

## Similarity between the Cue for Judgments of Learning (JOL) and the Cue for Test Is Not the Primary Determinant of JOL Accuracy

JOHN DUNLOSKY

*University of North Carolina at Greensboro*

AND

THOMAS O. NELSON

*University of Maryland at College Park*

Eventual memory performance is predicted more accurately when a person's judgment of learning (JOL) is delayed until shortly after studying an item than when made immediately after studying the item. According to the transfer-appropriate-monitoring hypothesis, this delayed-JOL effect arises because of the contextual similarity between the cue for the JOL and the cue for the memory test. In three paired-associate learning experiments, delayed JOLs were cued by the stimulus alone or by the stimulus-response pair, and the eventual test was associative recognition of stimulus-response pairs. Recognition of stimulus-response pairs was predicted more accurately when JOLs had been cued by the stimulus alone than when they had been cued by the stimulus-response pair, even though the latter was more similar than the former to the cue for the recognition test. Implications of these results, especially the lack of support for the class of theories emphasizing transfer-appropriate monitoring, are discussed for theories of the accuracy of JOLs. © 1997 Academic Press

The focus of the present research is on the accuracy of people's metacognitive monitoring, a topic that has recently received much interest in relation to the mechanisms that underlie metacognitive monitoring (e.g., Metcalfe & Shimamura, 1994; Nelson, 1992). We focused on judgments of learning (JOLs), which are people's predictions of the likelihood of eventual memory performance for recently studied items. Beginning with the seminal investigation of JOLs (Arbuckle & Cuddy,

1969), most investigations of JOLs either have demonstrated that the accuracy of JOLs is above chance and/or have evaluated factors that people may base JOLs on. However, few investigations have empirically evaluated explanations for why people's JOLs are accurate at predicting eventual memory performance, perhaps because until recently no manipulation had been demonstrated to substantially modulate JOL accuracy.

Recently, the interval between the study and JOL of an item has been shown to have an extraordinary effect on the accuracy of JOLs: Accuracy at predicting eventual paired-associate recall is relatively low when the interval between the study and JOL of an item is minimal (called an *immediate* JOL), whereas accuracy is much greater when a short interval (e.g., 30 s) occurs between the study and the JOL of an item (*delayed* JOL) (Nelson & Dunlosky, 1991). This delayed-JOL effect is robust across individuals (Nelson & Dunlosky, 1991) and also occurs (a) under various experimenter-controlled study activities (Dun-

This research was supported by Grant R01-MH32205 and a Research Scientist Award (K05-MH1075) from the National Institute of Mental Health to the second author. During the writing of the article, the first author was partially supported by a Research in Cognitive Aging Grant funded by PHS/NIH National Institute on Aging (5 T32 AG00175-07) to the Georgia Institute of Technology. We thank Lisa Tabor Connor for comments on a preliminary draft. Address reprint requests to Dr. John Dunlosky, 296 Eberhart Bldg, Department of Psychology, University of North Carolina at Greensboro, Greensboro, NC 27412, or to Dr. Thomas O. Nelson, Psychology Department, University of Maryland, College Park, MD 20742.

losky & Nelson, 1994), (b) when people predict eventual recognition performance (Thiede & Dunlosky, 1994), and (c) for young adults and older adults (Connor, Dunlosky, & Hertzog, 1994). Although several hypotheses have been proposed to explain JOL accuracy (e.g., Begg, Duft, Lalonde, Melnick, & Sanvito, 1989; Dunlosky & Nelson, 1992; Nelson & Dunlosky, 1991; Spellman & Bjork, 1992), none of them provides a complete account of the delayed-JOL effect (Nelson & Dunlosky, 1992).

Hypotheses for the delayed-JOL effect are constrained by a boundary condition described by Dunlosky and Nelson (1992). Namely, when JOLs are cued by the stimulus alone (e.g., “dog-table”) is presented during study, and the JOL is cued by “dog-?”, and retention is tested by “dog-?”), a potent delayed-JOL effect occurred, whereas when JOLs are cued by the stimulus-response pair (e.g., the cue is “dog-table”), no reliable effect occurred, with accuracy being no better for delayed JOLs than immediate JOLs. A critical question for theory is, Why is accuracy greater for delayed stimulus-alone JOLs than for either of the two kinds of immediate JOLs or for delayed JOLs cued by the stimulus-response pair?

The transfer-appropriate-monitoring (TAM) hypothesis we tested here is that people’s accuracy at predicting eventual memory performance increases as the similarity increases between the context of metacognitive monitoring and the context of the eventual memory test. This hypothesis is consistent with more process-oriented versions of TAM because similar contexts (as compared to less similar contexts) are generally thought to enhance performance (in this case, predictive accuracy) because they will increase the likelihood of matches among the underlying cognitive processes at the time of JOL versus at the time of the eventual criterion test. According to the TAM hypothesis, accuracy at predicting eventual paired-associate recall is greater for delayed JOLs cued by the stimulus alone than for the three other kinds of JOLs, because of the aforementioned four kinds of context for a JOL cue: The context of delayed JOLs cued by the stimulus alone is most similar to the

context of a paired-associate recall test. That is, both a paired-associate recall test and a delayed JOL cued by the stimulus alone occur well after study (unlike immediate JOLs) and both are cued by the stimulus alone.

Because this TAM hypothesis is based on the match between the cue for the JOL and the cue for the test, it provides a testable prediction about a new situation, namely, the effect of various kinds of cue for JOLs on the accuracy of people’s JOLs at predicting *associative recognition of stimulus-response pairs*. This retention test requires the subject to select a previously studied item (e.g., “dog-table”) from distractors in which the same stimulus (i.e., “dog”) is presented with responses that had been originally paired with other stimuli (Clark & Shiffrin, 1992). The context of an associative recognition test of stimulus-response pairs is more similar to delayed JOLs cued by the stimulus-response pair than to delayed JOLs cued by the stimulus alone. This is because both the recognition test and delayed JOLs cued by the stimulus-response pair are cued by the stimulus and response (in contrast to delayed JOLs cued by the stimulus alone). Therefore, the prediction from TAM is that accuracy at predicting eventual associative recognition will be greater when delayed JOLs are cued by the stimulus-response pair than when they are cued by only the stimulus.

To evaluate this prediction empirically, delayed JOLs were cued either by the stimulus alone or by the stimulus-response pair, and the subsequent retention test was associative recognition. A secondary goal was to include immediate JOLs to evaluate whether a delayed-JOL effect occurs when people predict associative recognition by making JOLs cued by the stimulus-response pair. This topic has not been explored previously.

## EXPERIMENT 1

### *Method*

#### *Materials*

Items were 66 concrete ( $C > 6.08$ ; norms from Paivio, Yuille, & Madigan, 1968), unrelated, noun-noun paired-associates (e.g., dog-

table). Apple II computers displayed instructions and items and recorded all responses.

### *Design, Subjects, and Task*

The kind of cue for JOLs (either the stimulus alone or the stimulus-response pair) was a between-subjects manipulation, and the interval between the study and JOL for a given item (immediate or delayed JOL, operationalized below) was a within-subjects manipulation.

One hundred undergraduates participated individually for extra credit. Fifty subjects were randomly assigned to each group by order of appearance. The task included one paired-associate study trial. The items were presented in a random order for 6 s/item. Subjects were instructed to try hard to learn the items and were instructed that their "task is to learn to recognize the second word of a given pair when prompted with the first word of that pair."

A subject-paced JOL was made for every item and was cued either by the stimulus alone or by the stimulus-response pair along with the query, "How confident are you that in about ten minutes from now you will be able to recognize the second word of the item when prompted with the first? (0 = definitely won't recognize, 20 = 20% sure, 40 . . . , 60 . . . , 80 . . . , and 100 = definitely will recognize)."

### *Procedure*

*List construction.* The first six items served as practice for study and were not included on the recognition test. The remaining items comprised two blocks of 30 items/block. Each item was randomly slated for an immediate or delayed JOL with the only restrictions being that 15 items in each block were slated for each kind of JOL and no more than 3 consecutive items were slated for the same kind of JOL.

Each immediate JOL occurred immediately after the offset of the study presentation for an item. The delayed JOLs of a given block occurred as follows: After the final study presentation or final immediate JOL of the block, the JOLs for the first third of the items pre-

sented during study that were slated for delayed JOLs occurred in random order, then the JOLs for the second third occurred in random order, and so on. This ensured that there were at least 10 other items between the study and JOL of every item slated for delayed JOLs.

*Recognition test.* Following the final delayed JOL of the second block there was an unrelated filler activity for 10 min (pilot data had shown that a 10-min retention interval and a 6-s/item presentation rate led to intermediate levels of recognition performance). Following the filler activity, the 12-AFC recognition test occurred for each of the 60 critical items. The order of items for recognition trials was determined as follows: The first 10 items presented during the study trial were randomized anew and presented for recognition, then the second 10 items presented during the study trial were randomized anew, and so on.

For each recognition trial, the correct response of an item and eleven foils (randomly chosen from the responses of other items that had been presented during study) were presented below the stimulus of the item. Thus, responses were repeated over trials, and the specific responses that were repeated was a random variable. The 12 alternatives were randomly ordered, and the subjects indicated the alternative that they believed had been paired with that stimulus during study. Omissions were not allowed, and recognition trials were subject-paced.

## *Results and Discussion*

### *Effects on Recognition Performance*

Although analysis of recognition performance was not central to the goals of the present research, recognition performance was analyzed because it is involved in analyzing JOL accuracy. For each subject, the proportion of stimulus-response pairs that were correctly chosen on the 12-AFC recognition test was calculated separately for items that had received immediate JOLs and for items that had received delayed JOLs. Means across individual's scores are reported in Table 1.

A mixed-design  $2 \times 2$  analysis of variance (ANOVA) was conducted to assess the effects

TABLE 1  
PERFORMANCE ON THE TESTS OF ASSOCIATIVE RECOGNITION

Condition	Experiment		
	1	2	3
Delayed JOLs			
Stimulus-alone cue	.52 (.03)	.57 (.04)	+ .81 (.02)
Stimulus-response cue	.72 (.04)	.75 (.04)	+ .91 (.02)
Immediate JOLs			
Stimulus-alone cue	.54 (.04)	— <sup>a</sup>	—
Stimulus-response cue	.60 (.04)	—	—

*Note.* Entries in the first two columns are mean proportion of correct forced-choice recognition of stimulus-response pairs (Experiments 1 and 2), and entries in the third column are mean gamma correlations that are used here as a measure of detection accuracy involving both hits and false alarms (Experiment 3). Standard errors of the mean are reported in parentheses.

<sup>a</sup> Immediate JOLs were not included in Experiments 2 and 3.

of kind of cue (between groups) and the interval between the study and JOLs (repeated measure). Main effects occurred both for JOL interval,  $F(1,98) = 18.77$ ,  $MS_e = 0.01$ , and for kind of cue,  $F(1,98) = 7.42$ ,  $MS_e = 0.12$ ,  $ps < 0.01$ . The interaction was also reliable,  $F(1,98) = 34.06$ ,  $MS_e = 0.01$ ,  $p < .01$ . The interaction occurred because performance was greater after delayed JOLs cued by the stimulus-response pair than after each of the three other kinds of JOL,  $ts > 3.50$ ,  $ps < .001$ , whereas performance did not reliably differ between these latter three conditions,  $ts < 1.60$ . Most important for empirically evaluating the TAM hypothesis, the present procedure yielded off-the-ceiling and off-the-floor test performance, which is necessary for analyzing JOL accuracy.

#### *Effects on the Accuracy of the Judgments of Learning*

*Accuracy of the relative aspects of the judgments of learning.* JOL accuracy for predicting the likelihood of recognizing one item versus another (henceforth called “relative accuracy”) was operationalized as a Goodman–Kruskal gamma correlation between JOLs and eventual performance on the 12-AFC recognition test. For each subject, one gamma was calculated for items that had immediate JOLs, and another was calculated for items that had

delayed JOLs. (Eleven subjects were excluded from this analysis because of indeterminate gammas, which occurred because of perfect recognition performance on every item—i.e., no variance in recognition performance across the items—for 10 subjects and no variance in JOLs for one subject.) For each condition, the mean across individual subjects’ gammas is reported in the first column of Table 2.

A mixed-design  $2 \times 2$  analysis of variance (ANOVA) was conducted as in the analysis of recognition performance. A main effect occurred for JOL interval,  $F(1,87) = 17.56$ ,  $MS_e = .183$ ,  $p < .001$ . As is evident from inspection of Table 2, a delayed-JOL effect occurred both when JOLs were cued by the stimulus alone (replicating Thiede & Dunlosky, 1994) and when JOLs were cued by the stimulus-response pair. The latter result is in contrast with the negligible delayed-JOL effect for JOLs cued by the stimulus-response pair when the criterion test is paired-associate recall (Connor et al., 1994; Dunlosky & Nelson, 1992).

A main effect did not occur for the kind of cue,  $F(1,87) = 0.08$ ,  $MS_e = .174$ ,  $p = .78$ , and the interaction was not reliable,  $F(1,87) = 1.83$ ,  $MS_e = .183$ ,  $p = .24$ . As suggested by these results, the mean gammas were not reliably different for delayed JOLs cued by the stimulus alone versus for delayed JOLs

TABLE 2

ACCURACY OF JUDGMENTS OF LEARNING (JOLs) AT PREDICTING EVENTUAL ASSOCIATIVE RECOGNITION FOR JOLs CUED BY THE STIMULUS ALONE VERSUS BY THE STIMULUS-RESPONSE PAIR

Condition	Experiment		
	1	2	3
Delayed JOLs			
Stimulus-alone cue	+ .70 (.05)	+ .69 (.04)	+ .63 (.03)
Stimulus-response cue	+ .60 (.07)	+ .38 (.04)	+ .51 (.04)
Immediate JOLs			
Stimulus-alone cue	+ .34 (.07)	— <sup>a</sup>	—
Stimulus-response cue	+ .35 (.07)	—	—

Note. Mean Goodman–Kruskal gamma correlations between JOLs and eventual recognition of the stimulus-response pairs, with standard errors of the mean in parentheses.

<sup>a</sup> Immediate JOLs were not included in Experiments 2 and 3.

cued by the stimulus-response pair,  $t(89) = 1.17, p = .25$ . However, even though this critical difference did not reach statistical reliability, the trend was in the opposite direction of that predicted by TAM, and therefore we reexamined this topic in Experiment 2.

*Accuracy of the absolute aspects of the judgments of learning.* Another way to evaluate the prediction from TAM is to evaluate the accuracy of the absolute aspects of the judgments of learning (henceforth called “absolute accuracy”). In contrast to the relative accuracy of JOLs at predicting the likelihood of eventual memory performance for one item relative to another, absolute accuracy is the degree to which people accurately predict the magnitude of eventual memory performance for a given subset of items to which they gave the same predicted value (Nelson, 1996). We compared the predicted percentage of correct recognition to the actual percentage of correct responses on the recognition test, with this comparison being made separately for every subset of items that had received a given JOL rating.

Absolute accuracy over all subsets of items is presented via calibration curves: For each subject, items receiving the same predicted percentage of correct recognition were aggregated, then the actual percentage of correct responses on the recognition test was calculated for that subset of items. Finally, the mean

(across subjects) actual percentage of correct responses on the recognition test was plotted as a function of the predicted percentage of correct recognition. To evaluate the prediction from TAM, we constructed a separate calibration curve for delayed JOLs cued by the stimulus alone and for delayed JOLs cued by the stimulus-response pair.<sup>1</sup> These are shown in the upper panel of Fig. 1.

Subjects showed underconfidence as indicated by the high levels of recognition performance for items given low JOL ratings. Such underconfidence is a typical finding for people’s predictions of recognition performance (e.g., Groninger, 1979; Thiede & Dunlosky, 1994) and indicates that subjects do not adjust their JOLs to account for the likelihood of choosing the correct answer by guessing (Thiede & Dunlosky, 1994). More relevant to the questions investigated in the present

<sup>1</sup> The absolute accuracy of immediate JOLs is not relevant to the prediction from TAM and comparisons between immediate JOLs versus delayed JOLs on this measure were inconclusive. Nevertheless, to provide a complete description of absolute accuracy, we included the values for immediate JOLs. For immediate JOLs cued by the stimulus alone, actual percentages of correct responses were 37, 45, 55, 66, 78, and 75 (for JOL ratings of 0, 20, 40, 60, 80, and 100, respectively). For immediate JOLs cued by the stimulus-response pair, actual percentage of correct responses was 46, 60, 63, 69, 79, and 88 (for ratings of 0, 20, etc.).

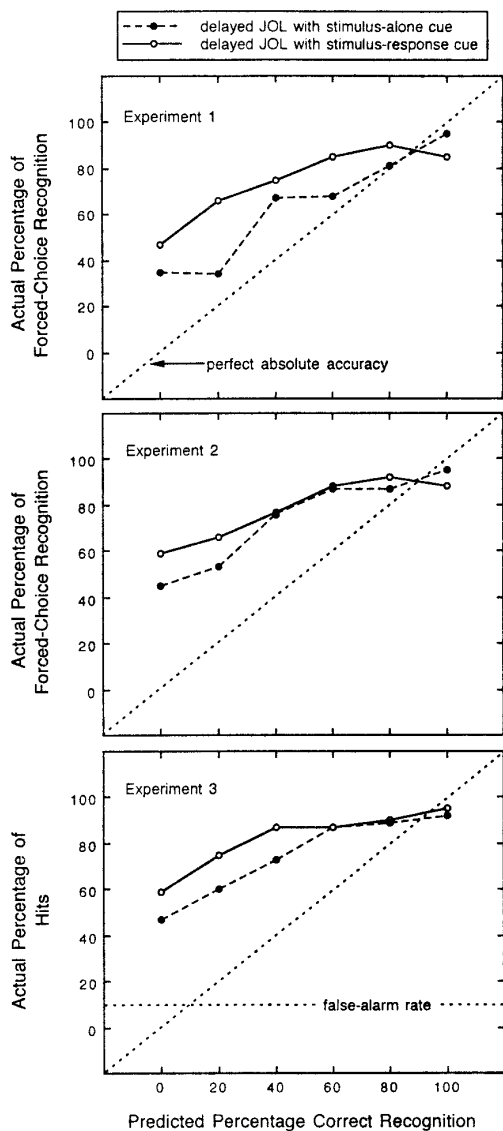


FIG. 1. Calibration curves for delayed JOLs in Experiment 1 (upper panel), Experiment 2 (middle panel), and Experiment 3 (lower panel). The main diagonal represents perfect absolute accuracy of predictions.

research, for all six values of JOLs (0%, 20%, etc.), the closeness of the predicted percentage of correct recognition to the actual percentage of correct recognition was greater for delayed JOLs cued by the stimulus alone than for delayed JOLs cued by the stimulus-response pair. Put differently, absolute accuracy was reliably greater when delayed JOLs were cued by the stimulus alone than when they were

cued by the stimulus-response pair,  $p = .03$ , sign test. These results reliably disconfirm TAM.

### Effects on the Judgments of Learning

*Magnitude of the judgments of learning.* For each subject, a median JOL was calculated separately for items that had immediate JOLs and for items that had delayed JOLs. The mean across individual's scores are reported in Table 3.

A mixed-design  $2 \times 2$  analysis of variance (ANOVA) was conducted as in the analysis of recognition performance. A main effect did not occur either for JOL interval,  $F(1,98) = 2.24$ ,  $MS_e = 183.1$ ,  $p = .14$ , or for the kind of cue,  $F(1,98) = 1.07$ ,  $MS_e = 927.6$ ,  $p = .30$ . The interaction was reliable,  $F(1,98) = 18.1$ ,  $p < .001$ . Follow-up  $t$  tests revealed that the magnitude of JOLs was not reliably different for immediate JOLs that were cued by the stimulus alone than for those that were cued by the stimulus-response pair,  $t(98) = 0.89$ . By contrast, the magnitude of delayed JOLs was reliably less when JOLs were cued by the stimulus alone than when they were cued by the stimulus-response pair,  $t(98) = 2.42$ ,  $p < .05$ . Implications of these findings are considered under General Discussion.

*Proportion of items that received each judgment-of-learning rating.* A finer-grained analysis of how people used the rating scale involves examining the proportion of items that had received each JOL rating. For each subject, we calculated the proportion of items that had received each JOL rating, and the means across individual subjects' proportions are shown in the upper panel of Fig. 2.

As evident from inspection of Fig. 2, two patterns of distributions occurred. For delayed JOLs cued by the stimulus alone, subjects used the extreme ratings more frequently than middle ratings. By contrast, for both kinds of immediate JOLs and for delayed JOLs cued by the stimulus-response pair, subjects made fewer extreme ratings than middle ratings. These patterns were confirmed by inferential tests (as in Dunlosky & Nelson, 1994). For each subject, we separately calculated the proportion of items that had received the two extreme JOL ratings (JOLs of 0

TABLE 3

THE MAGNITUDE OF JUDGMENTS OF LEARNING FOR JOLs CUED BY THE STIMULUS ALONE VERSUS BY THE STIMULUS-RESPONSE PAIR

Condition	Experiment		
	1	2	3
Delayed JOLs			
Stimulus-alone cue	27 (4.1)	30 (5.0)	31 (2.7)
Stimulus-response cue	39 (3.3)	44 (5.5)	38 (2.3)
Immediate JOLs			
Stimulus-alone cue	38 (3.0)	— <sup>a</sup>	—
Stimulus-response cue	34 (2.9)	—	—

*Note.* Mean (across subjects) of the individual subjects' median judgments of learning (predicted percentage likelihood of correctly remembering the items), with standard errors of the mean in parentheses.

<sup>a</sup> Immediate JOLs were not included in Experiments 2 and 3.

or 100) and the proportion of items that had received the two middle JOL ratings (JOLs of 40 or 60). For delayed JOLs cued by the stimulus alone, 38 of 50 subjects made more extreme than middle ratings,  $p < .001$ , sign test. By contrast, for immediate JOLs cued by the stimulus alone, 37 of 50 subjects made more middle than extreme ratings,  $p < .001$ ; for delayed JOLs cued by the stimulus-response pair, 34 of 50 subjects made more middle than extreme ratings,  $p = .01$ ; and for immediate JOLs cued by the stimulus-response pair, 34 of 50 subjects made more middle than extreme ratings,  $p < .01$ . That these latter three conditions produced the same distribution of JOLs is perhaps more evident when one considers that the 95% confidence interval for the means is narrow (about  $\pm .05$  for each mean) and includes the means from all three of these conditions at each rating.

TAM was not developed to account for the effects of different variables (such as the kind of cue for JOLs) on how people make JOLs, and hence it does not account for these findings. A hypothesis that can account for the effect of the kind of JOL cue on how people distribute JOLs across items is described under General Discussion.

## EXPERIMENT 2

The methodology from Experiment 1 was modified to be more favorable to TAM. Two aspects of the recognition test used in Experiment

1 may have made the context of the recognition test somewhat similar to the context of a recall test. First, although the 12-AFC recognition test in Experiment 1 was helpful for minimizing the role of correct guessing, when the number of alternatives on the recognition test becomes too large, recognition becomes similar to recall (cf. Davis, Sutherland, & Judd, 1961). Therefore, in Experiment 2 the number of alternatives was decreased from 12 to 7. Second, whereas in Experiment 1 the recognition test was cued by a stimulus alone (e.g., dog-?) presented above a list of alternatives (e.g., chicken, table . . .), in Experiment 2 the stimulus was presented beside every alternative (e.g., dog-chicken, dog-table . . .). This format made the context of the recognition test even more similar to the context of JOLs cued by the stimulus-response pair than to the context of JOLs cued by the stimulus alone.

Finally, more emphasis in the research design was devoted to the critical comparison between the delayed stimulus-alone JOLs versus the delayed stimulus-response JOLs. This was accomplished both by investigating only delayed JOLs, so as to have more relevant observations per subject, and by manipulating the kind of cue for JOLs as a within-subject variable.

## Method

### Materials

The 74 items were from the same pool as in Experiment 1. Apple II computers displayed

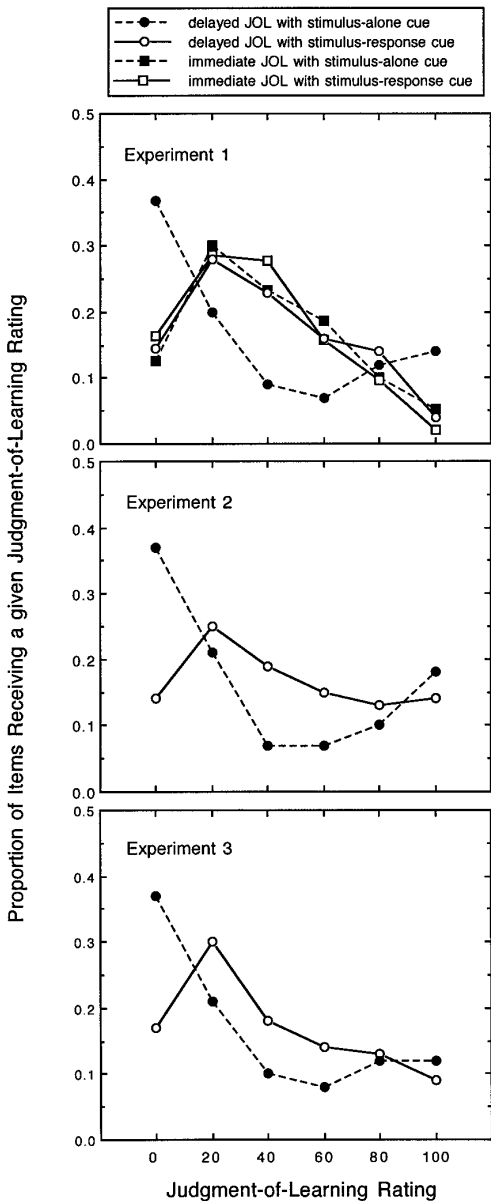


FIG. 2. The mean (across subjects) of the proportion of items that had received a given JOL rating in Experiment 1 (upper panel), Experiment 2 (middle panel), and Experiment 3 (lower panel), showing that people's use of the rating scale is qualitatively different for delayed JOLs cued by the stimulus alone versus for delayed JOLs cued by the stimulus-response pair.

instructions and items and recorded all responses.

#### Design, Subjects, and Task

The kind of cue (either the stimulus alone or stimulus-response pair) was a within-subject

manipulation. Thirty-nine individually tested undergraduates participated for extra credit. During the study trial, items were presented in a random order for 4 s/item. The procedure for cuing JOLs was the same as in Experiment 1. Besides instructing subjects to learn the items for an eventual test of recognition, subjects were shown an example: "If you originally studied dog-table, you will see dog-table mixed with other pairs (such as, dog-television and dog-crayon). Your job is to recognize the original pair, that is, dog-table."

#### Procedure

*List construction.* The first six items served as practice for study and were not included on the recognition test. The order of the kind of cue for JOLs was random except for the restrictions that over the entire list an equal number of each kind of cue occurred and no more than two consecutive trials could have the same kind of cue.

The presentation of eight other items intervened between the study and JOL for every item, thereby ensuring at least a 32-s interval between the study and JOL for a given item. The last eight items were not included on the recognition test and served only to ensure the eight-item interval between the study and JOL for every item that was included on the recognition test.

*Recognition test.* Following the final delayed JOL, there was an unrelated filler activity for 10 min, followed by 7-AFC recognition (pilot data had shown that a 10-min retention interval and 4-s/item presentation rate led to intermediate levels of recognition performance). For each recognition trial, the response of an item and six foils (chosen randomly from other responses of items presented during study) were each paired with the stimulus of the item (e.g., dog-apple, dog-table, dog-chicken . . .). As in Experiment 1, responses were repeated as foils and the specific responses that were repeated was a random variable. The pairs were randomly ordered, and subjects were asked to choose the pair that had been presented during study. The order of the items was randomized anew for recognition.

## Results and Discussion

### Effects on Recognition Performance

For each subject, the proportion of stimulus-response pairs that were correctly chosen on the 7-AFC recognition test was calculated separately for items that had received stimulus-response JOLs and for items that had received stimulus-alone JOLs. Means across individual's scores are reported in the second column of Table 1. As in Experiment 1, performance was greater for delayed JOLs that were cued by the stimulus-response pair than for those cued by the stimulus alone,  $t(38) = 7.75, p < .001$ .

### Effects on the Accuracy of the Judgments of Learning

*Relative accuracy of the judgments of learning.* For each subject, one gamma was calculated for items that had JOLs cued by the stimulus alone, and another was calculated for items that had JOLs cued by the stimulus-response pair. (Six subjects were excluded because of indeterminate gammas, which occurred because of perfect recognition performance for five subjects and no variance in JOLs for one subject.) The mean gammas are reported in the second column of Table 2. Results were reliably opposite to the prediction from the TAM hypothesis: Accuracy was reliably greater when delayed JOLs were cued by the stimulus alone than by the stimulus-response pair,  $t(32) = 2.54, p = .016$ .

*Absolute accuracy of the judgments of learning.* The absolute accuracy of JOLs was examined as in Experiment 1. A calibration curve for delayed JOLs cued by the stimulus alone and one for delayed JOLs cued by the stimulus-response pair are shown in the middle panel of Fig. 1. Absolute accuracy was reliably greater for delayed JOLs cued by the stimulus alone than for delayed JOLs cued by the stimulus-response pair: At all six values of JOLs the predicted percentage of recognition was closer to the actual percentage of correct forced-choice recognition for delayed JOLs cued by the stimulus alone than for delayed JOLs cued by the stimulus-response pair,  $p = .03$ , sign test. These results disconfirm TAM.

### Effects on the Judgments of Learning

*Magnitude of the judgments of learning.* A median JOL was calculated separately for items that had delayed JOLs cued by the stimulus alone and for items that had delayed JOLs cued by the stimulus-response pair. The mean across individual's scores are reported in the second column of Table 3. As in Experiment 1, the magnitude of JOLs was reliably greater for items that had delayed JOLs cued by the stimulus-response pair than for items that had delayed JOLs cued by the stimulus alone,  $p < .01$ , sign test.

*Proportion of items that received each judgment-of-learning rating.* The proportion of items that had received each JOL rating was calculated as in Experiment 1. Values for each rating are shown in the middle panel of Fig. 2. For JOLs cued by the stimulus alone, 30 subjects made more extreme than middle ratings,  $p < .01$ . Although the shape of the distribution for JOLs cued by the stimulus-response pair appeared the same as in Experiment 1, 18 subjects made more extreme than middle ratings and the same number made vice versa,  $p = 1.0$ . This lack of a reliable difference may be due to this distribution peaking at a rating of 20, which is not part of the present analysis. That different patterns of distributions emerged for the two kinds of delayed JOLs was substantiated by a reliable crossover interaction: Subjects made more extreme ratings (at both JOLs of 0 and 100) when JOLs were cued by the stimulus alone than by the stimulus-response pair, whereas subjects made fewer middle ratings when JOLs were cued by the stimulus alone,  $ps < .05$  via sign tests.

## EXPERIMENT 3

Although the context of the 7-AFC recognition test used in Experiment 2 was more similar to the context of stimulus-response JOLs than to the context of stimulus-alone JOLs, the context of the 7-AFC recognition test was still not identical to presenting the stimulus-response pair as the cue for the JOL. To make our evaluation even more favorable to the TAM hypothesis, we made the context of the

recognition test identical to the context of JOLs cued by the stimulus-response pair, by using a yes/no associative recognition test of each stimulus-response pair.

### *Method*

#### *Materials*

The 74 items were the same as used in Experiment 2. Apple II computers displayed instructions and items and recorded all responses.

#### *Design, Subjects, and Procedure*

As in Experiment 2, the kind of cue for JOLs (stimulus alone versus stimulus-response pair) was manipulated within subjects. One hundred twenty-five undergraduates participated individually for extra class credit.

The procedure was identical to the procedure of Experiment 2 except that yes/no recognition was used instead of 7-AFC recognition. In particular, following the final delayed JOL and the unrelated filler activity, the 60 critical items were randomly ordered with 60 distractors. The distractors were constructed by re-pairing each critical stimulus presented during study with a response that was randomly chosen (without replacement) from the responses originally paired with the critical stimuli (hence yielding 120 trials with each stimulus and each response being presented exactly two times: once as a foil and once as a critical item). Items were presented individually, and subjects were instructed to type "Y" if that item had been presented during study and to type "N" if that item had not been presented during study. Before study, subjects were instructed to learn each item for an eventual test of recognition and were presented with an example of yes/no associative recognition.

### *Results and Discussion*

#### *Effects of Recognition Performance*

For each subject, a gamma correlation (involving hits and false alarms) was computed as a measure of detection accuracy (Nelson, 1986, 1990). One gamma was computed for items that had JOLs cued by the stimulus-

response pair (mean hit rate = .79) and another for items that had JOLs cued by the stimulus alone (mean hit rate = .63). The false-alarm rate was identical (.12) for both conditions because distractors were identical for both conditions. Means across individual's gammas are reported in the third column of Table 1. As in the previous experiments, performance was greater after delayed JOLs that were cued by the stimulus-response pair than after those cued by the stimulus alone,  $t(124) = 8.43$ .

#### *Effects on the Accuracy of the Judgments of Learning*

*Relative accuracy of the judgments of learning.* For each subject, one gamma was calculated for items that had JOLs cued by the stimulus alone, and another was calculated for items that had JOLs cued by the stimulus-response pair. (Twelve subjects were excluded because of indeterminate gammas, which occurred because of perfect recognition performance for nine subjects and no variance in JOLs for three subjects.) The mean gammas are reported in the third column of Table 2. These results were reliably opposite to the prediction of the TAM hypothesis: Accuracy was reliably greater for delayed JOLs cued by the stimulus alone than for delayed JOLs cued by the stimulus-response pair,  $t(112) = 2.28$ ,  $p = .025$ .

*Absolute accuracy of the judgments of learning.* To examine absolute accuracy of JOLs, we constructed a separate calibration curve for delayed JOLs cued by the stimulus alone and for delayed JOLs cued by the stimulus-response pair. These curves are shown in the lower panel of Fig. 1. Results of absolute accuracy here were inconclusive: Only four of the five non-tied comparisons favored delayed JOLs cued by the stimulus alone,  $p = .37$ , sign test.

#### *Effects on the Judgments of Learning*

*Magnitude of the judgments of learning.* A median JOL was calculated separately for items that had delayed JOLs cued by the stimulus alone and for items that had delayed JOLs cued by the stimulus-response pair. The mean

across individual's scores are reported in the third column of Table 3. The magnitude of JOLs was reliably greater for items that had delayed JOLs cued by the stimulus-response pair than for items that had delayed JOLs cued by the stimulus alone,  $p < .01$ , sign test.

*Proportion of items that received each judgment-of-learning rating.* Mean values for each JOL rating are shown in the lower panel of Fig. 2. For JOLs cued by the stimulus alone, 87 subjects made more extreme than middle ratings,  $p < .001$ . For JOLs cued by the stimulus-response pair, 51 subjects made more extreme ratings, whereas 70 subjects made more middle ratings,  $p = .10$ . The difference in the shape of distributions shown in the lower panel of Fig. 2 was substantiated by a reliable crossover interaction: Subjects made more ratings of 0 and more ratings of 100 when JOLs were cued by the stimulus alone than by the stimulus-response pair, whereas subjects made fewer middle ratings when JOLs were cued by the stimulus alone than by the stimulus-response pair,  $ps < .01$ .

#### *Meta-Analyses of Judgment-of-learning*

##### *Accuracy across the Three Experiments*

For JOL accuracy, the critical comparison for evaluating the TAM hypotheses was statistically reliable in Experiments 2 and 3 but not in Experiment 1. Therefore, a meta-analysis was conducted on JOL accuracy. The results from this comparison were combined across all three experiments via the Fisher combined test (described in Wolf, 1986, pp. 18–19). All three  $p$  values were associated with the same direction of difference. The results of the meta-analysis yielded a reliable overall conclusion: JOL accuracy at predicting associative recognition is reliably greater for delayed JOLs cued by the stimulus alone than for delayed JOLs cued by the stimulus-response pair,  $\chi^2(6) = 18.44$ .

For absolute JOL accuracy, the critical comparison between the aforementioned prediction was statistically reliable in Experiments 1 and 2 but not in Experiment 3. Because all three  $p$  values were associated with the same direction of difference, a meta-analysis was conducted via the Fisher combined

test. Across experiments, absolute JOL accuracy at predicting associative recognition was reliably greater for delayed JOLs cued by the stimulus alone than for delayed JOLs cued by the stimulus-response pair,  $\chi^2(6) = 15.98$ . Thus, results of absolute accuracy provide converging evidence for the conclusion based on relative accuracy in which TAM was disconfirmed.

## GENERAL DISCUSSION

### *Implication for the Transfer-Appropriate-Monitoring Hypothesis*

The TAM hypothesis as the primary explanation for the delayed-JOL effect was disconfirmed by the following major finding: Accuracy of predicting eventual recognition of stimulus-response pairs was greater for delayed JOLs cued by the stimulus alone than for delayed JOLs cued by the stimulus-response pair. Although other versions of TAM have been discussed in the literature (e.g., Begg et al., 1989), they cannot account for the present findings unless extra auxiliary assumptions are developed that specify the underlying processes of JOLs and the underlying processes of the retention test. We do not know what empirically testable auxiliary assumptions will be sufficient. Some speculations about other possible bases for JOLs are provided next.

### *Other Hypotheses about the Delayed-JOL Effect*

This section of the General Discussion focuses on several hypotheses about the underlying bases of JOLs that are not disconfirmed by the current data and may help provide an explanation for them.

According to the *monitoring-retrieval hypotheses*, when a person predicts the likelihood of eventual memory performance, he or she monitors information retrieved from memory about the to-be-judged item (e.g., Dunlosky & Nelson, 1994; King, Zechmeister, & Shaughnessy, 1980; Lovelace, 1984; Nelson & Dunlosky, 1991; Spellman & Bjork, 1992). A core assumption of this class of hypotheses is that JOL accuracy is a func-

tion of “the degree to which information retrieved from memory about the to-be-judged item at the time of the JOL is predictive of eventual memory performance” (Dunlosky & Nelson, 1994, p. 561). In contrast to TAM, the similarity between the context (or processes) of JOLs and of the eventual test is not necessary for yielding high levels of predictive accuracy. This does not mean that the kind of eventual test per se will have no effect on the level of JOL accuracy (e.g., see Thiede & Dunlosky, 1994, for the effects of different kinds of retention test on JOL accuracy), but instead means that even if the context of JOLs and test do not overlap, high accuracy will arise if the information retrieved at the time of JOLs is highly correlated with eventual test performance (cf. explanation of the accuracy of feeling-of-knowing judgments by Koriat, 1993, p. 614).

The assumption that people monitor on-line retrieval of responses is evident in two hypotheses that have been recently developed to explain the delayed-JOL effect. According to a hypothesis developed by Spellman and Bjork (1992), the delayed-JOL effect occurs for JOLs cued by the stimulus alone because (a) the cue for the JOL elicits a covert attempt of retrieval, and (b) when such covert retrieval is successful and delayed, it increases the likelihood of eventual memory performance.<sup>2</sup> Put differently, in contrast to retrieval that occurs immediately after an item is studied, the covert retrieval elicited by delayed JOLs presumably determines eventual retrieval performance and hence turns people’s “predictions of future performance into a self-fulfilling prophecy” (Spellman & Bjork, 1992, p. 315). As mentioned by Nelson and Dunlosky (1992), one prediction from this hypothesis (hereafter referred to as the Spellman & Bjork hypothesis) is that eventual memory performance will be greater after delayed JOLs cued

by the stimulus alone than after immediate JOLs.

In contrast to this prediction, eventual recognition performance was statistically indistinguishable after immediate JOLs and delayed stimulus-alone JOLs (left-most column of Table 1), just as there was also no difference in recall after those conditions in Nelson and Dunlosky (1992). Thiede and Dunlosky (1994, Experiment 1) also had students make stimulus-alone JOLs in anticipation of an eventual recognition test. As in the present experiment, the delayed-JOL effect occurred, and memory performance was not reliably different after immediate JOLs than after delayed JOLs. These findings provide little or no support for the Spellman & Bjork hypothesis and imply that to whatever degree this hypothesis accounts for the boost in accuracy caused by delayed stimulus-alone JOLs, that degree is insufficient to account for the large effect, and other mechanisms are needed (e.g., recognition here was equivalent after immediate vs delayed stimulus-alone JOLs, but JOL accuracy was much higher for delayed stimulus-alone JOLs than for immediate JOLs). Furthermore, because the Spellman & Bjork hypothesis was developed to account for the accuracy of JOLs cued by the stimulus alone, modifications may be required if it is to provide any explanation of how the stimulus-response cue would affect predictive accuracy.

Another instantiation of the monitoring-retrieval hypotheses is the *monitoring-dual-memories hypothesis* (Nelson & Dunlosky, 1991), which is that a person evaluates retrieval both from short-term memory and from long-term memory for information about the to-be-judged item.<sup>3</sup> Information retrieved about the to-be-judged item may include an internal representation of the stimulus of the item and the response of the item and the association between them, and perhaps how the

<sup>2</sup> Although this hypothesis was developed to account for the accuracy of people’s delayed JOLs at predicting eventual recall performance, the hypothesis also seems applicable when the criterion test is associative recognition.

<sup>3</sup> The important distinction here is between assessing shorter-lasting memories versus longer-lasting memories. Thus, the monitoring-dual-memories hypothesis can be applied equally easily to models involving distinct memory stores or to models in which short-term memory is conceived as a subset of long-term memory (Cowan, 1993).

item was processed during study (Begg et al., 1989; Mazzoni & Nelson, 1995). When people make immediate JOLs, information about the item is still in short-term memory. This information will interfere with an assessment of information about the to-be-judged item in long-term memory, because people presumably are unable to distinguish between information about the to-be-judged item that is retrieved from short-term memory versus long-term memory. Note that retrieval of information from short-term memory does not necessarily interfere with retrieval of information from long-term memory, but instead interferes with a person's ability to evaluate whether any information retrieved about the item also had been retrieved from long-term memory (for further discussion of how interference may operate here see Dunlosky & Nelson, 1994). In accord with this assumption, evidence indicates that people retrieve information more quickly from short-term memory than from long-term memory (Wescourt & Atkinson, 1973). Because delayed JOLs occur after the to-be-judged item has been forgotten from short-term memory, information about that item is not present in short-term memory at the time of the delayed JOL; instead, short-term memory will contain information about only the recently intervening items, such that any information retrieved about the target item will come from long-term memory and therefore will be diagnostic of subsequent performance based on that same information in long-term memory.

In the present experiments, information retrieved at the time of delayed stimulus-alone JOLs was highly predictive of the criterion test of associative recognition, because retrieval of a response using the association between a stimulus and that response in long-term memory presumably is the basis not only of these delayed JOLs but also is a basis of associative recognition (according to Clark, 1992; Clark, Hori, & Callan, 1993; Clark & Shiffrin, 1992). In the terminology of SAM (Gillund & Shiffrin, 1984), delayed stimulus-alone JOLs and associative recognition may both be partially determined by the inter-item strength (parameter  $b$ ) between the two words

of a pair. In contrast to cuing delayed JOLs by the stimulus alone, when the stimulus-response pair is used to cue a delayed JOL, the pair (as a cue for the JOL) presumably enters short-term memory before information can be retrieved about the item from long-term memory and thereby will interfere with an assessment of information retrieved from long-term memory about that item (ala the mechanism described above for immediate JOLs). Because associative recognition is based on information only in long-term memory, accuracy is reduced for delayed JOLs that are cued by the stimulus-response pair (relative to delayed JOLs cued by the stimulus alone).

#### *How the Cue for the JOL Affects Memory and the Distribution of JOLs*

A complete theory of people's JOLs will account not only for the accuracy of JOLs but also for how the cue for JOLs affects other aspects of performance. Accordingly, we briefly discuss findings related to how the kind of cue for delayed JOLs (stimulus-response pair vs stimulus alone) affects eventual memory performance and the distribution of JOLs across items.

All three experiments yielded the same pattern of findings concerning eventual recognition performance, which was greater when delayed JOLs were cued by the stimulus-response pair than by the stimulus alone. One explanation for the effects of JOL cues on memory performance has been developed in detail elsewhere (Dunlosky & Nelson, 1992) and is based on the assumption that the cue for JOLs provides an opportunity either (a) for an extra study trial when the stimulus-response pair is the cue or (b) for a test trial when the stimulus alone is the cue. Although both kinds of cues may affect eventual memory performance, stimulus-alone cues will have a negligible effect on responses not retrieved at the time of JOL (cf. Modigliani, 1976). Hence, when relatively few responses are elicited by presenting the stimulus alone, eventual memory performance will be better when JOLs are cued by the stimulus-response pair than by the stimulus alone (as shown in

Table 1; for additional discussion, see Dunlosky & Nelson, 1992, p. 379).

The kind of cue for JOLs also had a robust effect on the magnitude of delayed JOLs: In all three experiments, the magnitude of people's delayed JOLs was greater when cued by the stimulus-response pair than when cued by the stimulus alone (Table 3). This outcome may be related to two other phenomena.

First, consider hindsight bias (Fischhoff, 1975; for a review see Hawkins & Hastie, 1990). People read a narrative that describes an historical event (e.g., the war between British and Gurkas in the mountains of Nepal) without revealing its outcome (i.e., who won the war). Based on the narrative, subjects are asked to judge the likelihood of occurrence of possible outcomes (e.g., British victory, Gurka victory and so on). Before making these judgments, one group is told that a particular event actually occurred (e.g., British victory). Telling subjects that a given outcome had occurred increases their belief that the events described in the narrative would have yielded that outcome (Fischhoff, 1975). Similarly, cuing JOLs with both the stimulus and response (analogous to providing an outcome) also yielded greater judgments than did cuing JOLs with only the stimulus (analogous to withholding the outcome).

Many mechanisms have been developed to account for hindsight bias. Because there is "evidence for the operation of each under some conditions" (Hawkins & Hastie, 1990, p. 320), we discuss only two that seem particularly relevant for JOLs. One hypothesis is that when shown the stimulus-response cue, subjects adjust all their ratings upward such as if they believed "they knew the pairings all along" (cf. panel labeled "Hindsight Effect" of Fig. 1 from Hoch & Loewenstein, 1989). Results presented in all three panels of Fig. 1 disconfirm this possibility: The distribution of ratings for stimulus-response JOLs was not shifted to the right of the distribution of ratings for stimulus-alone JOLs as if all of the stimulus-response JOLs were merely adjusted upward. Instead, stimulus-response JOLs were aggregated in the intermediate ratings, indicat-

ing that subjects were not confident that they knew the pairings.

Another hypothesis is based on the anchoring-and-adjustment heuristic in which subjects assign 100% as the likelihood for a reported outcome and then adjust downward (Fischhoff, 1975). In the present case, subjects may anchor off of a JOL of 100% for each item and adjust ratings downward (such as to incorporate their theory of retention; Maki & Berry, 1984), which seems more consistent with the distribution of stimulus-response JOLs shown in Fig. 1. Note, however, that differences exist in methods typically used to investigate hindsight bias versus JOLs, with the former being based on retrospective judgments (whereas JOLs are prospective judgments) and occurring when subjects are told to ignore the reported outcome (which is not the case for JOLs). Accordingly, the theory of hindsight bias would likely provide at most only a partial account of how people make JOLs.

Second, consider Jacoby and Kelley's (1987) anagram experiment<sup>4</sup> in which a person is asked how difficult it will be for others to solve a given anagram (e.g., *fscar*). These judgments of difficulty are made under one of two conditions. The anagram is either presented alone (i.e., *fscar*) or is presented with the solution (i.e., *scarf fscar*). Compared to anagrams presented alone (analogous to stimulus-alone JOLs), anagrams shown with solutions (analogous to stimulus-response JOLs) were judged as being easier for others to solve and yielded judgments that were less predictive of the normative difficulty of solving the anagrams. According to Jacoby and Kelley (1987), poor predictive accuracy occurs because providing people with the solution to a problem spoils their opportunity to base judgments on the subjective experience of solving the anagram. This interpretation may be mapped into the monitoring-retrieval hypothe-

<sup>4</sup> The potential relation between JOLs and this anagram experiment was first brought to our attention by R. Bjork (Dunlosky & Nelson, 1992, p. 380), who also recently noted the possible effects of hindsight bias in people's metacognitive judgments (Bjork, 1994, p. 199).

ses, with “subjective experience” here being a more general term for the on-line retrieval of responses that presumably occurs when people are cued to make JOLs. Also, the monitoring-dual-memories hypothesis provides a specific mechanism of how subjective experience is “spoiled” by providing a solution: Presenting the stimulus-response pair as a cue for JOLs adds noise to information retrieved from long-term memory. That is, presenting the stimulus-response cue still invokes subjective experience; however, this experience will be less indicative of the degree to which information about the to-be-judged item has been stored in long-term memory.

As evident from the discussion of hindsight bias, evaluating the entire distribution of JOLs across items (Fig. 2) is also informative. When JOLs were delayed and cued by the stimulus alone, people’s judgments were polarized toward the extreme ratings. According to one theory of retrospective confidence judgments (which may have common psychological bases with delayed JOLs; for empirical evidence see Thiede & Dunlosky, 1994), extreme confidence occurs when information about a response is directly retrieved (Gigerenzer, Hoffrage, & Kleinbölting, 1991). By this account, the polarization of ratings appears consistent with the notion that people make an on-line retrieval of responses when JOLs are delayed and cued by the stimulus alone. Put differently, the polarization of ratings for delayed JOLs is in accord with a core assumption of the monitoring-retrieval hypotheses; namely, people’s monitoring of memory at the time of delayed JOLs is based on retrieval of responses.

When making delayed JOLs cued by the stimulus-response pair, people used the intermediate ratings more often than the extreme ratings, which indicates less extreme confidence in the predictions. The monitoring-dual-memories hypothesis provides an interpretation of this low confidence. When the stimulus-response pair is presented as a cue for delayed JOLs, some of the information retrieved about an item comes from short-term memory, which people cannot distinguish from information retrieved about the item from long-term memory. Thus, although

an individual will retrieve most items when these kinds of JOLs are made (i.e., from short-term memory), such retrieval will not be highly indicative of which items will be retained until the eventual test of memory and hence yields reduced confidence in predicting which items will versus will not be eventually retrieved.

In summary, the focus of the present research was to empirically evaluate the TAM hypothesis as originally developed to account for the delayed-JOL effect (Dunlosky & Nelson, 1992). Although the present findings on JOL accuracy have disconfirmed TAM, various versions of monitoring-retrieval hypotheses are compatible with these findings. Future research is needed to determine which version of those hypotheses (or other hypotheses) will yield the best overall account of the bases and accuracy of JOLs and the delayed-JOL effect.

#### 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 and Language*, **28**, 610–632.
- BJORK, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe and A. P. Shimamura (Eds.), *Metacognition: knowing about knowing* (pp. 185–205). Cambridge: MIT press.
- CLARK, S. E. (1992). Word frequency effects in associative and item recognition. *Memory & Cognition*, **20**, 231–243.
- CLARK, S. E., HORI, A., & CALLAN, D. E. (1993). Forced-choice associative recognition: Implications for global-memory models. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **19**, 871–881.
- CLARK, S. E., & SHIFFRIN, R. M. (1992). Cuing effects and associative information in recognition memory. *Memory & Cognition*, **20**, 580–598.
- CONNOR, L. T., DUNLOSKY, J., & HERTZOG, C. (1994). *Aging and metamemory: Performance-level dependence of memory predictions*. Poster presented at the 35th Annual Meeting of the Psychonomic Society, St. Louis.
- COWAN, N. (1993). Activation, attention, and short-term memory. *Memory & Cognition*, **21**, 162–167.
- DAVIS, R., SUTHERLAND, N. S., & JUDD, B. R. (1961). Information content in recognition and recall. *Journal of Experimental Psychology*, **61**, 422–429.
- DUNLOSKY, J., & NELSON, T. O. (1992). Importance of the kind of cue for judgments of learning (JOLs) and

- the delayed-JOL effect. *Memory & Cognition*, **20**, 373–380.
- DUNLOSKY, J., & NELSON, T. O. (1994). Does the sensitivity of judgments of learning (JOLs) to the effects of various study activities depend on when the JOLs occur? *Journal of Memory and Language*, **33**, 545–565.
- FISCHOFF, B. (1975). Hindsight  $\neq$  foresight: the effect of outcome knowledge on judgment under uncertainty. *Journal of Experimental Psychology: Human Perception and Performance*, **1**, 288–299.
- GIGERENZER, G., HOFFRAGE, U., & KLEINBÖLTING, H. (1991). Probabilistic mental models: a Brunswikian theory of confidence. *Psychological Review*, **98**, 506–528.
- GRONINGER, L. D. (1979). Predicting recall: The “feeling-that-I-will-know” phenomenon. *American Journal of Psychology*, **92**, 45–58.
- GILLUND, G., & SHIFFRIN, R. M. (1984). A retrieval model for both recall and recognition. *Psychological Review*, **91**, 1–67.
- HAWKINS, S. A., & HASTIE, R. (1990). Hindsight: biased judgments of past events after the outcomes are known. *Psychological Bulletin*, **107**, 311–327.
- HOCH, S. J., & LOEWENSTEIN, G. F. (1989). Outcome feedback: hindsight and information. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **15**, 605–619.
- JACOBY, L. L., & KELLEY, C. M. (1987). Unconscious influences of memory for a prior event. *Personality and Social Psychology Bulletin*, **13**, 314–336.
- 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.
- KORAT, A. (1993). How do we know that we now? The accessibility model of the feeling of knowing. *Psychological Review*, **100**, 609–639.
- LOVELACE, E. A. (1984). Metamemory: monitoring future recallability during study. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **10**, 756–766.
- MAKI, R. H., & BERRY, S. L. (1984). Metacomprehension of text material. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **10**, 663–679.
- MAZZONI, G., & NELSON, T. O. (1995). Judgments of learning are affected by the kind of encoding in ways that cannot be attributed to the level of recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **21**, 1263–1274.
- METCALFE, J., & SHIMAMURA, A. P. (1994). *Metacognition: Knowing about knowing*. Cambridge: MIT press.
- MODIGLIANI, V. (1976). Effects on a later recall by delaying initial recall. *Journal of Experimental Psychology: Human Learning & Memory*, **2**, 609–622.
- NELSON, T. O. (1986). BASIC programs for computation of the Goodman-Kruskal gamma coefficient. *Bulletin of the Psychonomic Society*, **24**, 281–283.
- NELSON, T. O. (1990). Comparable measurement scales in task-comparison experiments. *Journal of Experimental Psychology: General*, **119**, 25–29.
- NELSON, T. O. (1992). *Metacognition: Core Readings*. Boston: Allyn and Bacon.
- NELSON, T. O. (1996). Gamma is a measure of the accuracy of predicting performance on one item relative to another item, not of the absolute performance of an individual item. *Applied Cognitive Psychology*, **10**, 257–260.
- 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., & DUNLOSKY, J. (1992). How shall we explain the delayed-judgment-of-learning effect? *Psychological Science*, **3**, 317–318.
- PAIVIO, A., YUILLE, J. C., & MADIGAN, S. A. (1968). Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of Experimental Psychology Monograph*, **76**, (1, Pt. 2).
- SPELLMAN, B. A., & BJORK, R. A. (1992). When predictions create reality: Judgments of learning may alter what they are intended to assess. *Psychological Science*, **3**, 315–316.
- THIEDE, K. W., & DUNLOSKY, J. (1994). Delaying students’ metacognitive monitoring improves their accuracy in predicting their recognition performance. *Journal of Educational Psychology*, **86**, 290–302.
- WESTCOURT, K. T., & ATKINSON, R. C. (1973). Scanning for information in long- and short-term memory. *Journal of Experimental Psychology*, **98**, 95–101.
- WOLF, F. M. (1986). *Meta-analysis: quantitative methods for research synthesis*. Beverly Hills: Sage Publications.

(Received February 6, 1995)

(Revision received January 10, 1996)