THE ROLE OF MEANINGFULNESS (m) IN PAIRED-ASSOCIATE VERBAL LEARNING

CLYDE E. NOBLE AND DELDON A. McNEELY

Louisiana State University

A series of investigations (4, 5, 6, 7, 8) has been concerned with identifying and quantifying some of the variables which determine performance in verbal learning. In the first of these studies, Noble (4) developed a rational scale of meaningfulness (m) consisting of 96 disyllabic nouns and paralongs ordered in terms of frequency of continued written associations in 60 sec. The m scale meets the need for an extensive range of verbal stimuli, from paralongs to frequently used words. It fulfills the requirement of high stability as well, having an estimated intergroup reliability coefficient of .993 (n = 119). Recently, Mandler (3) has applied this technique to 100 English syllables (n = 34), while Umemoto, Morikawa, and Ibuki (10) have done the same for 1,892 Japanese syllables (n ~ 100).

In his second investigation, Noble (5) found that variations in m produced significant effects upon serial verbal performance. Difficulty, measured by trials to mastery, was a decreasing curvilinear function of mean list m value, and there was an interaction between reactivity to m and Ss' initial ability. The relationship appeared to be negatively accelerated, but an exact test could not be performed with the serial task since only three satisfactory 12-item lists may be constructed from the sample provided by the m scale.

The purpose of the present experiment is to extend the analysis of meaning to the learning of paired associates. By controlling individual differences and eliminating the complicating effects of the ordinal positions of items, this method can evaluate more clearly than the serial method the nature of the difficulty-meaning relationship for specific S-R connections (5). The paired-associate task also permits the sampling of more empirical points than were available in earlier experimentation. Specifically, it is predicted (a) that the acquisition rate of single verbal habits will be a positive function of m, and (b) that an interaction will occur between meaningfulness and ability to learn.

METHOD

Apparatus.—The apparatus consisted of a Gerbrands memory drum set to expose individual items at a 3-sec. rate. An idler attachment permitted five sets of 10 pairs to be presented.

1 This paper is based in part upon a 1955 master's thesis investigation by the junior author under the guidance of the senior author. The latter reviewed the findings during a Symposium on "The Experimental Analysis of Verbal Behavior," Midwestern Psychological Association, May 1956. We are grateful to Drs. E. A. Bilodeau, G. A. Kimble, G. Mandler, and R. Thompson for helpful criticisms.

2 Since the reported (4) average unit reliability of .975 was based on the Z-transformed intercorrelations of four groups of 30 Ss each, the combined scale reliability cited above represents a necessary correction for quadrupled sample size. The logic of this calculation as applied to response-defined variables is presented elsewhere (9).

3 For kindly calling our attention to this article, we are indebted to Dr. Koji Sato of Kyoto University. The translation was performed by Mr. K. S. Kim.

4 Two important studies on this topic (1, 10), which appeared after our work was completed, will be discussed later.
First, a practice list was constructed in which 10 pairs of three-syllable adjectives were arranged in five different random sequences. Each sequence constituted a trial, and the series repeated on Trial 6. The experimental lists were constructed from items in Noble's (4) m scale. Eighteen bands were prepared, each band having a list of 10 pairs of two-syllable nouns or paralogs. Like the practice list, these were also presented in five different random trial sequences giving effectively 90 sequences. The available tables of random numbers did not meet certain logical specifications, so the following statistical restrictions were imposed: no run was repeated on adjacent trials; no run was repeated more than once in five trials; there were no triple runs; and no S-R pair occurred in the same position on adjacent trials. Frequencies of occurrence at each serial position were approximately equalized for the 10 m values, and there were no homonyms, synonyms, antonyms, assonants, or alliterates within a given list. The purpose of these restrictions was to minimize cues which would facilitate the learning of some pairs and not of others.

Individual items ranged in median m value from .25 to 9.13. Medians were employed rather than means to provide greater accuracy at the low end of the scale, as discussed in (6). The median m value of each S-R pair was computed by averaging the median m values of the two items composing it. Pooling all 18 experimental lists, the paired associates represented m values ranging from .29 to 8.54 in 10 approximately equal-interval steps. The over-all mean was 4.29. There were from 3 to 7 items for each pair, with a total of 48 different items used altogether. No item appeared in more than 6 lists in the same S or R position, and none occurred more than 12 times. These manipulations were designed to stabilize performance scores at each of the 10 points by averaging out any intrinsic differential effects of particular items. All lists were typed in capital letters on white paper tape.

In order to test for the stability of the experimental variable, a 5-point graphic rating schedule of association value (a') was administered individually to an independent group of 10 undergraduates. Noble's (4) entire list of words and paralogs was used, and Ss' instructions were to judge the meaningfulness of each item on the basis of estimated number of associations. By a technique described previously (5, 6), 96 median ratings were derived. This scale had a high test-retest reliability (r = .985), and the correlation between test (a') and criterion (m) was .871. The evidence thus indicated that the m scale, although standardized on a military population, could be used quite rigorously with the population available in this investigation.

Procedure.—The Ss were 62 men and 28 women students at Louisiana State University, enlisted on a voluntary basis and uninformed of the purpose of the experiment. Their ages ranged from 18 to 27 (mean = 21 yr.). All were naive with respect to paired-associate learning. Each S first learned the practice list to a criterion of 8 out of 10 correct responses, which was considered adequate to control the learning-to-learn factor in the present situation (11, p. 306). This was given during the same session as the experimental list and served to classify Ss into three pre-experimental ability levels (5). These levels, defined by the total trials required to reach the 8/10 criterion, were as follows: Slow > 17 (n = 28), Average 12–17 (n = 32), and Fast < 12 (n = 30).

The interval between onsets of successive stimuli was 6 sec.: 3 sec. for S alone and 3 sec. for S-R together. The intertrial interval was 9 sec. A 2-min. rest period was given between the completion of the practice list and the presentation of the experimental list. Five Ss were then assigned without bias to each of the 18 lists in a counterbalanced order of trial sequences. All 90 Ss were required to practice for 20 trials. On the first presentation of the list (Trial 1), S merely read the paired words aloud and standard pronunciations were established. Attempted anticipations of the correct responses began on the second presentation (Trial 2). The correction procedure was used.

For the analyses of variance a Type I design (2) was employed, in which the interaction of list and meaningfulness effects was confounded with individual differences. In this type of mixed factorial design inter-S comparisons are usually less precise than intra-S comparisons, hence individual differences were controlled in the case of m value and precision was sacrificed only in the ability and list evaluations.

Results

Figure 1 shows the percentage of correct responses as a function of practice with meaningfulness (m) as the parameter. The 10 original curves have been pooled by adjacent pairs for greater clarity of presentation. Each of the five categories in Fig. 1 thus denotes the mean median m value of two groups of 18 paired associates (n = 90). The functions are quite regular with no inversions until Trial 11. Consistent with previous data (5) from serial
learning, rate of acquisition is a positive function of $m$. Inspection of the forms of the individual curves reveals a trend toward positive acceleration with decreasing $m$ value.

A $5 \times 18$ matrix of the average percentage of correct responses during Trials 2–20 was prepared. The A classification represented the same meaning categories as in Fig. 1, while the B classification represented the various lists. Each cell mean was based upon the performance of 5 Ss. The analysis of variance is presented in Table 1. Meaningfulness has a highly significant effect on performance, as indicated by the $F$ ratio of 52.72 for $df = 4/288$. There are no significant differences among lists, which include individual differences. An interaction of marginal significance is present, but it is of secondary interest since the list classification was introduced primarily to formulate a more generalized estimate of the effects of $m$ value. We may attribute the observed A $\times$ B interaction to experimental error associated with contextual distortion within certain lists.

The difficulty-meaning relationship is shown in Fig. 2 where the mean total errors in 20 trials are plotted against the 10 average $m$ values. A simple test of the observed trend is provided by computing a $t$ ratio between the first and last points (2, p. 342). The ratio is 8.49, which for 89 $df$ is highly significant ($P < .001$) of a tendency for errors to decrease with increasing meaningfulness. Another way of expressing the monotonic relationship in Fig. 2 is by means of rank-order correlation; the value of rho is $- .988$.

The correlation between the number of trials to 8/10 correct responses on the practice list and the same criterion measure on the experimental lists is .50, so we may expect a clear separation of the Slow, Average, and Fast groups for the foregoing analysis. Table 2 presents the total average errors in the various meaning categories for each ability level. It will be noted that the over-all means are

---

**TABLE 1**

**Analysis of Variance of Percentage of Correct Responses in 20 Trials**

<table>
<thead>
<tr>
<th>Source</th>
<th>$df$</th>
<th>$MS$</th>
<th>$F$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Ss</td>
<td>89</td>
<td></td>
<td>1.20</td>
</tr>
<tr>
<td>Lists (B)</td>
<td>17</td>
<td>311.88</td>
<td></td>
</tr>
<tr>
<td>Error ($b$)</td>
<td>72</td>
<td>260.34</td>
<td></td>
</tr>
<tr>
<td>Within Ss</td>
<td>360</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Meaningfulness (A)</td>
<td>4</td>
<td>3046.72</td>
<td>52.72**</td>
</tr>
<tr>
<td>Interaction (A $\times$ B)</td>
<td>68</td>
<td>82.91</td>
<td>1.45*</td>
</tr>
<tr>
<td>Error ($\omega$)</td>
<td>288</td>
<td>57.79</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>449</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* $P = .05$.
** $P < .001$. 

---

Fig. 1. Acquisition curves for single S-R pairs as a function of meaningfulness ($m$).

Fig. 2. The difficulty-meaning relationship for 10 groups of paired associates.
TABLE 2  
MEAN NUMBER OF ERRORS IN 10 MEANING CATEGORIES FOR 3 PRE-EXPERIMENTAL ABILITY LEVELS DURING TRIALS 2-20

<table>
<thead>
<tr>
<th>Levels</th>
<th>n</th>
<th>.29</th>
<th>.79</th>
<th>1.74</th>
<th>2.98</th>
<th>3.92</th>
<th>5.88</th>
<th>5.65</th>
<th>6.51</th>
<th>7.40</th>
<th>8.54</th>
</tr>
</thead>
<tbody>
<tr>
<td>Slow</td>
<td>28</td>
<td>4.64</td>
<td>4.71</td>
<td>4.57</td>
<td>4.79</td>
<td>2.89</td>
<td>2.86</td>
<td>3.00</td>
<td>1.68</td>
<td>1.46</td>
<td>0.86</td>
</tr>
<tr>
<td>Average</td>
<td>32</td>
<td>4.28</td>
<td>3.38</td>
<td>2.53</td>
<td>3.56</td>
<td>2.66</td>
<td>2.09</td>
<td>1.38</td>
<td>1.31</td>
<td>1.06</td>
<td>1.03</td>
</tr>
<tr>
<td>Fast</td>
<td>30</td>
<td>1.73</td>
<td>2.47</td>
<td>1.67</td>
<td>1.67</td>
<td>1.07</td>
<td>1.57</td>
<td>0.73</td>
<td>0.57</td>
<td>0.33</td>
<td>0.47</td>
</tr>
</tbody>
</table>

Mean

<table>
<thead>
<tr>
<th>Value</th>
<th>.29</th>
<th>.79</th>
<th>1.74</th>
<th>2.98</th>
<th>3.92</th>
<th>5.88</th>
<th>5.65</th>
<th>6.51</th>
<th>7.40</th>
<th>8.54</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>3.55</td>
<td>3.52</td>
<td>2.92</td>
<td>3.34</td>
<td>2.21</td>
<td>2.17</td>
<td>1.70</td>
<td>1.19</td>
<td>0.95</td>
<td>0.79</td>
</tr>
</tbody>
</table>

the same as the difficulty values graphed in Fig. 2. Standard errors of the means are included to show the inverse relationship between variability and meaningfulness. This rank-order correlation is \( r = -.915 \).

According to Table 1 there are no differences among lists, so the ability classification may be regarded as of primary interest. Pooling Ss regardless of list, a \( 10 \times 3 \) matrix of error scores was prepared. The analysis of variance given in Table 3 again indicates the facilitative effect of \( m \) on performance as well as differences due to initial ability. The significant interaction variance confirms Noble's (5) findings in the serial learning situation and supports his hypothesis that slow learners are more sensitive to differences in meaningfulness than are relatively fast learners. A Bartlett test indicated unequal variances among the three ability levels; for \( df = 2, \chi^2 = 15.4 \) which is significant \( (P < .01) \). Judging from an investigation by Norton (2, p. 78), however, such heterogeneity does not necessarily invalidate the \( F \) test, particularly since the main effects are still significant under a doubled level of confidence. The Bartlett test is useful here because it verifies the inverse relationship between ability and variability, demonstrating once again that sources of difficulty (e.g., low \( m \) value) affect the variances as well as the means in verbal learning.

In Fig. 3 appears an analysis of the difficulty-meaning relationship as a function of stage of practice. As was observed in Fig. 1, the acquisition

TABLE 3  
ANALYSIS OF VARIANCE OF TOTAL ERRORS IN 20 TRIALS

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>MS</th>
<th>( F )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Ss</td>
<td>89</td>
<td>270.02</td>
<td>19.48**</td>
</tr>
<tr>
<td>Ability (B)</td>
<td>2</td>
<td>13.86</td>
<td></td>
</tr>
<tr>
<td>Error (b)</td>
<td>87</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within Ss</td>
<td>810</td>
<td>100.59</td>
<td>25.66**</td>
</tr>
<tr>
<td>Meaningfulness (A)</td>
<td>9</td>
<td>8.90</td>
<td>2.27*</td>
</tr>
<tr>
<td>Interaction (A ( \times ) B)</td>
<td>18</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Error (( e ))</td>
<td>783</td>
<td>3.92</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>899</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\* \( P < .01 \).  
** \( P < .001 \).
curves begin to level off around Trial 11 so the 3-trial parameter sequence is broken at that point. The errors for Trials 18–20 are plotted in order to describe the final learning period. From this graph it is evident that the phenomenon shown in Fig. 2 is primarily an initial effect which is quickly dissipated by the cumulative effects of practice.

Discussion

Two recently published experiments (1, 10) are closely related to the present findings. In conjunction with their application of the association-frequency method (4) to the Japanese language, Umemoto, Morikawa, and Ibuki (10) carried out a paired-associate investigation using three levels of meaningfulness ($n = 9$). Each list consisted of seven S-R pairs, with English letters as stimuli and scaled syllables as responses. Four different sequences were used and the items were presented at a 2-sec. rate. Consistent with the data of Noble (5) and with those reported here, difficulty varied inversely with $m$ value: the mean trials to criterion for the low, medium, and high lists were 20.9, 13.9, and 7.5, respectively.

The other study, designed partly to test the same hypothesis, is by Kimble and Dufort (1). Employing 10-pair lists drawn from the $m$ scale, they too obtained a generally negative relationship between trials to criterion scores and meaningfulness. In the first replication (Exp. IA), however, there was a flattening of the curve at the low end of the scale. Kimble and Dufort were led to state that “the paired-associate procedure introduces a variable which makes the less meaningful materials easy to learn” (1, p. 362). Despite the fact that this inversion did not occur in the second replication (Exp. IB), in which there were marked fluctuations at both ends, they proposed an explanation by analogy to the von Restorff “isolation” effect. They proceeded to ingenious demonstrations of other phenomena, such as heterogeneity in serial position curves (Exp. III), but unfortunately did not return to the central issue concerning us here. Kimble and Dufort seemed to be convinced that the difficulty-meaning relationship is nonmonotonic; i.e., it cannot be determined “in an uncomplicated way” (1, p. 361) by the paired-associate method.

The writers disagree on this point. Distortions resulting from a few different items appearing in diverse contexts can seriously obscure one’s estimate of the true function. In executing the present experiment, therefore, a variety of S-R pairs were employed in a mixed factorial design. In addition, rather extensive statistical precautions regarding frequencies and runs were taken in order to preclude other kinds of artifact. There was also an attempt to reduce variability due to learning-to-learn, type of score, and sample size. Whereas Kimble and Dufort used 40 $S$s with no pretraining, 2 lists, 12 different sequences, and a trials-to-mastery design, this experiment used 90 $S$s with pretraining, 18 lists, 90 different sequences, and a constant-trials design. The two studies are very similar otherwise, even to the extent of identical apparatus and presentation rates. Under the present conditions there is no reason to conclude that paired associates per se introduces an “isolation” factor. While it is true that a list $\times$ meaning interaction significant at about the 5% level was obtained, the effect is confounded with individual differences; indeed, this may help to clarify the Kimble-Dufort observations by identifying the mode of operation of the complicating factor.

More to the point are the relative magnitudes of the meaning and interaction variances. Given the precision indicated by Tables 1 and 2, the practical importance of the “isolation” effect is small when compared with the main effect of meaningfulness (2, p. 214). Furthermore, since no significant differences were found among lists, the generalization of this interaction to any hypothetical population of $S$s would
have to be so limited by qualifications as to be of small theoretical value. Accordingly it is interpreted as experimental error. In other words, the phenomenon Kimble and Dufort observed may be an artifact and not representative of the function one would obtain by infinite replications where items, positions, and sequences are of no interest.

It is notable that response variability in these studies is rather large. This may be due as much to the material as to the method. Perhaps if more uniform stimuli (e.g., three-letter syllables vs. two-syllable words) were employed, thus requiring more uniform responding (e.g., spelling vs. pronouncing), the resulting functions would show greater stability and freedom from disturbing artifacts. Kimble and Dufort's own evidence (1, p. 363), collected during preliminary investigation of the comparative effects of S and R meaningfulness (Exp. II), is relevant here. Before this hypothesis can be tested, however, new measurements of association value will be required. Lists such as those compiled by Glaze and Krueger have been very useful to students of verbal learning, but are now outdated. Mandler's (3) scale, which is similar to the $m$ scale operationally, might be suitable except that the range of syllable associations ($2.9$ to $5.3; M = 4.1, SD = .48$) is more restricted than that of paralogs and words ($0.99$ to $9.61; M = 3.66, SD = 2.38$).

The present results suggest that the difficulty of single verbal habits is a decreasing monotonic (sigmoid?) function of $m$. The theoretical explanation for the facilitative effect, however, is not clear. There have been a number of attempts to account for the meaning-learning correlation; various writers have alluded to such factors as "familiarity" (Ebbinghaus), "multiple associations" (James), "organization" (Koffka), "predifferentiation" (Gibson), and "transfer" (McGeoch). Actually, these terms are little more than descriptions of some of the related phenomena. A more strategic attack upon the problem would be to determine the relative contributions of the S and R terms to the variance in performance, as outlined by Noble (5, p. 443). Several lines of evidence reviewed elsewhere (5, 8) indicate that differential facilitation is primarily connected with responding (articulation) and that stimulus factors are of secondary importance (cf. also 10). Consideration of those studies together with the current discussion permits two final predictions: (a) the simple effect favoring high S-high R meaningfulness (e.g., $m = 6-9$) over low S-low R meaningfulness (e.g., $m = 0-2$) will be greater than for any other comparison, and (b) the difference in the main effect of $m$ on the response side will be greater than the comparable difference on the stimulus side. Given (a) and (b), the likelihood of intrinsic S-R interaction with respect to $m$ is slight.

Both these hypotheses are clearly at variance with the Kimble-Dufort position, yet there is one point of agreement. By assuming that "isolation" is a perceptual process, they reason (1, p. 363) that the learning of low S-high R pairs should be enhanced relative to that of high S-low R pairs. It is interesting that an S-R motor-patterning viewpoint, although contradictory otherwise, generates the same implication. It is also consistent with the methodological consideration that $m$ values are established by recall rather than recognition operations. Thus, using Fig. 2 as a guide, the expected ranking of the four basic conditions in order of decreasing difficulty is: 1. low S-low R, 2. high S-low R, 3. low S-high R, 4. high S-high R. By comparison, the configurational ground plan is at least incomplete for it neglects to establish continuity with the limiting cases, where the $m$ values of S and R covary. Due to the absence of a low S-low R condition in the Kimble-Dufort design (Exp. II), and also because of the presence of confounded (and possibly heterogeneous) interaction, their data do not constitute an appropriate test of the hypotheses presented above.
A simple $2 \times 2$ factorial design employing four independent groups of $S$s would provide unequivocal answers to these questions.

**Summary**

Certain relationships between performance in verbal learning and the attribute of meaningfulness ($m$) were investigated. Ninety college students practiced for 20 trials with 18 lists of 10 paired associates representing approximately equal intervals on the $m$ scale. As expected from serial learning research, rate of acquisition was a positive function of $m$, and the curves showing the percentage of correct responses tended toward positive acceleration with decreasing $m$ value. Difficulty measured by mean total errors was an inverse monotonic function of $m$, the rank-order correlation being $-0.988$. Variability also decreased with increasing meaningfulness.

There were significant differences in performance correlated with pre-experimental ability, as well as a negative relationship between ability and variability. Consistent with a second prediction, reactivity to $m$ interacted with ability to learn. A level of practice analysis revealed that the difficulty-meaning relationship was an initial effect which disappeared by Trial 11. Contrary to recent opinion, these findings indicate that the law relating difficulty and meaningfulness for specific S-R connections can be determined by the method of paired associates. Suggestions were offered for further research to specify more clearly the role of "perceptual" and "motor" factors in verbal learning.

**References**


(Received June 3, 1955; Revised July 5, 1956)