Bulletin of the Psychonomic Society
1992, 30 (2), 125-128

Memory-monitoring accuracy as influenced by the distribution of retrieval practice

JOHN J. SHAUGHNESSY
Hope College, Holland, Michigan

and

EUGENE B. ZECHMEISTER
Loyola University, Chicago, Illinois

An experiment was done to determine whether retrieval practice improved judgment-of-learning (JOL) accuracy when degree of learning was controlled. Fifty undergraduate students were asked to learn a long list of unrelated facts, with critical items presented either once or four times. The repetitions of critical items were retrieval prompts for half of the subjects (study-test) and additional study presentations (study-only) for the other half of the subjects. The subjects made JOL ratings after the last occurrence of critical items. Immediately after the study list, they were given a cued-recall test. Recall was comparable for once-presented items and repeated items across the two groups, but JOL accuracy was higher for repeated items in the study-test group. These results confirm that retrieval practice enhances JOL accuracy even when degree of learning is controlled.

Over 20 years ago, Tulving and Madigan (1970) mildly chastised memory researchers for paying too little attention to memory-monitoring abilities:

> No extant conceptualization, be it based on S-R associations or an information processing paradigm, makes provisions for the fact that the human memory system cannot only produce a learned response to an appropriate stimulus or retrieve a stored image, but it can also rather accurately estimate the likelihood of its success in doing it. (p. 477)

There was substantial evidence supporting their positive assessment of memory-monitoring accuracy. Hart (1965), using the feeling-of-knowing paradigm, had shown that subjects could accurately predict which nonrecalled items they would subsequently be able to recognize. Underwood (1966), using an ease-of-learning paradigm, had shown that subjects could predict (prior to learning the items) which paired-associate and free-recall items they would later find easiest or most difficult to learn. Arbuckle and Cuddy (1969) asked subjects who had just studied a given item to predict the likelihood that they would recall the item on a subsequent test (predictions of this type are now referred to as judgments of learning). They found that subjects accurately discriminated items that they would subsequently recall or not recall.

Results of memory research following Tulving and Madigan's (1970) call for more attention to memory-monitoring processes have led to a more mixed assessment of memory-monitoring accuracy (see, e.g., Begg, Duft, LaLonde, Melnick, & Sanvito, 1989; Glenberg, Wilkinson, & Epstein, 1982; Nelson & Leonesio, 1988). Because the present research involves the judgment-of-learning (JOL) paradigm, the review of more recent research will focus on this aspect of memory-monitoring accuracy. There have been several studies replicating JOL accuracy (e.g., Groninger, 1979; King, Zechmeister, & Shaughnessy, 1980; Lovelace, 1984). There have also been several studies, however, demonstrating JOL inaccuracy. Zechmeister and Shaughnessy (1980) found that subjects gave comparable JOL ratings to massed and distributed items even though the distributed items were much more likely to be recalled. Shaughnessy (1981) similarly found that subjects gave comparable JOL ratings to items studied under either maintenance or elaborative rehearsal even though the items studied under elaborative rehearsal were subsequently recalled more often by these same subjects. Cohen (1988) found substantial JOL accuracy for word recall but reported very low JOL accuracy for recall of subject-performed tasks (e.g., subject breaks a toothpick). Clearly, subjects are not always accurate in monitoring their memories.

Given that memory monitoring is not always accurate, it is reasonable to ask whether there are ways to improve or enhance memory-monitoring accuracy. There is some evidence that giving subjects test trials (retrieval practice) prior to their making judgments of learning increases the accuracy of their ratings (King et al., 1980; Lovelace, 1984). King et al. asked two groups of subjects to learn two lists of 24 paired associates. One group was given

The present experiment is a better controlled version of a similar experiment we reported in a paper presented at the Twentieth Annual Meeting of the Psychonomic Society in November 1979. The present experiment involved more lists so that a greater variety of facts were tested in each list function. Completion of this research project was facilitated by a Hope College faculty development grant made possible by the Matthew J. and Anne C. Wilson Foundation. Requests for reprints should be addressed to John J. Shaughnessy, Department of Psychology, Hope College, Holland, MI 49423.

three study-test cycles, and the other group was given five study trials. Both groups were then asked to make JOL ratings, which were followed by a final test trial. Both groups were then given a third list with only three study trials, which were followed by a JOL trial and a final test trial. On the first two lists, subjects given the study-test cycles showed consistently higher JOL accuracy. On the third list, however, the JOL accuracy of the study-test group decreased to the level of the group that was given only study trials on the first two lists. The latter finding suggests that retrieval practice provides specific information that enhances JOL accuracy. It is likely, for example, that subjects remember the outcome of the test trials and use this information at the time of making their JOL rating. Gardiner and Klee (1976) have shown that subjects are very accurate in assessing recall outcomes following a free-recall test.

There is one troublesome aspect of the King et al. (1980) findings that keeps them from being definitive evidence for the beneficial effect of retrieval practice on JOL accuracy. In their experiment, King et al. (1980) attempted to adjust for degree-of-learning differences between the groups with and without test trials by varying the number of study and test trials; nevertheless, degree-of-learning differences remained. Nelson, Leonesio, Shimamura, Landwehr, and Narens (1982) have shown that the accuracy of feeling-of-knowing judgments is affected by degree of learning. Although feeling-of-knowing judgments and JOLs are not highly correlated (Leonesio & Nelson, 1990), it is still reasonable to expect that JOL accuracy would be influenced by degree-of-learning differences. What is needed is a test of the effect of retrieval practice on JOL accuracy when degree of learning is controlled. That is what we set out to accomplish in the present experiment. We expected to find that retrieval practice does enhance JOL accuracy when degree-of-learning differences have been controlled.

METHOD

Materials

The study lists included 43 unrelated facts selected from a larger set used in a previous experiment. The 43 facts were selected to represent a midrange of recall likelihood. Each item was a relatively obscure fact from which a salient piece of information was tested—for example, "The Confederate general who fought against the Federals led by General Rosecrans in the Battle of Chickamauga in the American Civil War was General Bragg." The to-be-tested piece of information in each fact was highlighted by an underline. The retrieval prompt for this fact was "What is the name of the Confederate general who fought against the Federals led by General Rosecrans in the Battle of Chickamauga in the American Civil War?"

Of the 43 items, 25 were designated critical items and 18 were filler items. There were five types of critical items in each study list. Two types of critical items were presented once. Half of these items were rated for JOL, and half were not rated. The remaining three types of critical items were presented four times. The spacing of repetitions after the initial presentation of a repeated critical item followed one of the following schedules: 0-0-0—massed repetition of items with no other items intervening between successive repetitions; 5-5-5—constant distributed repetition with five other items occurring between successive repetitions; 1-5-9—expanding distributed repetition with increased spacing between successive repetitions. Landauer and Bjork (1978) had shown that recall was better following an expanding series of test trials. We included patterns similar to the ones they used so we could see if the sequence of test trials would affect JOL accuracy. There were five facts randomly assigned to each of the five types of critical items in the first study list. Four additional study lists were constructed by rotating the five sets of facts across the five critical-item functions.

The study lists had a total of 146 positions. Lists were constructed by dividing the body of the list (excluding a primacy section of 15 positions and a recency section of 4 positions) into five approximately equal sections. One of each of the five types of critical items was placed in each section, with the position of last occurrence of the five critical items kept as comparable as possible. The 18 filler items were used to complete the list. Two were presented six times, 12 were presented three times, and 4 were presented twice. At each frequency, half of the items were presented as massed repetitions and half as distributed repetitions. The filler items held the same positions across all five study lists.

Design and Subjects

The most important independent variable in this experiment was whether the subjects were given retrieval prompts during study (study-test group) or they were not (study-only group). Successive repetitions of 0-0-0, 5-5-5, and 1-5-9 critical items involved retrieval prompts for the study-test group and presentations of the facts for the study-only group. The five forms of the study list introduced a second independent variable, and 5 subjects were assigned, according to a block-randomized schedule, to the 10 groups defined by the list- and study-group variables. The 50 subjects were introductory psychology students at Hope College who volunteered to participate for research credit. The five types of critical items represent a within-subject variable, so the overall design was a 2×5×5 mixed factorial.

Procedure

All subjects were tested individually and they were told that their task was to try to learn as many facts as possible. The subjects were told that the critical piece of information in each fact would be underlined and that after all the facts had been presented they would be asked to recall the critical piece of information in response to a question. The subjects were told that some of the facts would be repeated; those in the study-test group were told that repetitions would appear as questions and that they should try to say the answer to themselves when the question appeared. The subjects were told that they would be given only a short time to study each fact (actual time was 6 sec). Each fact or question was presented on an index card, and the subjects were told to turn over one card each time they heard a tone from the tape recorder.

The subjects in both the study-only and the study-test groups made JOL ratings after the last occurrence of repeated critical items and after the only occurrence of half of the once-presented items. We asked the subjects to rate only half of the once-presented items and none of the filler items to keep them from trying to anticipate which items would be rated. The fact statement was presented for 6 sec while the subjects made the JOL rating. The cards for the to-be-rated items were marked RATE THIS FACT across the top of the card. The subjects made their JOL ratings by using a percent likelihood scale, but they were restricted to the responses 0, 20, 40, 60, 80, and 100. The subjects were informed that ratings above 50 were to be used for facts likely to be recalled and that ratings below 50 were to be used for facts not likely to be recalled. The subjects were instructed, however, that they should attempt to be as precise as possible when rating their probability of recall within these two major likelihood categories. The subjects were encouraged to ask questions about the JOL task and about the experiment in general before the experimenter began the study list.

RESULTS

The nonrated once-presented items and the massed items were included in this experiment to address research questions that are not the focus of this report. Therefore, only the data for the rated once-presented items and for the distributed repeated items (5-5-5 and 1-5-9) will be presented. All tests described as statistically significant were

Table 1
Recall Performance as a Function of Condition and Item Type

	Item Type								
	Once-presented			5-5-5			1-5-9		
Condition	*M*†	*SD*	% Recalled	*M*	*SD*	% Recalled	*M*	*SD*	% Recalled
Study only*	.84	.92	16.80	1.52	1.33	30.40	1.48	1.14	29.60
Study-Test*	.76	1.03	15.20	1.88	1.45	37.60	2.16	1.43	43.20

*n = 25. †Mean recall (maximum of 5).

evaluated by using a .05 level of significance. Error terms for F tests will be reported by the symbol MS_e.

Recall

The recall results were examined first to determine if comparable degrees of learning had been obtained in the study-only and study-test groups. The mean recall (a maximum of 5) with the corresponding standard deviations, along with the percent recall for the three types of critical items in the study-only and study-test groups, are presented in Table 1. The overall analysis of these data included the five study-list forms, but there was no main effect of forms, nor were there any interactions with the form variable. There was also no interaction of the study-condition variable and the item-type variable [$F(2,80) = 2.14$, $MS_e = .85$]. The main effect of item type was statistically significant [$F(2,80) = 18.32$, $MS_e = .85$]. Not surprisingly, an analytical comparison of the overall means showed that repeated items (1.76) were better recalled than were once-presented items [.80; $F(1,80) = 54.21$, $MS_e = .85$]. The effect that most directly tests for degree-of-learning differences is the main effect of study condition. Overall, the mean recall difference between the study-only group and the study-test group was not statistically significant [$F(1,40) = 1.23$, $MS_e = 3.13$]. Using the same error term, separate comparisons were made for the once-presented items and for the repeated items (combining 5-5-5 and 1-5-9). Neither comparison was statistically significant [$F(1,40)$ = .03 and 2.16, for once-presented items and for repeated items, respectively]. By careful selection of items and by systematic rotation of these items across list functions (and by a stroke of good fortune), comparable degrees of learning across study conditions have been obtained in this experiment.

Judgments of Learning

The JOL accuracy was measured by computing for each subject the mean JOL rating for recalled and for nonrecalled items within the categories of once-presented, 5-5-5, and 1-5-9 items. The effect of this variable (which will be called the recall variable) reflects the accuracy with which the subjects could discriminate at the time of study between items they subsequently did or did not recall.[1] The primary question of interest regarding JOL accuracy is the comparison of the accuracy for repeated items in the study-only and study-test groups. A preliminary analysis was done, however, before this primary analysis. The preliminary analysis was done to confirm that the study-only and study-test groups did not differ in JOL accuracy for once-presented items. Only 12 subjects in the study-only group and 10 subjects in the study-test group had entries for both recalled and nonrecalled once-presented items. The means for recalled and nonrecalled items in the study-only group were 50.2 and 31.7, respectively. The corresponding means in the study-test group were 54.5 and 39.8. Neither the main effect of study condition nor the interaction of study condition and the recall variable was statistically significant [$F(1,20)$ = .39 and .10, respectively]. The effect of the recall variable, however, was statistically significant [$F(1,20) = 7.24$, $MS_e = 417.4$]. These results indicate that the subjects were able to discriminate recalled and nonrecalled items and that the study-only and study-test groups did not differ in JOL accuracy for once-presented items.

Because of the higher recall levels for repeated items, more subjects were able to be included in the analysis of JOL accuracy for repeated items. The number of subjects included and the means and standard deviations for JOL ratings for recalled and nonrecalled items in the two study conditions are presented in Table 2. There were no differences between the JOL ratings for 5-5-5 and 1-5-9 items, so the ratings for repeated items were combined across these categories.

The most critical finding in the overall analysis of the data presented in Table 2 was the statistically significant interaction of study condition with the recall variable [$F(1,43) = 26.12$, $MS_e = 211.6$]. The difference between mean JOL ratings of recalled and nonrecalled items varied with the study condition. There was also a statistically significant main effect of the recall variable [$F(1,43) = 99.15$, $MS_e = 211.6$], but the main effect of study condition was not statistically significant [$F(1,43) = .66$]. The

Table 2
Mean Judgments of Learning for Recalled and Nonrecalled Repeated Items

	Recall Status			
	Recalled		Nonrecalled	
Condition	*M*	*SD*	*M*	*SD*
Study only*	65.3	17.0	50.4	17.2
Study-Test†	77.5	19.8	31.3	14.7

Note—Judgment-of-learning ratings were made on a 6-point scale with 0, 20, and 40 representing predictions of nonrecall and 60, 80, and 100 representing predictions of recall. *n = 25. †n = 23.

simple main effect for the recall variable was significant for both the study-only group [$F(1,43) = 11.80$] and for the study-test group [$F(1,43) = 113.43$]. The larger JOL mean difference in the study-test group than in the study-only group (with little or no difference in standard deviations across conditions) indicates that JOL accuracy is higher in the study-test group. An analysis of the relative sizes of the two simple main effects yielded $r^2 = .22$ for the study-only effect and $r^2 = .73$ for the study-test effect. The means in Table 2 show that the critical interaction results from both an increase in JOL ratings for recalled items and a decrease in JOL ratings for nonrecalled items.

DISCUSSION

The present results confirm that retrieval practice does improve JOL accuracy even when degree of learning has been controlled. In describing a theoretical framework for metamemory, Nelson and Narens (1990) argued that a feeling of knowing (FOK)

> does not reflect any monitoring of unconscious processes at all. Put another way, the FOK does not directly monitor a given unrecalled item in memory, but rather the FOK monitors recallable aspects related to that item, such as the item's acquisition history or partial/related recalled components. (p. 29)

It seems reasonable to extend their described process to JOL accuracy (see King et al., 1980). Success or failure on a test trial is likely to be a salient component of an item's acquisition history, and we know from research on memory for remembered events (Gardiner & Klee, 1976) that it is a memorable component. The subjects in the study-test group raised their JOL ratings for recalled items *and* lowered their JOL ratings for nonrecalled items relative to the ratings of repeated items made by the study-only subjects. Such rating changes in the study-test group are consistent with the idea that retrieval practice helps subjects to be more appropriately encouraged about the likelihood of recalling the items that will end up being recalled and more appropriately discouraged about the likelihood of recalling the items that will not end up being recalled.

Retrieval practice is not the only way to enhance memory-monitoring accuracy. Nelson and Dunlosky (1991) reported a situation in which JOL accuracy was remarkably high. This accuracy was achieved when the JOL rating was made after a filled delay following presentation of the to-be-remembered item. Memory-monitoring accuracy has also been enhanced in situations involving comprehension of text (Maki, Foley, Kajer, Thompson, & Willert, 1990). Calibration of comprehension was enhanced when the amount of processing required was increased by having subjects read text with deleted letters.

Tulving and Madigan (1970) judged that memory monitoring was "rather accurate." The present research leads us to state with a high degree of confidence that memory-monitoring accuracy can be improved through retrieval practice.

REFERENCES

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

Begg, I., Duft, S., LaLonde, P., Melnick, R., & Sanvito, J. (1989). Memory predictions are based on ease of processing. *Journal of Memory & Language*, **28**, 610-632.

Cohen, R. L. (1988). Metamemory for words and enacted instructions: Predicting which items will be recalled. *Memory & Cognition*, **16**, 452-460.

Gardiner, J. M., & Klee, H. (1976). Memory for remembered events: An assessment of output monitoring in free recall. *Journal of Verbal Learning & Verbal Behavior*, **15**, 227-233.

Glenberg, A. M., Wilkinson, A. C., & Epstein, W. (1982). The illusion of knowing: Failure in the self-assessment of comprehension. *Memory & Cognition*, **10**, 597-602.

Groninger, L. M. (1979). Predicting recall: The "feeling-that-I-will-know" phenomenon. *American Journal of Psychology*, **92**, 45-58.

Hart, J. T. (1965). Memory and the feeling-of-knowing experience. *Journal of Educational Psychology*, **56**, 208-216.

King, J. F., Zechmeister, E. B., & Shaughnessy, J. J. (1980). Judgments of knowing: The influence of retrieval practice. *American Journal of Psychology*, **93**, 329-343.

Landauer, T. K., & Bjork, R. A. (1978). Optimum rehearsal patterns and name learning. In M. M. Gruneberg & R. N. Sykes (Eds.), *Practical aspects of memory* (pp. 625-632). New York: Academic Press.

Leonesio, R. J., & Nelson, T. O. (1990). Do different metamemory judgments tap the same underlying aspects of memory? *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **16**, 464-470.

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

Maki, R. H., Foley, J. M., Kajer, W. K., Thompson, R. C., & Willert, M. G. (1990). Increased processing enhances calibration of comprehension. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **16**, 609-616.

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin*, **95**, 109-133.

Nelson, T. O., & Dunlosky, J. (1991). When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: The delayed JOL effect. *Psychological Science*, **2**, 267-270.

Nelson, T. O., & Leonesio, R. J. (1988). Allocation of self-paced study time and the "labor-in-vain effect." *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **14**, 676-686.

Nelson, T. O., Leonesio, R. J., Shimamura, A. P., Landwehr, R. F., & Narens, L. (1982). Overlearning and the feeling of knowing. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **8**, 279-288.

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and some new findings. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 26, pp. 125-173). San Diego: Academic Press.

Shaughnessy, J. J. (1981). Memory monitoring accuracy and modification of rehearsal strategies. *Journal of Verbal Learning & Verbal Behavior*, **20**, 216-230.

Tulving, E., & Madigan, S. A. (1970). Memory and verbal learning. *Annual Review of Psychology*, **21**, 437-484.

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

Zechmeister, E. B., & Shaughnessy, J. J. (1980). When you know that you know and when you think that you know but you don't. *Bulletin of the Psychonomic Society*, **15**, 41-44.

NOTE

[1] Nelson (1984) has recommended that the Goodman-Kruskal gamma be used to assess judgment-of-learning accuracy. In our experiment, we had only five observations for each type of critical item for each subject. When we tried to use gamma as a measure of accuracy, there were only 8 subjects in the study-only group and 3 in the study-test group who had values of gamma for both once-presented items and repeated items. This resulted primarily from the large number of zero entries for once-presented items. We could include more subjects in an analysis of repeated items only. The mean gamma scores for repeated items did show more accuracy in the study-test group (.78) than in the study-only group (.63), but this difference was not statistically significant. We also chose to use the mean-difference measure of JOL accuracy because it allowed us to do further analyses to explore the source of the increased accuracy in the study-test group.

(Manuscript received September 20, 1991.)