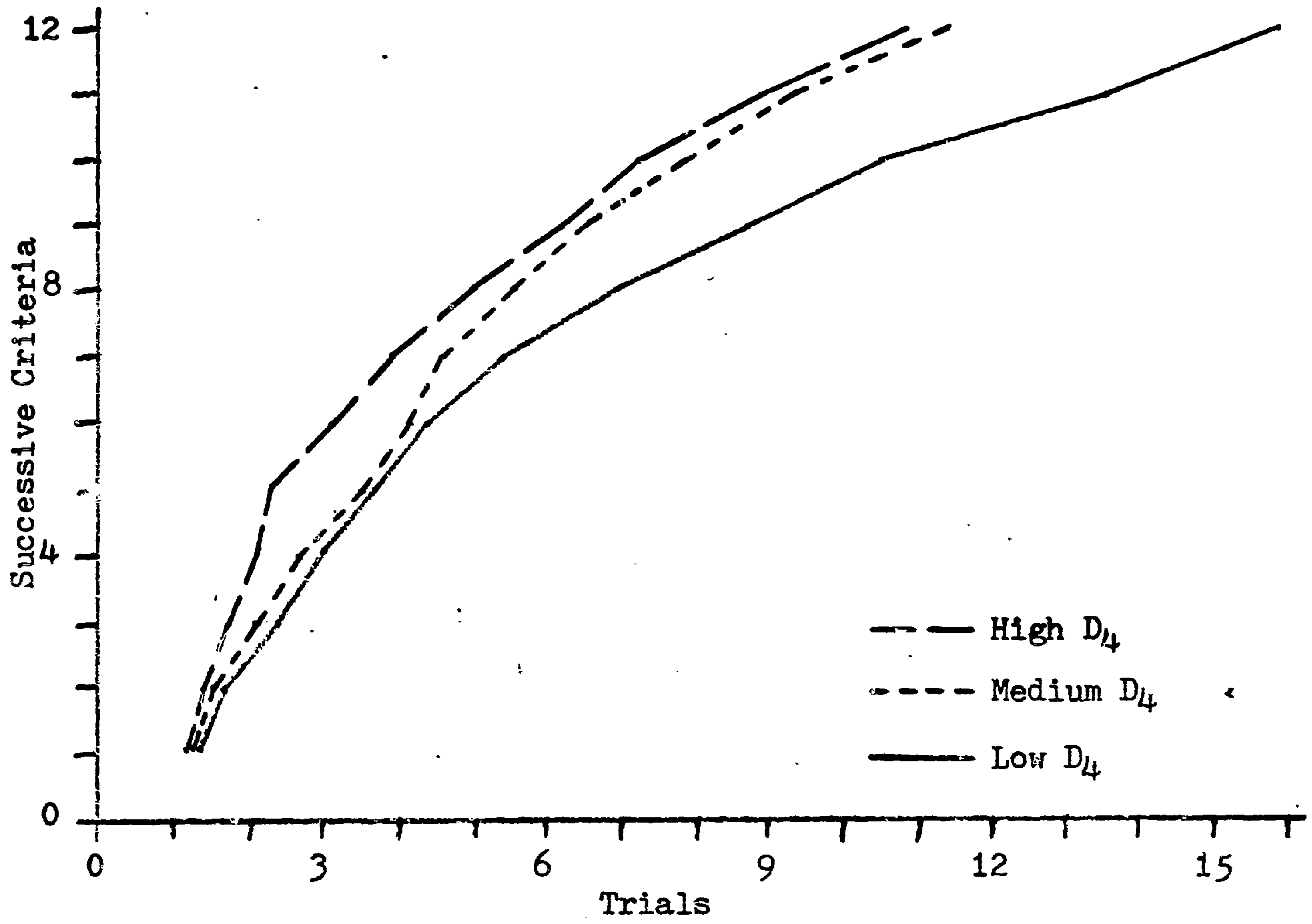

Fig. 1. Mean number of trials to successive criteria in Experiment I.

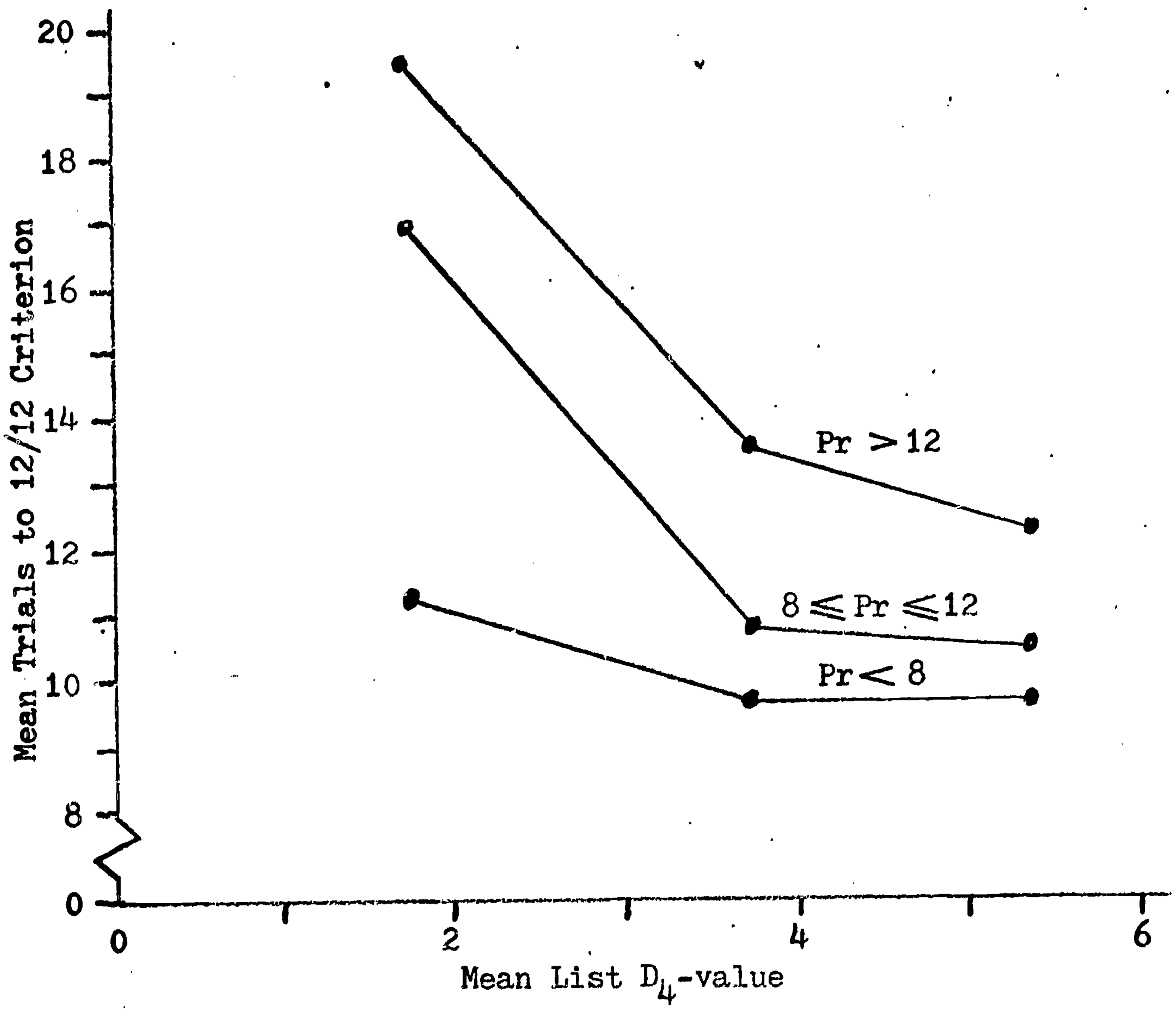

Fig. 2. The difficulty-meaning relationship as a function of number of trials to 7/12 criterion on a practice list in Experiment I.

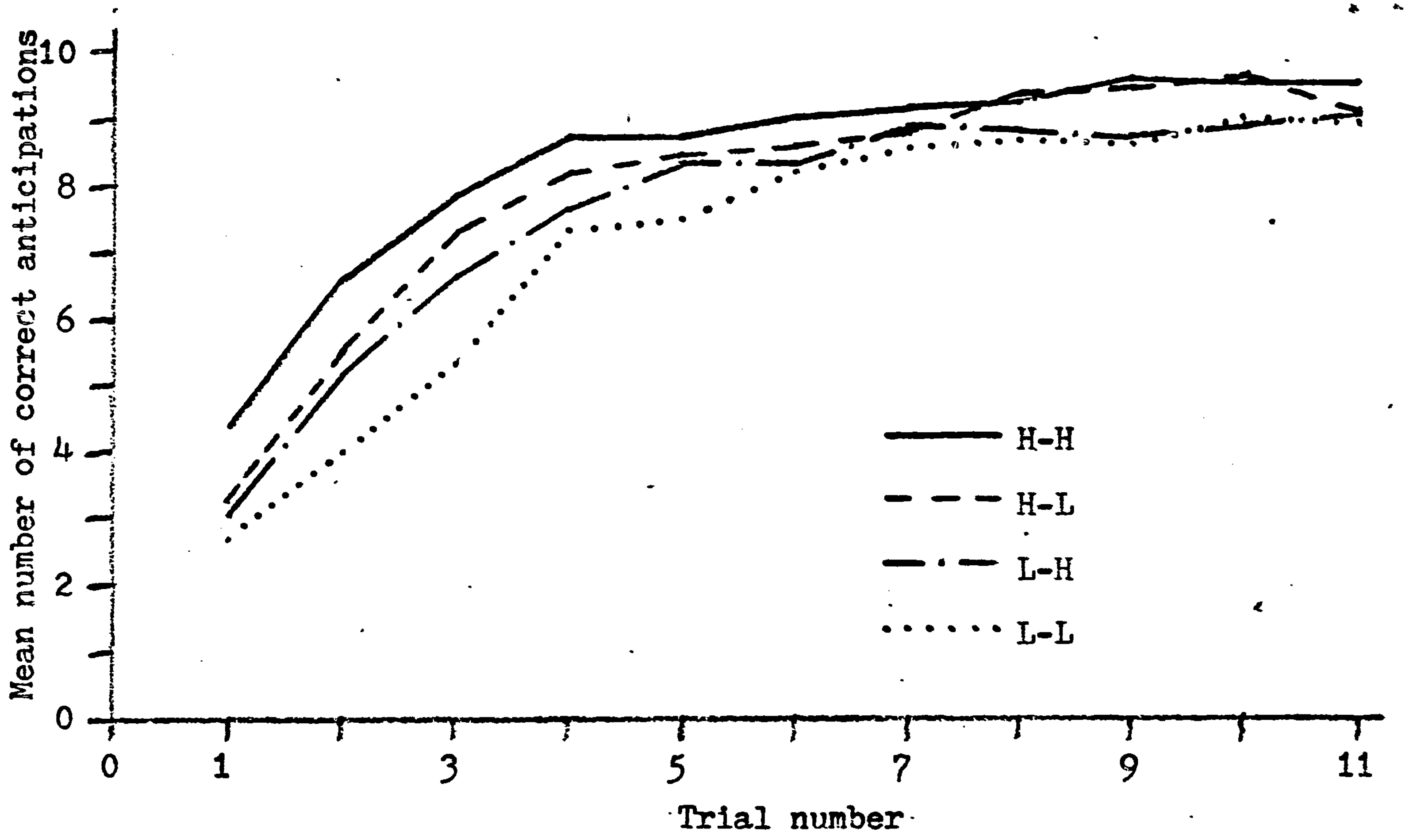

Fig. 3. Mean number of correct anticipations as a function of the number of trials of practice in Experiment II.