Bulletin of the Psychonomic Society
1984, 22 (6), 497-500

Metamemory: Monitoring future recallability in free and cued recall

EUGENE A. LOVELACE
University of Virginia, Charlottesville, Virginia

Data are reviewed from three studies in which individuals rated, at time of study, the likelihood that they would later recall each to-be-remembered item. In all conditions, individuals show a very substantial ability to predict those items that they later succeed in recalling. Although performance on this variety of metamemory task was similar in several respects for free-recall and cued-recall tasks, one consistent difference emerged. In the cued, or paired-associate, task, individual differences in recall performance were related to individual differences in the use of the predictive rating scale; those who recalled more items used higher values on the rating scale. For free recall, no such relationship was observed.

Metamemory refers to knowledge of one's own memory processes. This includes several types of memory monitoring (see Cavanaugh & Perlmutter, 1982, for a recent review). The particular variety of episodic memory monitoring of interest here is the ability to predict, at time of study, the relative memorability of individual items on a later memory test. Previous studies have shown that college students are fairly good at making such predictions for free recall (Groninger, 1979; Underwood, 1966), for paired-associate (PA) learning (Arbuckle & Cuddy, 1969; King, Zechmeister, & Shaughnessy, 1980; Lovelace, 1984), and for recall of the remainder of sentences when cued with the initial three words (Lovelace, 1984).

The basic procedure for this sort of monitoring requires that the individual subject rate each to-be-remembered item at the end of the study period as to how likely he or she thinks it is that that item will be recalled on a later memory test. For the two experiments to be described here, and the one experiment from an earlier study to be cited in detail (Lovelace, 1984, Experiment 1), the stimulus materials were all common English words chosen from the Paivio, Yuille, and Madigan (1968) norms. The following average values are very close to the specific values for each list in each of those three experiments. Imagery ranged from a low of around 2.60 to a high of around 6.75, with a mean of about 5.65 (SD ≈ 1.10). Concreteness ranged from around 2.95 to 7.35, with a mean of about 5.75 (SD ≈ 1.38). Meaningfulness ranged from around 4.50 to 8.55, with a mean of about 6.40 (SD ≈ .95). Frequency ranged from around 21 to AA; with A and AA treated as 50 and 100, respectively, the lists had means of about 63 (SD ≈ 29).

The major purpose of the present paper is to report an interesting difference in college students' performance on this monitoring task for free recall versus cued (PA) recall of words. The metamemory differences observed as a function of task reveal that individuals who had higher average ratings showed superior recall performance in the cued-recall task but not in free recall. This difference was discovered rather than sought; it was not the focus of the research design, and the analyses revealing this finding were not motivated by a hypothesis deduced from a theoretical viewpoint.

EXPERIMENT 1

This experiment was designed to explore the relative accuracy of prediction of later recall in a free-recall task as a function of whether or not the individual had prior practice with the prediction rating and free-recall tasks. This entailed simply having the same individual rate and then attempt free recall of words for two successive lists.

Method

Materials. Two lists of 90 words each were selected from the Paivio et al. (1968) norms, as previously described.

Subjects. Thirty-four students enrolled in an introductory psychology course at the University of Colorado participated in groups ranging in size from two to seven.

Procedure. The two lists were projected at an 8-sec rate. The individuals made their predictions of later recall during the single 8-sec study interval, using a 6-point rating scale that was described as follows: to circle 1 "means you believe that you *almost surely* will *not* recall the word," 2 means "you believe you probably will not recall it," 3 means you "think you may not recall it," 4 means "you think you may recall it," 5 means "you believe that you probably will recall it," and 6 means "you are *quite certain* that you *will* recall it."

Between the study-rating presentation and the recall attempt, to minimize immediate memory contributions, the students were given 2 min to complete as many multiple-column subtraction problems as they could. They were then allowed 4 min to write down as many words as they could recall. They were instructed "if you think that a word was on the list, but you are not sure, go ahead and put it down, but do not give wild guesses."

After a brief rest, this was followed by the study-rating trial of the second list (List 2); the individuals were explicitly informed prior to this list that they would never again be asked to recall the words from the first list (List 1). The filler task following this list was a multiplica-

The author's mailing address is: Department of Psychology, Gilmer Hall, University of Virginia, Charlottesville, VA 22901.

tion task for 2 min which was followed by 4 min of written recall of List 2.

Results and Discussion

The distributions of ratings were very similar on the two lists; the proportion of all ratings falling in values 1 to 6, respectively, were .06, .20, .29, .25, .13, and .08 for List 1 and .13, .23, .27, .20, .09, and .08 for List 2. The individuals did not substantially change their use of the rating scale as a result of their experience with attempting recall of List 1.

For *all* individuals, on both lists, the mean ratings for items recalled were greater than the mean ratings for items not recalled; the consistency of this difference indicates success in the monitoring task. Although the difference score (mean rating for items recalled minus mean rating for items not recalled) might be taken as an index of predictive accuracy, it does not take cognizance of the range or dispersion of the individual's ratings. To quantify the predictive accuracy for each individual, a predictive accuracy quotient (PAQ) was calculated in a fashion similar to that introduced by Zimmerman, Broder, Shaughnessy, and Underwood (1977) and employed by King et. al. (1980). The PAQ score for an individual is that difference score divided by the pooled variance of the ratings of recalled and of non-recalled items for that individual.

The mean PAQ score did not change appreciably from List 1 (mean = .60, SD = .29) to List 2 (mean = .66, SD = .30) ($t < 1$). Clearly, the ability to monitor the relative recallability of words in this free-recall task was about equally good on the two lists. This is consistent with the observation of King et al. (1980) that they found "little or no improvement in prediction accuracy across lists" (p. 339) in a PA task. The curve in Figure 1 labeled FR-1 provides a summary of the combined data from the two lists by showing the conditional probabilities of recall given ratings.

In an earlier study (Lovelace, 1984, Experiment 1) employing 40 pairs of words with these same stimulus characteristics in a PA task, it was found that individuals who recalled greater numbers of items tended to use higher ratings than did those who recalled fewer words. Pearson correlations of individuals' overall mean ratings with the number of items they recalled were .28, .49, and .49 for three different study conditions of that experiment. A similar analysis of the data for the present study yielded weak *negative* correlations, −.16 for List 1 and −.04 for List 2. These combined results suggest that individuals have some sense of whether they will do well or poorly on the PA task in some absolute sense, or at least relative to other students, and use different portions of the rating scale accordingly. For the free-recall task, on the other hand, the individuals appear insensitive to individual differences in their ability on the task.

One way in which the present free-recall study differed from the earlier PA study, however, in addition to the nature of the recall task, is that ratings were made on the *single* trial in the present experiment, whereas in two of the three conditions of the earlier PA study, the pairs were seen more than one time before the ratings were given. For those two conditions, the correlations (both .49) were significant. The weakest, and only nonsignificant, correlation in the earlier study (.28) was found for the one condition in which ratings were made on a single study trial.

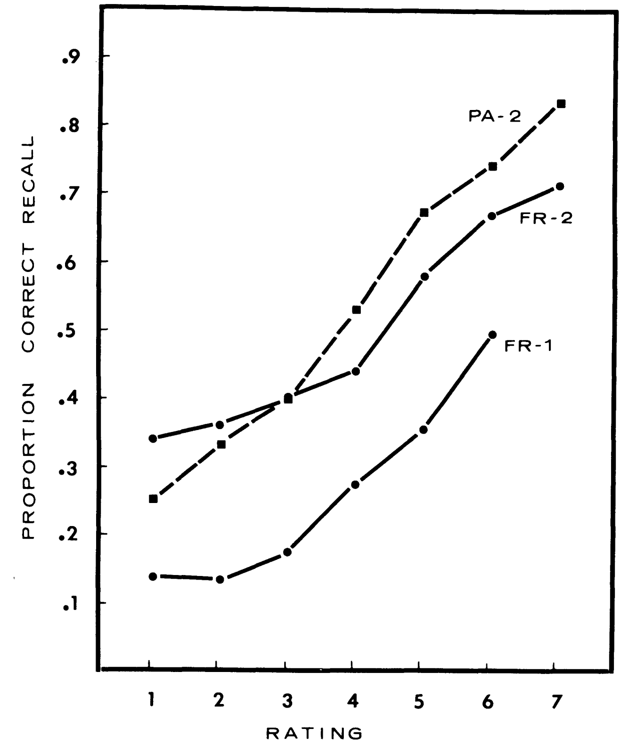

Figure 1. The mean conditional probabilities of recalling items as a function of ratings they received at time of study.

The data reported below from Experiment 2, however, indicate that the low correlations observed for the free-recall task in Experiment 1 did not result simply from the fact that the ratings were made on the initial and only study trial, but rather appear to be characteristic of that task.

EXPERIMENT 2

In this study, the same words constituted the to-be-remembered units in either a 60-word free-recall task or a 60-pair PA task. In both tasks, the predictive ratings were made on the fourth study trial, with no test trials having been administered.

Method

Materials. The 120 nouns used ranged in length from five to nine letters; synonymity or obvious strong associative relations between words was minimized by inspection. The words were divided into two lists of 60 words each; one of these constituted the materials for free recall, and materials for a PA task were generated by pairing each word in one list with a word from the other list.

Subjects. Eighty students who were enrolled in summer session courses at the University of Virginia participated in groups of nine or fewer.

Design and Procedures. Forty students were assigned to each of two experimental conditions: cued recall and free recall. In both conditions, stimuli were presented by slide projector at a 3-sec rate for three consecutive study trials and then at a 5-sec rate on a fourth study trial, on which ratings of the memorability of the items were made. Effect of primacy and recency positions were minimized by randomly reordering materials from trial to trial. In the free-recall condition, the 60 words appeared on the screen singly; in the cued-recall condition, the words appeared as pairs, one word above the other. In the free-recall condition, the individuals were told that they would later be given a blank piece of paper and asked to write down all the words they could recall; in the cued-recall condition, they were told that they would later be shown either the top or the bottom word of the pair and asked to recall the other word. Prior to the fourth study trial, each individual was instructed that he or she was to "rate each word (pair) for how well you think you will be able to recall that particular word (pair)" by circling a number from 1 to 7 for each item. (Instructions anchoring the ratings were similar to those presented above for Experiment 1, with the added midpoint rating of 4 as a neutral or "uncertain" judgment.)

In the free-recall condition, the individuals were given a sheet with 60 lines and told they had 5 min to recall as many words as they could; they were encouraged to adopt a low criterion for response, being told "if you think a word was on the list, but are not certain, go ahead and put it down." In the cued-recall condition, a single test trial was given at a 6-sec rate, with the upper member of half the pairs being presented as the recall cue and the lower member of the remaining pairs being presented. For each cue word, the subjects were strongly urged to write down a response, giving their "best guess" if they were uncertain.

Results and Discussion

The distributions of ratings were quite similar for the two recall tasks. The proportion of all ratings falling at each value from 1 to 7 were .07, .10, .13, .15, .18, .16, and .21 for the free-recall task and .09, .11, .16, .16, .22, .14, and .13 for the cued recall (PA) task.

The mean PAQ score for the free-recall condition (mean = .65, SD = .30) did not differ significantly from that of the cued-recall condition (mean = .73, SD = .41) [t(78) = 1.58]. In Figure 1, the curves labeled PA-2 and FR-2 display the conditional probabilities of correct recall given ratings from 1 to 7 for the two tasks. The cued-recall condition clearly produced a steeper function there, despite the absence of a significant difference in PAQ scores. Note that, in Figure 1, each individual does not contribute equally to all points on a curve, but only to those for ratings that that person used. If those individuals recalling the greatest number of items also use a higher set of ratings than do the people who recall fewer items, the slope is increased due to these *inter*individual effects. There is clear evidence for such in the PA task; the Pearson correlation of overall mean ratings with number of words the person recalled was .49 ($p < .01$). For the free-recall task, however, the correlation was near zero ($r = .04$), confirming the lack of relationship seen in Experiment 1 above. Clearly, the significant difference in these correlations ($Z = 2.15$, $p < .05$) may be more reasonably attributed to some aspect of the tasks than to the number of study trials prior to rating.

Lower correlations can always result from problems of restriction of range. Although the means of the individuals' mean overall ratings were fairly similar for free and cued recall (4.58 and 4.24, respectively) and were close to the midpoint of the scale (4.0), the standard deviation of individuals' overall mean ratings was somewhat greater for the cued-recall (SD = .76) than for the free-recall (SD = .60) tasks. Furthermore, the range of these individuals' mean ratings was greater for cued recall (2.35 to 6.75) than for free recall (3.28 to 5.68). On the recall performance side, again the mean proportions correctly recalled were similar for free-recall (.54) and cued-recall (.56) tasks, but the standard deviation of individuals' mean recall scores was somewhat greater for cued recall (.24) than for free recall (.17); also, the range of mean proportions was greater for cued recall (.15 to .97) than for free recall (.27 to .83)

Although these statistics are of the sort one might expect, given the differences in correlations of mean overall rating and proportion recalled, they do not address why the PA task leads to greater individual differences in the use of the rating scale. One might be tempted to ascribe the correlation in the PA case simply to the fact that individuals who use higher ratings thus exhibit or even create a greater motivation to perform well; there is no obvious reason, however, why such a motivational account should not similarly apply to the free-recall task. It is possible that the differential use of the rating scale has to do with the similarity of each task, as individuals perceive them at time of study, to other, real-world tasks or situations in which the individuals have knowledge of their own absolute or relative performance on past occasions. That is, if it is supposed that the free-recall task presents a situation that is more novel and less structured than the PA task, individuals consequently are less likely to feel high or low confidence in their ability to perform this sort of task, on the basis of prior performance in other situations. Thus, more individuals in the free-recall task will stick closer to the middle of the rating scale.

A second possibility is that the basis for selecting a particular mean level of ratings is the same for both tasks, but it is determined by a factor that is relevant to only one of the two tasks. For example, suppose that individuals' knowledge of their relative status with respect to some general trait, such as verbal IQ, was the basis of their choice of level in the use of the rating scale for both tasks. It might then be that the trait, as they have knowledge of it, really taps some aspects of verbal associability, which is very salient for a PA task but is not a trait that is particularly predictive for free recall (cf. Underwood, Boruch, & Malmi, 1978, for the notion that paired associates and free recall tap different factors in individual differences for memory performance).

GENERAL DISCUSSION

Two other findings reported by Lovelace (1984) are confirmed in the data from the present experiments. First is the observation that no relationship exists between amount recalled and ability to discriminate on the metamemory monitoring task. The Pearson correlations of number of words individuals correctly recalled with their PAQ scores in Experiment 1 were $r = -.01$ and $r = .05$ for Lists 1 and 2, respectively. For the free-recall condition of Experiment 2, this correlation was .23, whereas, for the PA task, it was −.15. None of these correlations approach statistical significance.

The second finding is that there is a substantial idiosyncratic component of the predictive ratings, a kind of privileged knowledge of the effectiveness of one's own encoding processes. If there is no privileged information involved in making the ratings, that is, if each individual is not responding on the basis of his or her unique encoding operations but rather on the basis of properties intrinsic to the materials and common to all individuals, then a given individual's recall might be equally well predicted from the rating of any individual. In the absence of privileged knowledge of the learner's encodings, then, the distribution of the values of PAQs should be about the same whether the recall of each individual is related to his or her own predictive ratings or to those of a randomly yoked individual.

This was not the case. In Experiment 1, the mean PAQs for Lists 1 and 2 dropped from .60 and .66, respectively, when based on the individual's own ratings, to .26 and .18 when based on ratings of a randomly yoked individual. Similarly, the free-recall condition of Experiment 2 showed a drop from a mean PAQ of .65 based on one's own ratings to .12 for the randomly yoked; the cued-recall (PA) task showed a drop from .73 to .26. In all cases, these drops were so consistently present across subjects that they were significant by a simple sign test (all ps < .01). Evidence for such an idiosyncratic component also has been demonstrated in previous studies of free recall (Underwood, 1966) and paired associates (Lovelace & Marsh, in press; Rabinowitz, Ackerman, Craik, & Hinchley, 1982).

In summary, although this variety of predictive memory monitoring is similar in many respects for cued and free recall, the two tasks clearly differ in the extent to which individual differences in recall test performance are reflected in differential use of the prediction rating scales. In cued (PA) recall, those individuals who recall the greatest number of items tend to have used higher mean ratings at time of prediction; for free recall of word lists, on the other hand, there is no relationship between the mean level of ratings and the amount recalled. This effect was interpreted as being due to the differential extent to which individuals' knowledge of their performance on other memory tasks is relevant to their ability to perform in this particular task. The greater novelty and lack of structure in the free-recall task reduces the value that knowledge of individual differences in prior memory tasks has for use of the predictive rating scale.

REFERENCES

Arbuckle, B. F., & Cuddy, L. L. (1969). Discrimination of item strength at time of presentation. *Journal of Experimental Psychology*, **81**, 126-131.

Cavanaugh, J. C., & Perlmutter, M. (1982). Metamemory: A critical examination. *Child Development*, **53**, 11-28.

Groninger, L. D. (1979). Predicting recall: The "feeling-that-I-will-know" phenomenon. *American Journal of Psychology*, **92**, 45-48.

King, J. F., Zechmeister, E. B., & Shaughnessy, J. J. (1980). Judgments of knowing: The influence of retrieval practice. *American Journal of Psychology*, **83**, 329-343.

Lovelace, E. A. (1984). Metamemory: Monitoring future recallability during study. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **10**, 756-766.

Lovelace, E. A., & Marsh, G. R. (in press). Prediction and evaluation of memory performance by young and old adults. *Journal of Gerontology.*

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

Rabinowitz, J. C., Ackerman, B. P., Craik, F. I. M., & Hinchley, J. L. (1982). Aging and metamemory: The roles of relatedness and imagery. *Journal of Gerontology*, **37**, 688-695.

Underwood, B. J. (1966). Individual and group predictions of item difficulty for free learning. *Journal of Experimental Psychology*, **71**, 673-679.

Underwood, B. J., Boruch, R. F., & Malmi, R. A. (1978). Composition of episodic memory. *Journal of Experimental Psychology: General*, **107**, 393-419.

Zimmerman, J., Broder, P. K., Shaughnessy, J. J., & Underwood, B. J. (1977). A recognition test of vocabulary using signal-detection measures, and some correlates of word and nonword recognition. *Intelligence*, **1**, 5-31.

(Manuscript received for publication June 18, 1984.)