The test-retest reliability of free association norms¹

S. JOYCE BROTSKY, SAN FERNANDO VALLEY STATE COLLEGE
MARIGOLD L. LINTON, SAN DIEGO STATE COLLEGE

146 Ss gave free associations to 158 stimulus words on two occasions separated by 10 weeks. The probability of responding with an identical associate (IA) after 10 weeks was .32. A correlation of .84 was found between number of Primary (P) responses and number of IAs for the 146 Ss. A high positive correlation was found between dominance of the Primary (DP) and number of IAs for the 158 words, while a negative correlation was found between frequency in the language of the stimulus word (LF) and number of IAs. The coefficient of multiple correlation yielded $R_{1,23} = .92$, indicating that a high degree of reliability can be predicted for words with low LF and high DP.

The frequency with which a response is given to a verbal stimulus on a free association test has usually been taken as an index of the strength of the association between the stimulus and response. The most widely used index of associative strength is DP, the number of times the most frequent response occurs to a given stimulus word in a normative sample. Although DP is a simple and objective measure, it has not provided the means for making precise predictions in more than a few limited types of verbal learning situations (Marshall & Cofer, 1963).

When the validity of a test instrument is low, it is customary to investigate the reliability of that instrument. Although the Test-Retest reliability (TRR) of free association norms would seem to be a basic question, little research has been done in this area. In a recent study by Hall (1966), six-letter adjectives were presented individually to Ss, and after seven or 21 days, Ss gave IAs with a probability of approximately .50. Although Hall concluded that the value of .50 "presumably provides sufficient stability to enable investigators to utilize word association responses in their experimental analysis of learning, transfer, and retention" (1966, p. 80), this conclusion remains open to question.

In view of the fact that normative data are frequently used as the basis for predictions of verbal behavior, it seemed important to determine the TRR of the Kent-Rosanoff (K-R) list (Russell & Jenkins, 1954) and other polysyllabic words using the standard procedures employed by such investigators as Russell & Jenkins (1954); Bousfield, Cohen, Whitmarsh, & Kincaid (1961); and Palermo & Jenkins (1964) in the collection of their norms.

Method

The Ss were 198 undergraduates between the ages of 17 and 28. All Ss had been born in continental United States and had learned English as their first language.

The Ss participated as part of a course requirement at San Fernando Valley State College. The data for the present study were taken from 146 Ss, since 52 Ss were excluded for failure to take both Test and Retest or failure to meet the age or language requirements.

The stimulus booklet consisted of 158 words, 25 per page, each followed by a line on which S was instructed to write the first word that came to mind. The first 100 words on the Test were a randomization of the K-R list, while the last 58 words were chosen from the Connecticut Free Associational Norms (Bousfield et al, 1961). The Retest consisted of a second randomization of the same words. All Ss received the words in the same serial order.

The instructions and procedures were essentially those used by Russell & Jenkins (1954) in the collection of the Minnesota norms. The Test and Retest took place in a large lecture hall during class time. The E wrote the time on the blackboard at 1 min intervals, and Ss were instructed to write the time of completion at the bottom of their test booklets. No specific mention was made during the Retest of the fact that the class had participated in a similar experiment on an earlier occasion.

Results and Discussion

A set of norms was developed for the 146 responses to the first presentation of the 158 stimulus words. On the basis of these norms, the P response to each word was found and DP was determined. DP ranged from 12 for MARK-WRITE to 120 for WEB-SPIDER. The number of P responses given by any S ranged from 14 to 94 with a mean of 55, a median of 57, and a standard deviation of 14.99.

The TRR of the associations was found by determining if the same response was given on the Test and Retest. For the purposes of scoring, a response that was plural in form was considered identical to the singular. After checking each Retest word as identical (IA) or different (D), tallies were made for each S and for each word. The number of IAs ranged from 16 for the stimulus word SCENE to 127 for the stimulus word SPOOL. The number of IAs across Ss ranged from 8 to 99 with a mean of 51, a median of 52, and a standard deviation of 15.20. The mean proportion of IAs was .33 for the 100 K-R words and .31 for the 58 words taken from the Connecticut norms. This rather small difference between the proportions was statistically reliable ($F = 10.76$, $df = 1/144$, $p < .01$).

The number of different responses (DR) made by the 146 Ss to each stimulus was determined. To avoid

arbitrary decisions, all variations in spelling (other than the plural) were taken as separate entries. DR ranged from eight responses to SPOOL to 77 responses to TROUBLE with a mean of 43, a median of 39, and a standard deviation of 13.28.

The time required to complete the Test (T-1) ranged from 7 min to 17 min while the time required to complete the Retest (T-2) ranged from 6 to 14 min.² There was a strong positive correlation between T-1 and T-2 ($r = .81$, $df = 143$, $p < .01$).

It was predicted that Ss who tended to give a large number of P responses on the Test would tend to give more IAs and complete the word association task more quickly than Ss who gave fewer P responses on the Test. The correlation between number of Ps on the Test and number of IAs was statistically reliable ($r = .84$, $df = 144$, $p < .01$), as was the negative correlation between number of Ps and T-1 ($r = -.27$, $df = 144$, $p < .01$). The correlation between the number of IAs and T-2 was not statistically reliable ($r = .15$, $df = 143$, $p > .01$).

In order to determine if Ss had a greater tendency to give IAs to high DP words than to low DP words, a correlation was computed between DP and number of IAs for each of the 158 words and yielded $r = .88$, $df = 156$, $p < .01$. The correlation between DP and LF as indexed by the Thorndike-Lorge word count (1944) did not differ significantly from chance ($r = -.10$, $df = 156$, $p > .01$), but the correlation between LF and number of IAs yielded $r = -.34$, $df = 156$, $p < .01$; when DP and LF were used as multiple predictors of IA, $R_{1.23} = .92$, $df = 155$, $p < .01$. This relationship suggested that a high degree of reliability can be predicted for those words that have a low frequency in the language and a highly dominant P.

A salient characteristic of the response hierarchy of any given stimulus word is the number of different words evoked by that stimulus (Linton & Brotsky, 1967). A high negative correlation ($r = -.71$, $df = 156$, $p < .01$) was found between DR and DP. A similar relationship ($r = -.75$, $df = 156$, $p < .01$) was found between DR and IA.

The TRR of .32 found in the present study was quite low, despite the fact that Hall (1966) found a higher degree of reliability in his study. Although it is possible that the difference in the results of the two studies might be due to differences in procedure or the longer Test-Retest interval used in the present experiment (10 weeks vs 3 weeks), a more likely explanation for the discrepancy may be that the words used in the present study had a much higher average LF than those used in Hall's study. In Hall's study, 27 of the 56 stimulus words had a frequency count of 1 to 5 occurrences per million. In the present study only four stimulus words fell into this category, while an additional eight words had a

frequency count of 6 to 10 occurrences per million. The average TRR for these 12 words was about .49, considerably higher than the average TRR for the 158 words in the stimulus set. Thus, the negative relationship between frequency in the language and TRR coupled with the large proportion of high frequency words used in the present study could readily account for the low TRR.

The occurrence of a negative correlation between TRR and LF was unexpected, particularly since Hall found no relationship between reliability and frequency. It seems likely that high frequency words acquire many associates whose probability of occurrence is high, resulting in competition among responses and low TRR. Thus, it might be argued that when DP is low and there are many other words in the hierarchy (high DR) there would be many possibilities for intrusions during the Retest, and TRR would necessarily be low. When DP is high and there are few DRs, as would be the case with low frequency words, there would be less likelihood of intrusions during the Retest, and TRR would necessarily be high.

The results of the present research suggest that it may be possible to form lists of stimulus words which would tend to evoke the same responses from Ss over considerable periods of time. It would be necessary that these words be carefully chosen, however, if reliability is considered a necessary prerequisite for valid predictions in studies of verbal behavior.

References

- Bousfield, W. A., Cohen, B. H., Whitmarsh, C. A., & Kincaid, W. D. The Connecticut free associational norms. *Tech. Rep. No. 35*. Contract N onr 631 (00) Office of Naval Research, 1961.
- Hall, J. F. The reliability of free association responses. *Psychon. Sci.*, 1966, 4, 79-80.
- Linton, Marigold L., & Brotsky, S. Joyce. Word associations: Oral vs. written stimuli and responses. Paper read at West. Psychol. Assoc., San Francisco, May, 1967.
- Marshall, G. R., & Cofer, C. N. Associative indices as measures of word relatedness: a summary and comparison of ten methods. *J. verbal Learn. verbal Behav.*, 1963, 1, 408-421.
- Palermo, D. S., & Jenkins, J. J. *Word association norms: grade school through college*. Minneapolis: University of Minnesota Press, 1963.
- Russell, W. A., & Jenkins, J. J. The complete Minnesota norms for responses to 100 words from the Kent-Rosanoff Word Association Test. *Tech. Rep. No. 11*, Contract N8 Onr-66216, 1954.
- Thorndike, E. L., & Lorge, I. *The teacher's word book of 30,000 words*. New York: Teacher College, Columbia University, 1944.

Notes

1. This research was supported in part by a Research Grant to the first author from the San Fernando Valley State College Foundation and by a Faculty Research Award to the second author from the San Diego State College Foundation.
2. The S with the largest T-1 did not record his T-2 as instructed.