Bulletin of the Psychonomic Society
1982, Vol. 19(1), 11-14

A comparison of three methods of measuring recognition memory in high and low academic achievers

RONALD LEY and JURGEN KARKER
State University of New York, Albany, New York 12222

This study compared three commonly used methods of measuring recognition memory (single-item, embedded-item, and multiple-choice tests) under conditions in which the targets and distractors were the same for all tests. The multiple-choice test resulted in higher recognitions and lower false recognitions than the other two tests. High academic achievers had higher recognition scores than low achievers on all tests, but the interaction between achievement level and type of test showed the high achievers to have lower false recognitions only under the multiple-choice test. Significant correlations between word assessment characteristics of the targets (associative frequency, associative reaction time, and pronunciability) and recognition scores were discussed in the context of corresponding correlations with free recall.

The fact that the method by which memory is measured has an effect on performance is readily apparent in recognition tests of memory, in which the relative ease or difficulty of correctly identifying a list of words or other verbal units depends on the quantity and quality of the distractors (recognition lures) included among the targets. That is, for a given list of words presented for study under a standardized procedure, the number of words correctly identified during the recognition test (targets and distractors randomly dispersed) will depend on the ratio of distractors to targets (e.g., Underwood, 1972) and on the similarity between the distractors and the targets with respect to semantic and formal structural characteristics of the targets (e.g., McNulty, 1965). Thus, in experiments in which recognition is measured by target-distractor discrimination tests, recognition scores for a standard list might be either greater or lesser than recall for the same list, depending on the selection of distractors. This problem is not a new one (e.g., Müller, 1913/1970), but it appears to be a frequently overlooked one, as evidenced by the number of experiments that purport to compare recall with recognition, as contrasted with the few experiments that Brown (1976) reports may have been successful in such a comparison.

In addition to distractor variables, another factor that might affect recognition scores is the format of the recognition test. According to Kausler (1974), the three most commonly used formats for measuring recognition are (1) the forced-choice single-item test, in which targets and distractors are presented at one time and subjects are instructed to identify "old" items (targets) and "new" items (distractors) (e.g., Seward, 1928), (2) the forced-choice embedded-item test, in which targets and distractors are listed on a sheet of paper and subjects are instructed to identify old items (e.g., Peixotto, 1947), and (3) the multiple-choice test, in which targets and distractors are presented in sets, each of which contains one target and two distractors, and subjects are instructed to identify old items (e.g., Shepard, 1967). The question raised here is, if targets and distractors are held constant, do recognition scores and false recognition scores differ under the three recognition test formats? The answer to this question is important for anyone interested in comparing the results of experiments in which different tests of recognition have been used. The primary purpose of the present experiment was to answer this question.

The importance of individual differences in the construction of theories has been emphasized by Underwood (1975), and a recent study of free recall by Dean and Ley (1977) illustrates Underwood's (1975, p. 128) proposal that nomothetic theories should be formulated "in a way that will allow an immediate individual difference test." The assumption that memory (recall and recognition) is a basic factor underlying intelligence is implicit, if not explicit, in theories of intelligence and academic achievement, as well as measures of intelligence and achievement. The secondary purpose of the present study was to determine whether subjects who differ in academic achievement (the upper and lower thirds of a class of high school seniors), and who presumably differ in intelligence as well, show uniform differences among the three measures of recognition. In view of the large number of college students (the upper third of high school seniors) used in studies

This research was supported in part by a grant-in-aid and research fellowship from the Research Foundation of the State University of New York to the first author. Requests for reprints should be sent to Ronald Ley, State University of New York at Albany, 1400 Washington Avenue, Albany, New York 12222.

0090-5054/82/010011-04$00.65/0

of memory, an interaction between methods of measurement and individual differences in achievement would have implications for research designed to test theories of recognition memory and research designed to measure intelligence in terms of recognition memory tests (Carroll, 1978; Carroll & Maxwell, 1979).

One of the important distinctions between recognition and recall is the difference in the effect of word frequency; high-frequency words are recalled more readily than are low-frequency words, but the opposite is often the case in recognition (e.g., Gorman, 1961; Schulman & Lovelace, 1970; Shepard, 1967). In the present experiment, we tested the hypothesis that the effect of frequency on recognition might be a function of the type of recognition test. If this is the case, correlations between frequency assessment measures and recognition scores should differ from test to test.

METHOD

Subjects and Design

The 144 subjects were seniors from an Albany area high school, 72 of whom were randomly selected from the upper third of their class, with respect to grades, and 72 from the bottom third. None of the subjects had previous experience in verbal learning studies. The design of the experiment was a 3 by 2 factorial in which the three methods of measurement (single-item, embedded-item, and multiple-choice tests) and two levels of academic achievement (high achievers and low achievers) were between-subjects factors.

Materials

Thirty-six target items and 72 distractors were randomly selected from Ley and Tesiny's (1975) list of 382 words and paralogs, with the single restriction that none of the items shared formal (structural) similarity with respect to their first three letters; that is, no two of the first three letters of any of the 108 items (36 targets plus 72 distractors) were the same.

The materials of the single-item test consisted of 48 decks of 108 Hollerith computer cards, each of which contained either one target or one distractor printed on the top row of the card. The materials for the embedded-item test consisted of 48 sheets of 8.5 x 11 in. paper, each of which contained four columns of 27 items per column. The column and row locations of the 36 targets and 72 distractors were randomly determined. The materials for the multiple-choice test consisted of 48 decks of 36 Hollerith computer cards, each of which contained one target and two distractors printed on the top row of the cards. The ordinal position of items was randomly determined for each card.

The materials used for the presentation of the target items during the study period consisted of 36 35-mm slides, each of which contained a single target item. The slides were projected on a film screen by means of a Kodak Carousel slide projector. Presentation rate was controlled by the internal timer of the projector.

Procedure

During the study phase, subjects were seated facing the film screen. All subjects received uniform instructions that required that they rehearse silently each unit as it appeared on the screen. The subjects were given four practice trials for the purpose of familiarizing them with the types of verbal items that were to be studied and with the 4-sec presentation rate.

Following the study phase, subjects were given brief instructions relevant to their respective recognition task. The 48 subjects in the embedded-item group were instructed to read down the typed sheet and to circle with a pencil those items presented during the study phase. The 48 subjects in the multiple-choice group were instructed to go through the deck and to circle the item, from among the three printed on each card, which was presented during the study phase. The 48 subjects in the single-item group were instructed to sort their decks (looking at each card only once), by placing cards containing items presented during the study phase in one pile and cards containing other items in a second pile.

RESULTS

Three performance measures were calculated for each subject: recognitions (total number of target items correctly identified as "old"), false recognitions (total number of distractors incorrectly identified as "old"), and corrected recognitions (total number of recognitions minus total number of false recognitions). Table 1 gives the means and standard deviations of the performance scores for the high and low achievers under the three recognition tests.

The analysis of variance of the recognition data indicated that recognition scores under the multiple-choice test (mean = 28.30) were significantly greater than those obtained under the embedded-item test (mean = 24.70) and the single-item test (mean = 25.14) [$F(2,138) = 6.40$, $p < .002$, $MSe = 30.79$], but according to Tukey's Honestly Significant Difference (HSD) test, the difference (D) between the means of the embedded-item and single-item tests ($D = 1.35$) was not significant ($p > .05$). The analysis also indicated that the recognition scores for the high achievers (mean = 28.57) were significantly greater than those for the low achievers (mean = 23.53) [$F(1,138) = 28.11$, $p < .001$]. The interaction was not significant [$F(2,138) = .61$, $p > .05$].

The analysis of variance of the corrected recognition scores followed the same pattern the recognition scores

Table 1
Mean Performance Scores for High and Low Achievers Under Three Recognition Tests

	Recognition Test						
	SI		EI		MC		
AG	Mean	SD	Mean	SD	Mean	SD	GM
			Recognitions				
High	27.21	5.01	27.78	5.64	30.71	3.96	28.57
Low	23.08	6.15	21.63	5.63	25.88	6.58	23.53
Mean	25.14		24.70		28.30		
			False Recognitions				
High	7.16	6.46	4.70	3.42	1.63	2.08	4.50
Low	6.29	4.70	4.33	3.88	4.50	3.11	5.04
Mean	6.72		4.52		3.06		
			Corrected Recognitions				
High	20.05	4.53	23.08	5.73	29.08	5.51	24.07
Low	16.79	7.24	17.30	5.25	21.38	7.38	18.49
Mean	18.42		20.19		25.23		

Note–SI = single item, EM = embedded item, MC = multiple choice; AG = achievement group; GM = grand mean.

followed: Scores under the multiple-choice test (mean = 25.23) were significantly greater than those under the embedded-item test (mean = 20.19) and the single-item test (mean = 18.42) [$F(2,138) = 14.87$, $p < .001$, MSe = 42.35], and the difference between the means of the embedded-item and single-item tests (D = 1.77) was not significant. Furthermore, the corrected recognition scores for the high achievers (mean = 24.07) were significantly greater than those for the low achievers (mean = 18.49) [$F(1,138) = 27.03$, $p < .001$], and the interaction was not significant [$F(2,138) = 1.10$, $p > .05$].

The analysis of the false recognition scores for the three tests showed a pattern consistent with the corrected recognition scores; that is, the multiple-choice test resulted in the lowest number of errors (mean = 3.06), the single-item test resulted in the highest number (mean = 6.72), and the embedded-item test resulted in the intermediate number (mean = 4.52) [$F(2,138) = 9.39$, $p < .001$, MSe = 17.42]. Furthermore, according to Tukey's test, differences between all three pairs of means were significant ($p < .05$). Although the difference between the false recognition scores of the high achievers (mean = 4.50) and low achievers (mean = 5.04) was not significant [$F(1,138) = .61$, $p > .05$], the interaction between achievement and type of test was marginal [$F(2,138) = 2.86$, $p = .059$]. A perusal of the data of Table 1 makes it clear that this observed interaction is attributable to the relatively small number of false recognitions for the high achievers under the multiple-choice test (mean = 1.63) compared with the other two high-achievement groups and with the low achievers under the multiple-choice test. Tukey's test showed that the observed differences that favored the low achievers over the high achievers under the single-item test (D = −.87) and embedded-item test (D = −.37) were not significant, whereas the difference between low and high achievers under the multiple-choice test (D = 2.87) favored the high achievers and was significant ($p < .05$).

Since the targets used in the present experiment consisted of items selected from a list rated for associative reaction time (Taylor & Kimble, 1967) and pronunciability (Ley & Karker, 1974), as well as for associative frequency (Locascio & Ley, 1972), correlation coefficients between these assessment characteristics and recognition scores for the high and low achievers under each of the three recognition tests were computed. All of the 18 coefficients, which are given in Table 2, were significant ($p < .025$, df = 34), but a test of the absolute difference between the largest correlation (r = .67) and the smallest (r = −.36) was not significant ($z = 1.48$, $p > .05$), thus none of the differences between any pair of correlations was significant. Nevertheless, strong trends among the coefficients suggest that the observed magnitudes of the correlations vary from test to test but not between achievement groups within each test; the largest correlations occur under the embedded-item test (except for pronunciability for low achievers, for which r = −.36) and the smallest under the single-item test. The trends also suggest that associative frequency and associative reaction time are about equally better predictors of recognition than is pronunciability under the single-item and multiple-choice tests, but that the differences in predictive accuracy are attenuated under the single-item test. Another notable trend is the remarkable similarity between pairs of correlations for high and low achievers for each assessment measure under each test, except for pronunciability under the embedded-item test.

Table 2
Correlation Coefficients Between Assessment Characteristics of the Target Items and Recognition Test Scores (df = 34)

Assessment Characteristic	Recognition Test					
	EI		MC		SI	
	H	L	H	L	H	L
Associative Frequency	.67	.61	.52	.55	.48	.53
Associative RT	−.64	−.62	−.55	−.52	−.41	−.42
Pronunciability	−.59	−.36	−.44	−.40	−.40	−.40

Note—EI = embedded item, MC = multiple choice, SI = single item; H = high achievers, L = low achievers, r = .33, p = .025; r = .42, p = .005.

DISCUSSION

The data of the present experiment show clearly that measuring memory by means of the multiple-choice recognition test results in significantly higher scores than measurement by means of forced-choice tests when distractors are constant. However, the two forced-choice tests (single item and embedded item) resulted in very similar scores: The small observed differences between the single-item and embedded-item tests for recognition scores (D = .44) and for corrected recognition scores (D = −1.77) can most reasonably be attributed to sampling error. Although all three tests differed from each other on false recognition scores, the multiple-choice test resulted in significantly fewer false recognitions than did eight of the other two. It is obvious that the multiple-choice test is the most lenient of the three, and in view of the significant difference between the false recognition scores of the forced-choice tests, a meaningful comparison of recognition with recall of the same list could not be made without specification of the method of measurement. It may be that any test of recognition that employs distractors precludes comparison of recognition with recall, unless the same set of distractors can be reasonably assumed to be equally available during the test of recall.

The correlations between the assessment characteristics of the targets with recognition scores under the three tests may provide an indirect comparison of recognition with recall. In an earlier study (Ley & Karker, 1976) in which we measured free recall of a sample of 36 items selected from the same list as the targets used in the present study, we found a correlation of .71 between associative frequency and recall, a coefficient that is larger than any of the correlations between associative frequency and recognition of the present study, albeit the differences fell short of significance (e.g., D = .71 − .48, $z = 1.48$, $p = .069$). However, the correlation of .71 was significantly greater ($p < .05$) than all of the correlations between pronunciability and recognition, except that for high achievers under the embedded-item test (r = −.59). Additional data from the earlier study showed recall to correlate −.56 with associative reaction time and −.41 with pronunciability, coefficients that approximate closely corresponding correlations of the present study.

Although these assessment data do not allow for neat, clear-cut distinctions between recognition and recall, the trends support the general finding that frequency effects are stronger in recall than in recognition, but, contrary to the "frequency paradox," there was no inverse effect for recognition.

The data that show that the high achievers had higher recognition scores and corrected recognition scores than the low achievers are consistent with the notion that memory is a strong component of intelligence. Similarly, the lower false recognition scores on the multiple-choice test for the high achievers than for the low achievers was expected, but the finding that the high achievers did not have lower false recognition scores on the forced-choice tests was not. Whatever the explanation for the interaction, the important fact is that differences in false recognition scores between high and low achievers depend on the method by which recognition is measured.

REFERENCES

BROWN, J. An analysis of recognition and recall and of problems in their comparison. In J. Brown (Ed.), *Recall and recognition.* London: Wiley, 1976.

CARROLL, J. B. How shall we study individual differences in cognitive abilities: Methodological and theoretical perspectives. *Intelligence*, 1978, **2**, 87-115.

CARROLL, J. B., & MAXWELL, S. E. Individual differences in cognitive abilities. In M. R. Rosenzweig & L. W. Porter (Eds.), *Annual review of psychology.* Palo Alto, Calif: Annual Reviews, 1979.

DEAN, J., & LEY, R. Effects of associative encoding on free recall in high and low verbal associators. *Journal of Experimental Psychology: Human Learning and Memory*, 1977, **3**, 316-324.

GORMAN, A. M. Recognition memory for nouns as a function of abstractness and frequency. *Journal of Experimental Psychology*, 1961, **61**, 23-29.

KAUSLER, D. H. *Psychology of verbal learning and memory.* New York: Academic Press, 1974.

LEY, R., & KARKER, J. Pronunciability ratings of 319 CVCVC words and paralogs previously assessed for meaningfulness and associative reaction time. *Bulletin of the Psychonomic Society*, 1974, **3**, 421-424.

LEY, R., & KARKER, J. Meaningfulness, associative reaction time, recognition latency, and pronunciability in free recall. *Bulletin of the Psychonomic Society*, 1976, **8**, 231-232.

LEY, R., & TESINY, E. Associative reaction time, meaningfulness, and pronunciability ratings of 382 words and paralogs. *Bulletin of the Psychonomic Society*, 1975, **6**, 645-648.

LOCASCIO, D., & LEY, R. Scaled-rated meaningfulness of 319 CVCVC words and paralogs previously assessed for associative reaction time. *Journal of Verbal Learning and Verbal Behavior*, 1972, **11**, 243-250.

MCNULTY, J. A. An analysis of recall and recognition processes in verbal learning. *Canadian Journal of Psychology*, 1965, **19**, 188-195.

MÜLLER, G. E. Zur Analyse der Gedächtnistätigkeit und des Vorstellungsverlaufes, III. Teil. In W. Kintsch (Ed.), *Learning, memory, and perceptual processes.* New York: Wiley, 1970. (Originally published, 1913.)

PEIXOTTO, H. E. Proactive inhibition in the recognition of nonsense syllables. *Journal of Experimental Psychology*, 1947, **37**, 81-91.

SCHULMAN, A. I., & LOVELACE, E. A. Recognition memory for words presented at a slow or rapid rate. *Psychonomic Science*, 1970, **21**, 99-100.

SEWARD, G. H. Recognition time as a measure of confidence. *Archives of Psychology*, 1928, **16**, 1-54.

SHEPARD, R. N. Recognition memory for words, sentences, and pictures. *Journal of Verbal Learning and Verbal Behavior*, 1967, **6**, 156-163.

TAYLOR, J., & KIMBLE, G. Association value of 320 selected words and paralogs. *Journal of Verbal Learning and Verbal Behavior*, 1967, **6**, 744-752.

UNDERWOOD, B. J. Word recognition memory and frequency information. *Journal of Experimental Psychology*, 1972, **94**, 276-283.

UNDERWOOD, B. J. Individual differences as a crucible in theory construction. *American Psychologist*, 1975, **30**, 128-134.

(Received for publication October 1, 1981.)