Does the Sensitivity of Judgments of Learning (JOLs) to the Effects of Various Study Activities Depend on When the JOLs Occur?

JOHN DUNLOSKY AND THOMAS O. NELSON

University of Washington

Judgments of learning (JOLs) made immediately after items are studied have been shown to be insensitive to the way in which eventual memory performance is affected by (a) imagery versus rote rehearsal and (b) distributed versus massed repetitions. One explanation is that JOLs made immediately after study assess transient information that affects JOLs but is not predictive of eventual memory performance. Accordingly, we hypothesized that if the JOLs are delayed until the transient information about the to-be-judged item has dissipated, they might more accurately assess the effects of the study activities on subsequent retention. Our two experiments confirmed that hypothesis. The magnitude of delayed JOLs was greater after interactive imagery than after rote rehearsal (Experiment 1) and was greater after distributed repetitions than after massed repetitions (Experiment 2). Also, the distributions of JOLs indicated greater confidence (polarization) for delayed JOLs than for immediate JOLs, and the accuracy of predicting item-by-item retention was greater for delayed JOLs than for immediate JOLs in every condition (rote rehearsal, interactive imagery, single presentations, massed repetitions, and distributed repetitions). Thus people's timing of their JOLs is critical for several aspects of metacognition. © 1994 Academic Press, Inc.

Two major aspects of metacognition are monitoring and control (Flavell, 1979; Nelson, 1992), and the interplay between those aspects is important for theories of metacognition. Nelson and Narens (1990, see especially their Fig. 4) proposed a theory of subject-controlled study that emphasized (a) the selection of study activities, such as the kind of rehearsal strategy, and (b) judgments of learning (JOLs) in which people predict their likelihood of eventual memory performance on recently studied items. People supposedly select a study activity from their metacognitive library of activities (also called "metacognitive knowledge" in Flavell, 1979), use the study activity to learn new items, and throughout study, assess their learning via JOLs.

JOLs have been investigated by many researchers (e.g., Arbuckle & Cuddy, 1969; Bauer, Kyaw, & Kilbey, 1984; Begg, Duft, Lalonde, Melnick, & Sanvito, 1989; Dunlosky & Nelson, 1992; Groninger, 1976, 1979; King, Zechmeister, & Shaughnessy, 1980; Kroll, Jaeger, & Dornfest, 1992; Leonesio & Nelson, 1990; Lovelace, 1984; Lovelace & Marsh, 1985; Nelson & Dunlosky, 1991; Rabinowitz, Ackerman, Craik, & Hinchley, 1982; Shaughnessy, 1981; Shaughnessy & Zechmeister, 1992; Vesonder & Voss, 1985; Zechmeister & Shaughnessy, 1980). If the learner's JOLs indicate slower progress than expected, the learner may switch to another strategy. For instance, imagine a student who is exploring different study activities to learn foreign-language/English translation equivalents (e.g., *ardhi*-soil) and who makes a JOL after studying each item. If greater memory is produced by one study activity than another, but if the student's JOLs do not detect that difference, then during sub-

This research was supported by Grant R01-MH32205 and a Career Development Award (K05-MH1075) from the National Institute of Mental Health to the second author. We thank Earl Butterfield and Keith Thiede for comments on a preliminary draft. Reprint requests should be addressed either to Dr. John Dunlosky, who is now at the School of Psychology, Georgia Institute of Technology, Atlanta, GA 30332-0170, or to Dr. Thomas O. Nelson, Psychology (NI-25), University of Washington, Seattle, WA 98195.

0749-596X/94 $6.00

Copyright © 1994 by Academic Press, Inc.
All rights of reproduction in any form reserved.

sequent study trials the student may not use the more effective study activity. Accordingly, one critical determinant of the final outcome of subject-controlled study activities is the accuracy of JOLs.

Previous research has found that people's JOLs are inaccurate at predicting which study activities produce greater eventual recall. For instance, Shaughnessy (1981; Experiment 2) instructed students to study paired-associate items via either interactive imagery or rote rehearsal. Immediately after the presentation of each item, a JOL occurred (henceforth, called an *immediate* JOL). As expected from earlier research (e.g., Bower & Winzenz, 1970), Shaughnessy (1981) found that eventual recall was greater for items studied under interactive-imagery instructions than under rote-rehearsal instructions. However, he found no reliable difference in immediate JOLs about items studied under those two kinds of instructions. Rabinowitz et al. (1982) replicated that pattern of findings. Why did the instructions that produced greater eventual recall have a negligible effect on JOLs?

In another situation (Zechmeister & Shaughnessy, 1980), students studied items that were presented twice, either by distributed or massed repetitions. After the second presentation of a given item, an immediate JOL occurred. As found in earlier research (e.g., Peterson, Hillner, & Saltzman, 1962), recall was greater for items that had distributed repetitions rather than massed repetitions. However, the magnitude of JOLs was not reliably different for items that had distributed versus massed repetitions. Why did the distributed repetitions that produced greater eventual recall have a negligible effect on JOLs?

One answer to the above questions can be derived from what here is called the *monitoring-retrieval hypothesis* (cf. Nelson & Dunlosky, 1991): When a person assesses the likelihood of eventual memory performance, he or she monitors the information retrieved from memory about the to-be-judged item. However, if information retrieved about the to-be-judged item at the time of the JOL is not predictive of eventual memory performance, then the JOLs will not be predictive of the effect of various kinds of study activity on that performance.

Relevant to the above hypothesis is the finding from Smith, Baressi, and Gross (1971. Figure 1) that when the interval between the study and test for an item is minimal (e.g., an interval of less than 5 sec), recall is no better for items studied by interactive imagery than by rote rehearsal, even though recall at longer retention intervals is greater for items studied by interactive imagery than by rote rehearsal. Similarly, Peterson et al. (1962, Table 3) found that when the interval between the study and test for an item is minimal (e.g., less than 5 sec), recall is no better for items that had distributed rather than massed repetitions, even though recall at longer retention intervals is greater for items that had distributed repetitions.

Thus at the time when immediate JOLs occur, memory performance may not yet be indicative of the effects that those study activities will have on eventual memory performance. Presumably, this occurs because there is a negligible difference between the short-term memories for items studied via one versus another kind of study activity, even though there will be a substantial difference between the long-term memories for those items, and information is retrieved faster from short-term memory than from long-term memory (Wescourt & Atkinson, 1973). (The important distinction here is between assessing memories that are short-lasting—e.g., typically lasting no more than 30 sec—versus longer-lasting, and the monitoring-retrieval hypothesis can be applied equally easily to memory models that postulate a "working memory" rather than a "short-term memory.")

The monitoring-retrieval hypothesis suggests that people's JOLs might be predictive of the eventual effects of these various kinds of study activity. Given that people base their JOL on the memory status of the item at the time that the JOL occurs, then if people delay their JOL until the to-be-judged item has been forgotten from short-term memory, their delayed JOL will be predictive of the effects of the aforementioned study activities on eventual memory performance. This is because any information retrieved about the item during a delayed JOL will be from long-term memory and therefore will be predictive of eventual recall.

Another hypothesis—called the *inability hypothesis*—for why the aforementioned study activities have not affected people's JOLs was suggested by Rabinowitz et al. (1982) when they concluded, "Imagery increased participants' recall, but not their recall predictions. . . . [the participants] appear *unable* to monitor the effectiveness of actually performing these cognitive operations" (p. 694, italics added). A well-known statement by Nisbett and Wilson (1977) is consistent with an even more general notion of a deficiency in self-monitoring: "People may *have little ability* to report accurately about their cognitive processes" (p. 241, italics added). According to the inability hypothesis, people might be incapable of monitoring the effect that these specific study activities have on eventual memory performance. A related version of the hypothesis is that people have the ability to monitor these effects but are unable to use that monitoring appropriately when making JOLs (cf. Murphy, Sanders, Gabriesheski, & Schmitt, 1981). The two versions of this inability hypothesis are indistinguishable in the outcomes predicted here and therefore are considered together.

The prediction from the inability hypothesis is that regardless of the timing of the JOLs, the magnitude of JOLs will be the same (1) after interactive-imagery instructions versus after rote-rehearsal instructions and (2) after massed repetitions versus after distributed repetitions. Although the inability hypothesis can explain those previous findings, it does not hold for situations in which people overtly choose between two study activities for learning foreign-language vocabulary (Pressley, Levin, and Ghatala, 1984) or in which people are instructed to use interactive imagery versus separate imagery (Begg et al., 1989, Experiment 2). Each of those situations is considered next.

Pressley, Levin, and Ghatala (1984) had adult subjects study via the keyword mnemonic and via rote rehearsal. Although the keyword mnemonic produces greater eventual recall, prior to studying the items and even after studying without any overt test the subjects were no more likely to choose to study via the keyword mnemonic than via rote rehearsal. By contrast, subjects who had an overt delayed test after studying via both study activities were more likely to choose the keyword mnemonic over rote rehearsal.

Begg et al. (1989, Experiment 2) had people study noun-noun paired associates (e.g., ocean–tree) under either interactive-imagery instructions or separate-imagery instructions and make either immediate JOLs cued by the stimulus-response (i.e., "ocean–tree") or delayed JOLs cued by the stimulus alone (i.e., "ocean–"). Although eventual recall was greater for items studied under interactive-imagery instructions than under separate-imagery instructions, only the delayed JOLs were predictive of this effect.

The above two experiments show that the inability hypothesis has at least some restrictions in terms of the domain to which it applies, but this does not necessarily imply that the inability hypothesis cannot account for the failures of people's JOLs to assess the advantage of imagery over rote rehearsal or the advantage of distributed over massed repetitions. Different explanations might be needed for different kinds of study activities. Moreover, and particularly relevant for theory, the findings from the above two experiments are not analytic concerning the way in which people's metacognitions were affected. The Pressley et al. (1984) outcome could have been due either to having an overt test or to having a delay between study and the assessment of learning; the Begg et al. (1989) outcome could have been due either to the delay of the JOLs or to the use of stimulus-alone cues versus stimulus-response cues. The importance of the latter distinction was not established until Dunlosky and Nelson (1992) discovered that JOL accuracy is affected both by the timing of the JOLs *and* by the kind of cue for JOLs; namely, increased JOL accuracy for predicting recall performance occurs only when the JOLs are delayed *and* when the JOLs are cued by the stimulus alone. That discovery also helps to explain why the use of stimulus-response delayed JOLs by Shaughnessy (1981, Experiment 3) did not yield a reliable difference in the magnitude of JOLs for imagery versus rote rehearsal.

Accordingly, the cue for the JOLs in our experiments was always the stimulus alone, which has the maximal effect on JOL accuracy (Begg, Martin, & Needham, 1992; Dunlosky & Nelson, 1992). To have a more analytic design in which the only variation is the delay between study and the JOL, the same cue was used for immediate JOLs and delayed JOLs, and both of those kinds of JOLs were made prior to any overt test of recall. Thus the notion was that if our experiments yielded a positive result of immediate versus delayed JOLs, we would have a more precise idea of the critical factor for improving the metacognitive monitoring of the long-term effectiveness of different study activities.

Besides our primary goal of examining how the magnitude of JOLs is affected by different kinds of study activities, a secondary goal was to explore whether those study activities affect the accuracy of JOLs at predicting eventual recall on one versus another item receiving a given study activity. Begg et al. (1989, Experiment 2) found that item-by-item JOL accuracy was unaffected by the kind of study activity, and they concluded, "Within ways of studying, predictions of which items will succeed and fail are equally accurate" (p. 630). Therefore we wanted to explore whether the study activities in our research would all produce the same degree of item-by-item JOL accuracy. We wondered if there might be some study activity that affected both the overall level of recall and also the item-by-item JOL accuracy.

In addition to the above primary and secondary goals, our research had a tertiary goal. When subjects are not instructed to use a particular kind of mnemonic strategy, the accuracy of a delayed JOL is much greater than the accuracy of an immediate JOL, even when the cue is the same for both JOLs (called "the delayed-JOL effect" in Dunlosky & Nelson, 1992, and Nelson & Dunlosky, 1991). We wanted to determine whether each of the study activities we planned to investigate would also yield a delayed-JOL effect.

We investigated the above for two kinds of study activities that JOLs have previously been only insensitive to. In Experiment 1, the subjects studied under either interactive-imagery instructions or rote-rehearsal instructions. In Experiment 2, the subjects studied under either massed or distributed repetitions.

Experiment 1

Method

Materials

Items were 66 concrete ($C \geq 6.08$; norms from Paivio, Yuille, & Madigan, 1968), unrelated, noun–noun pairs. Apple II computers displayed instructions and items and recorded all responses.

Subjects, Design, and Task

Thirty-six students from the University of Washington participated individually to receive extra course credit. The interval between the study and JOL for an item (immediate or delayed) and the kind of study activity (interactive-imagery instructions versus rote-rehearsal instructions) were within-subject manipulations.

The task began with one paired-associate study trial. Presentation rate was 10 sec/item, and subjects were instructed to study the items so that they could recall the second word when prompted with the first. Two sec before and during the presentation of each item either the word "IMAGERY" or the word "REPEAT" was presented. Subjects were instructed to form an interactive image between the two words of an item when "IMAGERY" was presented. Subjects were instructed to repeat an item aloud until its offset when "REPEAT" was presented. A dummy tape recorder was switched to "record" in view of the subject to encourage compliance with the rote-rehearsal instructions.

A self-paced JOL was made for each item and was prompted with only the stimulus (e.g., if "ocean–tree" had been presented during study, the cue for the JOL would be "ocean–") and the query "How confident are you that in about ten minutes from now you will be able to recall the second word of the item when prompted with the first? (0 = definitely won't recall, 20 = 20% sure, 40. . . , 60. . . , 80. . . , and 100 = definitely will recall)."

List Construction

For each subject, the items were randomly ordered for presentation. The first six items constituted a practice list; none of these six items had a recall trial. The remaining sixty items comprised two blocks of 30 items/block.

To counterbalance the order of items receiving interactive-imagery versus rote-rehearsal instructions across subjects, we yoked subjects by order of appearance. For the first subject of a pair of yoked subjects, items were randomly assigned to a study activity with the constraint that fifteen items in the first block were presented under each kind of study activity and that fifteen items in the second block were presented under each kind of study activity. If the order for the first subject was "imagery, repeat, repeat, imagery. . .", then the order for the yoked subject was the opposite and would be "repeat, imagery, imagery, repeat. . . ."

Immediate JOLs versus delayed JOLs were randomly assigned to items except for the restrictions that (1) 15 immediate JOLs and 15 delayed JOLs occurred for each of the two study activities and (2) the same kind of JOL did not occur for more than two consecutive items that were presented under a given study activity. An immediate JOL for an item immediately followed the offset of that item. To ensure at least a 30-sec interval between the study and delayed JOL for an item, delayed JOLs occurred as follows: After the final immediate JOL or study trial of a given block, JOLs occurred for the first third of items presented for study within that block that were slated to receive delayed JOLs (order of presentation was randomized anew from study to delayed JOLs). Next, JOLs occurred for the second third of items slated to receive delayed JOLs within that block, followed by the JOLs for the last third of items.

Paired-Associate Recall

Recall trials followed the last delayed JOL. The stimulus was presented, and the subjects were asked to type the second word of the item. The recall trials were self-paced, and subjects were not permitted to omit a response. To minimize the role of spelling errors, if the first three letters of an answer were correct, it was scored as correct. The order of items was randomized anew from study to test in the following manner. The first sixth of items presented during study were randomly ordered and presented for recall, followed by the second sixth of items, followed by the third sixth of items, and so forth.

Results and Discussion

In Experiments 1 and 2, differences reported as reliable had $p < .05$.

Recall Performance

A prerequisite to evaluating the predictions about JOLs is that recall must differ for items that were presented under interactive-imagery instructions versus rote-rehearsal instructions. Accordingly, analyses of recall are reported first. For each subject, the proportion of correct recall was calculated within each kind of study activity for items that had immediate JOLs and for items that had delayed JOLs. Means across subjects within each of those four conditions are reported in the top portion of Table 1.

A 2 × 2 repeated-measures analysis of variance (ANOVA) was conducted to assess the effects of study activity and the timing of JOLs. There was a main effect of the timing of JOLs ($F(1, 35) = 14.74$, $MS_e = .01$). Although recall was reliably greater here for items that had delayed versus immediate JOLs, this effect has not been robust. In two previous experiments, recall was not greater after delayed JOLs than after immediate JOLs (Dunlosky & Nelson, 1992; Nelson & Dunlosky, 1991). The magnitude of recall has even been greater after immediate JOLs (Dunlosky & Nelson, 1992), a trend that is in the opposite direction of the present finding. Also, the magnitude of the present effect is relatively small (a difference of only .07), and so it will not be discussed further. (For hypotheses concerning how JOLs may affect eventual memory performance see Dunlosky & Nelson, 1992, and Zechmeister & Shaughnessy, 1980.)

There was a main effect of study activity ($F(1, 35) = 71.87$, $MS_e = .06$), and the interaction was not reliable ($F(1, 35) = 0.22$, $MS_e = 0.01$). The results for interactive imagery versus rote rehearsal are in accord with previous research (e.g., Smith et al., 1971): Eventual recall was greater for items studied under interactive-imagery instructions than under rote-rehearsal instructions, and this difference occurred regardless of the timing of JOLs. Thus, the prerequisite occurred that was necessary to allow a test of the hypotheses about JOLs.

Magnitude of Judgments of Learning

For each subject, a median JOL was calculated within each kind of study activity for items that had immediate JOLs and for items that had delayed JOLs. Means across the subjects' medians were calculated within each of those four conditions and are shown in Fig. 1.

Two orthogonal planned comparisons (derived from the hypothesis discussed above) were conducted. (1) So as to test the predictions empirically, the magnitude of delayed JOLs for items studied under inter-

TABLE 1
PROPORTION OF CORRECT RECALL PERFORMANCE

	Time of the JOL		
Kind of study activity	Immediate	Delayed	Overall
Experiment 1			
Interactive imagery	.55 (.05)	.63 (.04)	.59
Rote rehearsal	.22 (.04)	.28 (.04)	.25
Overall	.39	.46	
Experiment 2			
Distributed repetitions	.49 (.04)	.55 (.04)	.52
Massed repetitions	.38 (.04)	.46 (.04)	.42
Single presentation	.42 (.05)	.39 (.04)	.41
Overall	.43	.47	

Note. Main entries are means of proportion of correct recall; entries in parentheses are standard errors of the mean.

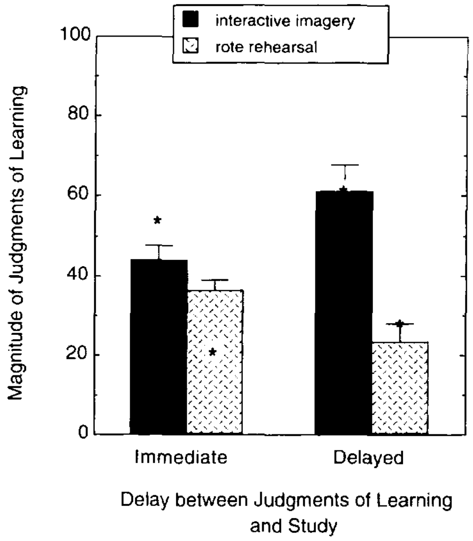

FIG. 1. Mean (across subjects) of the individual subjects' magnitude of judgments of learning, which is the median judgment-of-learning rating given by a subject within each condition in Experiment 1. The standard error of the mean (bars) and the level of mean percentage of correct recall (asterisks) are shown for each condition.

active-imagery instructions was compared to the magnitude of delayed JOLs for items studied under rote-rehearsal instructions. (2) Magnitude of immediate JOLs for items studied under interactive-imagery instructions was compared to the magnitude of immediate JOLs for items studied under rote-rehearsal instructions.

Delayed JOLs were predictive of the effect of interactive imagery versus rote rehearsal on recall: As indicated in Fig. 1, the magnitude of delayed JOLs was reliably greater for items studied under interactive-imagery instructions than under rote-rehearsal instructions (*t*(35) = 5.92). The magnitude of immediate JOLs was also reliably greater for items studied under interactive-imagery than under rote-rehearsal instructions (*t*(35) = 2.76). These findings disconfirm the hypothesis that people are unable to monitor the effects of these study activities.

A discrepancy occurred between the present research and previous research, because Rabinowitz et al. (1982) and Shaughnessy (1981) found that the magnitude of immediate JOLs did not differ reliably for items studied under these two study activities. This discrepancy may have occurred because of any of several differences in experimental designs. For instance, Rabinowitz et al. (1982) manipulated study activities between subjects and Shaughnessy (1981, Experiment 2) manipulated study activities via a blocked-list design. In the present research, study activities were manipulated within subjects via a mixed-list design. Such differences may contribute to the results, because metacognitive judgments may be less sensitive to various study activities that are manipulated via between-subjects designs or blocked-list designs than via mixed-list designs (Carroll & Nelson, 1993). Consistent with this possibility, Begg et al. (1989) found that immediate JOLs were not sensitive to the effects of separate imagery versus interactive imagery in a between-subjects design but were sensitive to those effects in a mixed-list design.

One explanation for this difference in sensitivity is that the criteria people use to judge items are more likely to shift between conditions in between-subjects designs than in within-subjects designs (Carroll & Nelson, 1993) and are more likely to shift between conditions that occur in different blocks than between conditions that are randomized across items within a list. However, even within a mixed-list design people's immediate JOLs are not always sensitive to the effects of various study activities (Zechmeister & Shaughnessy, 1980), so additional factors are probably also relevant.

Regardless of the particular explanation, it is evident in Fig. 1 that the magnitude of the difference between JOLs for interactive imagery versus rote rehearsal is greater for delayed JOLs than for immediate JOLs. This difference is meaningful rather than scale-dependent (in the sense of Townsend & Ashby, 1984). For items studied under interactive-imagery instructions, the magnitude of JOLs was reliably *greater* for delayed than immediate JOLs ($t(35) = 3.84$). By contrast, for items studied under rote-rehearsal instructions, the magnitude of JOLs was reliably *less* for delayed than immediate JOLs ($t(35) = 3.01$). Put differently, a reliable crossover interaction occurred.[1] This outcome confirms the hypothesis that the effect of these study activities on metacognitive monitoring is greater when people make delayed JOLs than immediate JOLs.

Comparison of the magnitude of judgments of learning to the level of recall. For comparing the magnitude of JOLs to the level of recall, we added asterisks to Figure 1 to indicate the mean percentage of correct recall within each condition. Within both kinds of study activity, the 95% confidence intervals for the magnitude of delayed JOLs (i.e., approximately two times the standard errors) contain the mean percentage of correct recall within each condition. By contrast, the 95% confidence intervals for the magnitude of immediate JOLs do not contain the mean percentage of correct recall within each study activity. Thus, people's delayed JOLs are not only more sensitive to the qualitative effects of the study activities but also are more sensitive to the quantitative effects of the study activities on eventual recall.

Item-by-Item Accuracy of Judgments of Learning

In contrast to the magnitude of JOLs, which is the median JOL assigned to items within a specific condition, item-by-item JOL accuracy is the degree to which a person's JOLs are predictive of his or her memory performance for one item versus another. As in previous research on JOLs (e.g., Nelson & Dunlosky, 1991), JOL accuracy was operationalized as a Goodman–Kruskal gamma correlation between JOLs and recall (for rationale see Nelson, 1984). For each subject and within each kind of study activity, one gamma was calculated between JOLs and recall for items that had immediate JOLs, and another was calculated for items that had delayed JOLs. (16 indeterminate gammas occurred: 8 for immediate JOLs and 8 for delayed JOLs.) Means across subjects within each of the four conditions are reported in the top portion of Table 2.

A 2 × 2 repeated-measures ANOVA was conducted to assess the effects of study activity and the timing of JOLs. There was a main effect of the timing of JOLs ($F(1, 23) = 67.63$, $MS_e = .15$): Accuracy was substantially greater for delayed than immediate JOLs for these experimenter-instructed study activities. Although there was not a reliable main effect of study activity ($F(1, 23) = .94$, $MS_e = .20$) and the interaction effect also was not reliable ($F(1, 23) = .59$, $MS_e = .17$), a trend occurred in which mean accuracy was lower after interactive imagery than after rote rehearsal. Similarly, Rabinowitz et al. (1980) found that the accuracy of immediate JOLs was reliably lower after interactive imagery than after rote rehearsal, and Shaughnessy (1981) found a trend in the same direction. People may be relatively poor at assessing the me-

[1] This conclusion based on the above comparisons of cell means can also be examined via a repeated-measures analysis of variance of the interaction. Although both kinds of analysis are appropriate, comparisons of cell means were conducted because they stemmed directly from the critical predictions of the hypotheses. This comparison of cell means (versus using analysis of variance) to test for an interaction follows Toothaker's (1993) recent conclusion that "the issue of tests on interaction effects versus tests on cell means has been discussed and resolved in favor of tests on cell means because they are easier to interpret, deal with hypotheses that are closer to the original hypotheses tested by most researchers, and contain the total impact on the subjects of both main effects and interaction" (p. 79). Even though Toothaker (1993) concludes that an off-the-shelf analysis of variance is less appropriate than the cell-means analysis reported above, we note that a 2 × 2 repeated-measures analysis of variance also yielded a reliable interaction between study activity and timing of JOLs ($F(1, 35) = 30.16$, $MS_e = 273.55$).

TABLE 2
GOODMAN–KRUSKAL GAMMA CORRELATION BETWEEN JUDGMENTS OF LEARNING AND EVENTUAL RECALL PERFORMANCE

	Time of the JOL		
Kind of study activity	Immediate	Delayed	Overall
Experiment 1			
Interactive imagery	+.10 (.09)	+.72 (.08)	+.39
Rote rehearsal	+.29 (.11)	+.93 (.02)	+.63
Overall	+.18	+.83	
Experiment 2			
Distributed repetitions	+.14 (.14)	+.71 (.09)	+.43
Massed repetitions	+.12 (.13)	+.83 (.06)	+.51
Single presentation	+.20 (.13)	+.91 (.05)	+.54
Overall	+.21	+.83	

Note. Main entries are mean gammas; entries in parentheses are standard errors of the mean.

morial advantages of one image versus another and may instead base their JOLs at least partly on aspects of the images that do not affect eventual memory performance (e.g., bizarreness; see Kroll et al., 1992).

Recall as a Function of Judgment-of-Learning Rating

For each subject, the proportion of items correctly recalled was calculated separately at each level of JOL rating (i.e., for items that had received a JOL of 0, for items that had received a JOL of 20, and so forth). At each JOL rating, means across individual subjects' proportions were calculated and are reported in Table 3.

The magnitude of correct recall increased monotonically with JOL rating for delayed JOLs, but there were some inversions for immediate JOLs. This positive relationship between JOLs and recall is consistent with the item-by-item accuracy described above.

Inspection of Table 3 also suggests how the timing of JOLs affected item-by-item accuracy. For instance, consider items studied via rote rehearsal. In the case of delayed JOLs, subjects rarely recalled an item when they had made a JOL rating of 0, and they rarely failed to recall an item when they had made a JOL rating of 100. However, a different pattern occurred for immediate JOLs: Although the likelihood of recall was low when people gave a JOL rating of 0, when they gave a high JOL rating (e.g., 60 or 80 or 100), they were more likely to not recall the item than to recall it.

TABLE 3
MEAN PROPORTION OF CORRECT EVENTUAL RECALL AS A FUNCTION OF JUDGMENT-OF-LEARNING (JOL) RATING

	Judgment-of-learning rating					
Study activity	0	20	40	60	80	100
Interactive imagery						
Immediate JOLs	.24 (.09)	.56 (.07)	.60 (.06)	.58 (.07)	.67 (.07)	.73 (.10)
Delayed JOLs	.23 (.06)	.41 (.09)	.58 (.11)	.75 (.08)	.78 (.08)	.92 (.02)
Rote rehearsal						
Immediate JOLs	.12 (.06)	.14 (.04)	.20 (.05)	.32 (.07)	.38 (.09)	.11 (.11)
Delayed JOLs	.03 (.01)	.13 (.06)	.34 (.10)	.62 (.10)	.67 (.10)	.96 (.03)

Note. Main entries are mean proportion of correct recall for items receiving a given judgment-of-learning rating; entries in parentheses are standard errors of the mean.

Proportion of Items Receiving Each Judgment-of-Learning Rating

A finer-grained analysis of how people use the rating scale when making JOLs is provided by examining the proportion of items that received each JOL rating. For each subject, the proportion of items that received each JOL rating was calculated separately for immediate versus delayed JOLs and for imagery versus rote. The mean across the individual subjects' proportions is shown in Fig. 2 for each JOL rating.

For delayed JOLs, the subjects displayed more polarization in their ratings, using extreme values of the scale more frequently than middle values. For immediate JOLs, by contrast, the subjects used middle values more often than extreme values. These different patterns of using the JOL rating scale were confirmed by inferential statistical tests. When we computed for each subject the proportion of items that had received either of the most extreme JOL ratings (JOL or 0 or 100) and the proportion of items that had received either of the middle JOL ratings (JOL of 40 or 60), the following four outcomes occurred: (a) For delayed JOLs on items studied via interactive imagery, 31 subjects made more extreme than middle JOLs, whereas only 5 subjects had the opposite pattern; (b) for immediate JOLs on items studied via interactive imagery, 30 subjects made more middle than extreme JOLs, whereas 7 subjects had the opposite pattern; (c) for delayed JOLs on items studied via rote rehearsal, 34 subjects made more extreme than middle JOLs, whereas only 4 subjects had the opposite pattern; and (d) for immediate JOLs on items studied via rote rehearsal, 26 subjects made more middle than extreme JOLs, whereas only 10 subjects had the opposite pattern. All four of those outcomes are reliable by a sign test.

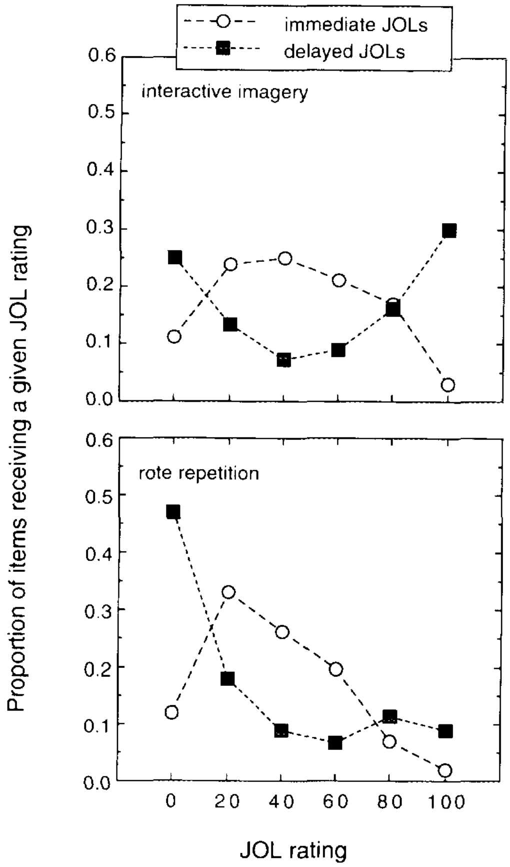

FIG. 2. The mean (across subjects) of the proportion of items that had received a given JOL rating (for each of the six possible JOL ratings), showing that the distribution of JOL ratings is different when the JOLs are immediate versus delayed (Experiment 1).

The above patterns have a straightforward interpretation. When people make delayed JOLs (rather than immediate JOLs), they are more confident that they know (JOL = 100) or don't know (JOL = 0). The relative lack of confidence when making immediate JOLs is made manifest when people assign proportionately more items to the middle values of the JOL rating scale (JOL = 40 or 60). The reduced confidence for immediate JOLs may occur because although people are able to retrieve almost every item at the time of the immediate JOLs, they are also aware that not every item will be retrieved on the eventual recall test, even though the particular items that will eventually be forgotten cannot yet be identified (cf. Griffin & Tversky's, 1992, explanation for the aggregation effect in retrospective confidence judgments). By contrast, the successful retrieval of an item at the time of a delayed JOL may be viewed by the person as strongly indicative of subsequent successful recall of that item, and unsuccessful retrieval may be viewed as strongly indicative of subsequent unsuccessful recall. This interpretation is also consistent with the finding that JOL ratings are more polarized when they follow test trials that occur shortly after study than when they are made in the absence of test trials (Lovelace, 1984).

Another implication of the different distributions for immediate versus delayed JOLs shown in Fig. 2 is that the greater sensitivity of delayed JOLs to various study activities does not arise from a unidirectional shift in people's JOLs across all items; for instance, people are not merely adding some constant percentage to each delayed JOL for items studied under interactive-imagery instructions.

Experiment 2

In Experiment 2, we investigated how delayed JOLs and immediate JOLs are affected by studying items via massed or distributed repetitions. Single-presentation items were also included, so as to provide a baseline for evaluating the effects of repetition.

Method

Stimuli and Apparatus

Items were 48 concrete, unrelated, noun-noun paired associates (e.g., "ocean–tree"). Apple II computers displayed instructions and items and recorded all responses.

Subjects, Design, and Task

Fifty-six undergraduates from the University of Washington participated individually to receive extra course credit. The interval between the study and JOL for an item (immediate or delayed) and the kind of study activity (massed repetitions, distributed repetitions, or a single presentation) were within-subjects manipulations.

The task included one paired-associate study trial. Presentation rate for a single presentation was 4 sec/item. Massed or distributed repetitions were comprised of two 4-sec repetitions. For massed repetitions, the second repetition of an item immediately followed its first repetition. For distributed repetitions, the two repetitions of an item were separated by eight other 4-sec presentations.

A subject-paced JOL occurred sometime after the final presentation of an item. A JOL was prompted with the stimulus alone (e.g., if "ocean–tree" had been presented during study, the cue for the JOL would be "ocean–") along with the query below. Following Zechmeister and Shaughnessy (1980), who had earlier investigated JOLs after massed versus distributed repetition, the JOL we investigated in this experiment was a Likert-type rating scale (rather than the percentile rating scale in Experiment 1). The query for the JOL was, "How well do you think you have learned to respond with the second word of the item when prompted with the first word above (1 = not learned at all. . . . 6 = extremely well learned, TYPE 1, 2, 3, 4, 5, OR 6)."

List Construction

For each subject, the 48 items were randomly ordered for presentation. Thirty-six of the items were separated into three blocks of 12 items/block; four items in a given block were assigned to each of the three kinds of study activity. The remaining twelve items were excluded from the recall trials: Six were presented for practice before the three 12-item blocks, and six were presented at the end of the blocks so as to fill the interval between the study and delayed JOLs for some items (described below).

The order of the three kinds of study activity was counterbalanced across subjects. For half of the subjects, the twelve items of each block were presented in a forward sequence that is illustrated by the following string of letters: AbcDeEfghHBCiIJF-

GKmM. Each of the 20 positions represents one presentation of an item, each of the 12 letters represents a unique item, and capital letters represent the final repetition of an item. For example, items A and D had a single presentation, B and C had distributed repetitions, and E and H had massed repetitions. The other half of the subjects received a sequence that was the reverse order of the above sequence.

For immediate JOLs, the JOL for an item occurred immediately after the final study of that item. For delayed JOLs, the interval between the final study of the item and the JOL was filled with 8 other presentations of items, and therefore the interval between the study and delayed JOL for an item was at least 32 sec. Six filler items were placed at the end of the list to ensure an 8-item interval between the study and delayed JOL for every critical item assigned to receive delayed JOLs.

Immediate versus delayed JOLs were randomly assigned to items with the restrictions that (1) within a given block and for items getting each kind of study activity, one half of the items received delayed JOLs and the other half received immediate JOLs, and (2) two JOLs were never made without at least one study trial intervening between them.

Paired-Associate Recall

Recall trials immediately followed the final delayed JOL and occurred as in Experiment 1. Within each block of items presented for study, the order of recall trials was randomized anew. The items of the first block presented for study were presented first for recall, followed by the items of the second block, and so forth.

Results and Discussion

Recall Performance

As in Experiment 1, we present the analysis of recall before the analysis of JOLs. For each subject, the proportion of correct recall was calculated within each of the three kinds of study activity for items that had immediate JOLs and for items that had delayed JOLs. Means across subjects within each of those six conditions are reported in the lower portion of Table 1.

A 3 × 2 repeated-measures analysis of variance was conducted to assess the effects of study activity (single versus massed versus distributed repetitions) and the timing of JOLs. There was a reliable main effect of study activity ($F(2, 110) = 8.35$, $MS_e = .05$). As expected from previous research (e.g., Peterson et al., 1962), recall was greater for items that had distributed than massed repetitions ($t(55) = 4.13$), which was a prerequisite for testing the hypotheses about metacognition. Recall was also greater for distributed repetitions than single presentations ($t(55) = 5.53$). The difference between massed repetitions and single presentations was not reliable ($t(55) = 1.56$, $p = .12$), which is in agreement with the previous literature, where some experiments yielded a reliable difference between massed versus single repetitions but others did not. The main effect of the timing of JOLs ($F(1, 55) = 1.90$, $MS_e = .06$) and the interaction ($F(2, 110) = 2.12$, $MS_e = .04$) were not reliable.

Magnitude of Judgments of Learning

For each subject, a median JOL was calculated within each kind of study activity for items that had immediate JOLs and for items that had delayed JOLs. Means across the subjects' medians were calculated for each of those six conditions and are reported in Fig. 3. (Note: Unlike the case for Fig. 1, the mean percentage of correct recall is not shown in Fig. 3, because the rating values in Experiment 2 do not correspond to any particular percentages of recall.) Two orthogonal, planned comparisons were conducted as in Experiment 1.

The magnitude of delayed JOLs was reliably greater for items that had distributed than massed repetitions ($t(55) = 2.07$). Thus, delayed JOLs were predictive of the effect of massed versus distributed repeti-

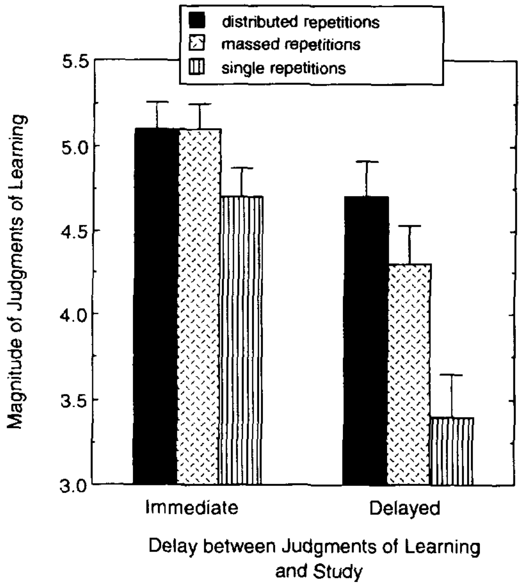

FIG. 3. Mean (across subjects) of the individual subjects' magnitude of judgments of learning, which is the median judgment-of-learning rating given by a subject within each condition. The standard error of the mean (bars) is shown for each condition in Experiment 2.

tions on eventual recall, which disconfirms the hypothesis that people are unable to monitor the effects of these study activities.

The magnitude of immediate JOLs was not reliably different for items that had distributed versus massed repetitions ($t(55) = 0$). In contrast to delayed JOLs, immediate JOLs were not predictive of the greater eventual recall produced by distributed versus massed repetitions.[2] This finding about immediate JOLs for paired associates extends the earlier finding of a negligible effect of massed versus distributed repetitions on metamemory when immediate JOLs were made about the free recall of individual words (Zechmeister & Shaughnessy, 1980).

By the following argument, our findings disconfirm the hypothesis that people will always have greater JOLs for items that are easier to process than for items that are harder to process (cf. Begg et al., 1989). First, Mazzoni and Cornoldi (1993) reported that people allocate less study time to items that are easier to process during study than to items that are harder to process. Second, people allocate less study time to the second presentation of an item having massed repetitions than to an item having distributed repetitions (Shaughnessy, Zimmerman, & Underwood, 1972), suggesting that processing is easier for items having massed repetitions. However, we found that people's JOLs were not greater after massed than distributed repetitions, regardless of the timing of those JOLs. Other disconfirmation of the hypothesis that JOLs are based on ease of processing has occurred for items learned via bizarre versus common imagery (Kroll et al., 1992). Although ease of processing may be one basis for JOLs (Begg et al., 1989), other bases are also used when people make JOLs, and in some circumstances these other bases may overshadow ease of processing as the basis for JOLs.

Unplanned post-hoc comparisons were also conducted to investigate the effects of presenting items once versus twice on the magnitude of JOLs; a Bonferroni correction was used to maintain α at .05. The magnitude of immediate JOLs was lower for items that had a single presentation than for items that had distributed repetitions ($t(55) = 4.03$) or massed repetitions ($t(55) = 4.49$). Zechmeister and Shaughnessy (1980) reported the same pattern of findings concerning immediate JOLs.

The magnitude of delayed JOLs was also lower for items that had a single presentation than for items that had distributed repetitions ($t(55) = 5.30$) or massed repetitions ($t(55) = 3.70$). Thus, both delayed JOLs and recall were sensitive to the kind of

[2] The analysis of this interaction (analogous to the planned comparisons of cell means) based on a 3 × 2 repeated-measures analysis of variance was not reliable ($F(2, 110) = 1.60$, $MS_e = 8.77$). However, because this analysis may be less sensitive and less appropriate than the planned comparisons of cell means (for reasons discussed in Footnote 1), our conclusions are based on the simple-effects tests in which massed versus distributed repetitions had a reliable effect on delayed JOLs but did not have a reliable effect on immediate JOLs.

study activity (cf. recall after these kinds of presentations under the column labeled "Delayed" in Table 1).

Item-by-Item Accuracy of Judgments of Learning

Item-by-item JOL accuracy was analyzed as in Experiment 1. For each subject and within each condition of presentation, one gamma was calculated for items that had immediate JOLs and another was calculated for items that had delayed JOLs. Means across subjects within each of the six conditions are reported in the lower portion of Table 2.

A 3 × 2 analysis of variance was conducted to assess the effects of study activity and the timing of JOLs; Greenhouse and Geisser's epsilon estimate was used to adjust degrees of freedom, as discussed in Myers and Well (1991). (One hundred indeterminate gammas occurred: 53 for immediate JOLs and 47 for delayed JOLs. Thus, many subjects were dropped from the repeated-measures ANOVA, including some subjects who did not have indeterminate gammas in every condition. However, the present analysis and an analysis in which no data were dropped yielded identical conclusions.) There was a main effect of the timing of JOLs ($F(1, 10) = 24.56$, $MS_e = 0.43$): JOL accuracy was greater for delayed JOLs than immediate JOLs. The main effect of study activity was not reliable ($F(2, 20) = 1.36$, $MS_e = 0.22$), but the interaction was close to being reliable ($F(2, 20) = 3.03$, $MS_e = 0.19$, $p = .08$). We note that a posthoc t test showed that the accuracy of delayed JOLs was reliably greater after single presentations than after distributed repetitions ($t(75) = 2.03$); the JOL accuracy may go down during the simultaneous assessment of two traces arising from distributed repetitions (cf. "multiplexing" of memory traces in Hintzman & Block, 1971).

Eventual Recall as a Function of Judgment-of-Learning Rating

For each subject, the proportion of items correctly recalled was calculated separately at each level of JOL rating, and the means across individual subjects' proportions were calculated and are reported in Table 4. For delayed JOLs, the proportion of correct recall increased monotonically with the JOL rating, with only one inversion (see fourth row of data in Table 4). By contrast, for immediate JOLs, the pattern showed many inversions, especially for massed and distributed repetitions. This helps to show why the item-by-item accuracy was lower for immediate JOLs than for delayed JOLs. The difference between those conditions in

TABLE 4
MEAN PROPORTION OF CORRECT EVENTUAL RECALL AS A FUNCTION OF JUDGMENT-OF-LEARNING (JOL) RATING

	Judgment-of-learning rating					
Study activity	1	2	3	4	5	6
Distributed repetitions						
Immediate JOLs	.50 (.50)	.18 (.10)	.25 (.10)	.45 (.08)	.57 (.08)	.54 (.05)
Delayed JOLs	.03 (.03)	.21 (.09)	.24 (.11)	.33 (.09)	.71 (.08)	.73 (.05)
Massed repetitions						
Immediate JOLs	.13 (.11)	.21 (.10)	.42 (.12)	.38 (.09)	.44 (.08)	.39 (.05)
Delayed JOLs	.04 (.03)	.04 (.04)	.43 (.10)	.33 (.13)	.58 (.10)	.78 (.04)
Single presentation						
Immediate JOLs	.11 (.11)	.19 (.09)	.23 (.08)	.33 (.07)	.48 (.08)	.40 (.06)
Delayed JOLs	.03 (.02)	.04 (.04)	.23 (.09)	.32 (.12)	.61 (.11)	.75 (.06)

Note. Main entries are mean proportion of correct recall for items receiving a given judgment-of-learning rating; entries in parentheses are standard errors of the mean.

the gammas for JOL accuracy does not seem to be due merely to particularly high accuracy for only one of the six delayed JOL ratings but rather to more uniformly higher accuracy for the delayed JOLs than for the immediate JOLs.

Proportion of Items Receiving Each Judgment-of-Learning Rating

For each subject, the proportion of items that received each JOL rating was calculated. The means across individual subjects' proportions are shown in Fig. 4.

As in Experiment 1 (cf. Fig. 2), when subjects made delayed JOLs they used extreme values of the scale more frequently than middle values: (a) For distributed repetitions, 43 subjects made more extreme than middle delayed JOLs, whereas only 10 subjects made more middle JOLs. (b) For massed repetitions, 41 subjects made more extreme than middle delayed JOLs, whereas only 8 subjects made more middle JOLs. (c) For single presentations, 45 subjects made more extreme than middle delayed JOLs, whereas only 6 subjects made more middle JOLs. Each of those three outcomes is reliable by a sign test. Also consistent with Experiment 1, a greater proportion of items received a delayed JOL of 1 than an immediate JOL of 1: (a) For distributed repetitions, 24 subjects made more JOLs of 1 for delayed JOLs than for immediate JOLs and vice versa for 1 subject. (b) For massed repetitions, 25 subjects made more JOLs of 1 for delayed JOLs than for immediate JOLs and vice versa for 2 subjects. (c) For single presentations, 39 subjects made more JOLs of 1 for delayed JOLs than for immediate JOLs and vice versa for no subjects. Each of those outcomes is reliable by a sign test.

However, in contrast to Experiment 1 and as is evident in Fig. 4, for immediate JOLs the subjects did not always use the middle values more often than both of the extreme values. For consistency with Experiment 1, first we will mention the omnibus tests: (a) For massed repetitions, 36 subjects made more extreme than middle immediate JOLs, and 13 subjects made more middle JOLs ($p < .05$ by a sign test). (b) For distributed repetitions, 34 subjects made more middle than extreme immediate JOLs, and 19 subjects made more extreme JOLs ($p = .05$). (c) For single presenta-

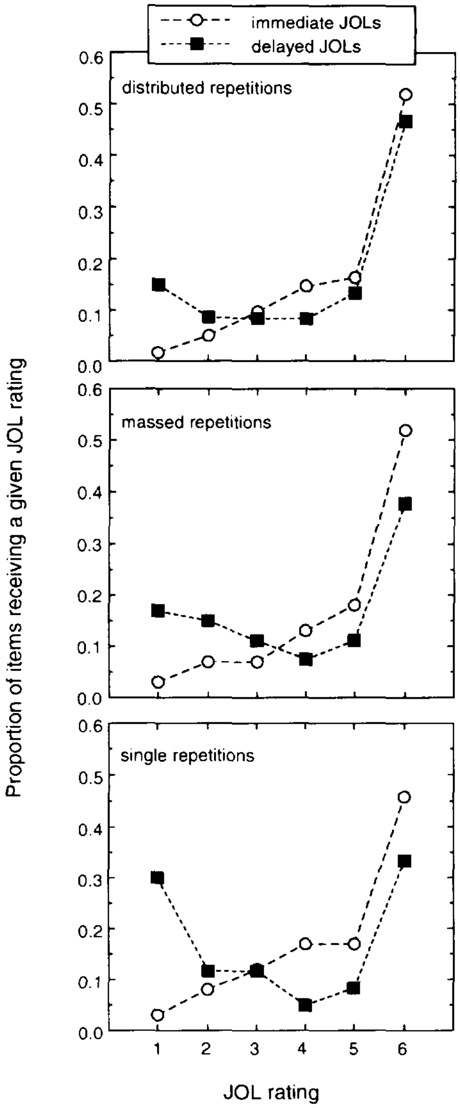

FIG. 4. The mean (across subjects) of the proportion of items that had received a given JOL rating (for each of the six possible JOL ratings), showing that the distribution of low-to-middle JOL ratings is different when the JOLs are immediate versus delayed (Experiment 2).

tions, 31 subjects made more extreme than middle immediate JOLs, and 21 subjects made more middle JOLs ($p = .21$).

The similar results from the above omnibus test notwithstanding, there is a difference between Experiment 1 and Experiment 2 in the distribution of immediate JOLs. This difference is due not so much to the usage of the low or intermediate values of the JOL scale but rather to the usage of the high values of the JOL scale. In particular, when the immediate JOLs are examined separately at the low and high ends of the JOL scale, the following patterns emerge. At the low end of the scale: (a) For distributed repetitions, 19 subjects made fewer JOLs of 1 than JOLs of 3, and vice versa for 1 subject ($p < .05$). (b) For massed repetitions, 12 subjects made fewer JOLs of 1 than JOLs of 3, and vice versa for 7 subjects ($p = .36$). (c) For single presentations, 22 subjects made fewer JOLs of 1 than JOLs of 3, and vice versa for 5 subjects ($p < .05$). Those patterns in Fig. 4 are qualitatively similar to the analogous patterns in Experiment 1 (see Fig. 2).

By contrast, at the high end of the JOL scale the pattern for immediate JOLs was different: (a) For distributed repetitions, 36 subjects made more JOLs of 6 than JOLs of 4, and vice versa for 14 subjects ($p < .05$). (b) For massed repetitions, 38 subjects made more JOLs of 6 than JOLs of 4, and vice versa for 10 subjects ($p < .05$). (c) For single presentations, 32 subjects made more JOLs of 6 than JOLs of 4, and vice versa for 15 subjects ($p < .05$). This qualitative difference between Experiment 1 and Experiment 2 at the high end of the JOL scale may arise from the prompts for the JOLs. A JOL rating of "100%" (in Experiment 1) meant that an item definitely would be recalled 10 min after the JOL, whereas a JOL rating of "6" (in Experiment 2) meant that an item was extremely well learned, with no necessary implication of complete certainty or of a memory that would last for at least 10 minutes. For instance, the subjects may not have incorporated their theory of retention (Maki & Berry, 1984) into their JOLs in Experiment 2 to take into account the possibility that even extremely well-learned items can be forgotten, and sometimes quite quickly.

General Discussion

A question central to the present research was, If eventual memory performance will be greater after one kind of study activity than after another, can people predict that difference? Our findings showed that the magnitude of delayed JOLs was greater for items that were studied under interactive-imagery instructions than under rote-rehearsal instructions (Experiment 1) and was greater for items that had distributed than massed repetitions (Experiment 2). In both experiments, the effect of study activities was also greater on delayed JOLs than on immediate JOLs. Thus, people's on-line item-by-item JOLs can be predictive of the effects of various study activities, but the timing of the JOLs is important.

Implications for Theories of Metacognition

One relatively uninteresting explanation for the present findings is that the extra predictive accuracy for delayed JOLs is due to the interval between JOLs and the criterion test, with accuracy increasing as this interval becomes shorter. In the present experiments, this interval was typically shorter for delayed JOLs than for immediate JOLs. However, this explanation is unlikely to account for the present findings for at least two reasons. First, in terms of rationalistic argument, the functional difference between a retention interval of, say, 10 min for items having delayed JOLs versus 10.5 min for items having immediate JOLs would seem to be quite small (cf. Weber's Law and the negative acceleration in forgetting curves). Second, in terms of relevant empirical evidence from previous research, predictive accuracy is greater for delayed than immediate JOLs even when the interval between JOLs and recall is *greater* for delayed than immediate JOLs (Nelson & Dunlosky, 1991). Thus, the critical variable modulating JOL accuracy does not seem to be the interval between the JOL and the criterion test so much as the interval between study and the JOL.

A more likely explanation for the present findings is that when people make a JOL, they monitor information retrieved from memory about the to-be-judged item (the "monitoring-retrieval hypothesis"). People presumably map this information onto the scale for JOLs via rules that relate the information to eventual memory performance (e.g., "I'll give greater JOLs to items I can recall versus those I can't recall, because I have a better chance of subsequently recalling the former"). This hypothesis is also consistent with previous findings that people base their JOLs at least partly on the outcome of overt tests of the to-be-judged items (King et al., 1980; Lovelace, 1984; Shaughnessy & Zechmeister, 1992). Thus predictive accuracy should increase as a joint function of (a) the degree to which information retrieved from memory about the to-be-judged item at the time of the JOL is predictive of eventual test performance, and (b) the degree to which the aforementioned rules reflect the relationship between the information retrieved and eventual test performance.

The retrieval of responses during delayed JOLs is also highly predictive of recall of one item relative to another. For instance, Runquist (1983, Experiment 2) had subjects study paired-associate items for three sec/pair. After all pairs had been studied, subjects had an initial test of recall (after approximately the same time interval as for our delayed JOLs). Twenty minutes later, subjects had a final test of recall. The probability of final recall for an item that was retrieved on the initial test was near perfect (mean = .96), whereas the probability of final recall for an item that was not retrieved on the initial test was virtually nil (mean = .04). By contrast, the retrieval of responses during immediate JOLs may be less predictive of eventual recall (cf. Craik, 1970).

The proposition that people base JOLs on the retrieval of responses is also consistent with other research on JOLs. Narens, Jameson, and Lee (1994) had subjects make stimulus-alone delayed JOLs (as in the present experiments). A subthreshold prime occurred 250 ms prior to each JOL and consisted of either a nonsense word or the correct response for the item. People's JOLs were greater after response primes than after nonsense primes, and Narens et al. (1994) concluded that this occurred because response priming increased the retrievability of responses when JOLs were made. In other investigations of JOLs (King et al., 1980; Lovelace, 1984), subjects studied items and had initial recall tests, followed by JOLs and final recall tests. JOLs were related more to recall on initial test trials than to recall on final test trials, suggesting that the JOLs are based on the retrieval of responses that occur prior to or during the JOLs.

Retrieval of responses also plays a central role in other uses of memory (e.g., see the distinction between memory as tool vs object; Jacoby & Kelly, 1987) and in theories of other kinds of metacognitive judgments. For instance, Costermans, Lories, and Ansay (1992) and Nelson and Narens (1990) theorized that retrospective-confidence judgments are based on the latency of retrieval, and Koriat (1993, 1994) has recently proposed a theory in which feeling-of-knowing (FOK) judgments are based on the overall retrievability of partial information about responses and stimuli, regardless of whether that partial information is correct or incorrect.

By contrast, other researchers (e.g., Metcalfe, Schwartz, & Joaquim, 1993; Reder & Ritter, 1992; Schwartz & Metcalfe, 1992) have de-emphasized the importance of retrieval for metacognitive judgments and have instead emphasized what they called a *cue-familiarity hypothesis,* wherein people base FOK judgments on the familiarity of the stimulus cues. Although such a hypothesis may be useful in accounting for data about FOKs, it may be less useful in accounting for data about JOLs. For instance, Narens et al. (1994) reported that priming of stimulus cues does not affect the magnitude of delayed JOLs. Thus different theoretical mechanisms may be needed to explain how people make FOKs versus JOLs, perhaps with FOKs based mostly on stimulus familiarity and with JOLs based mostly on retrieval of the response. Such a dissociation is potentially important for general theories of metacognition and should be investigated in future empirical research. Different bases for FOKs versus JOLs would not be a surprise, because Leonesio and Nelson (1990) already reported evidence implying that FOKs and JOLs are based on some (unknown at that time) different aspects of memory. What needs to be determined is the particular aspect(s) of memory underlying each of those kinds of metacognitive monitoring.

Besides basing JOLs on the on-line retrieval of responses, another kind of information that people could monitor when they make JOLs is their *a priori* metacognitive knowledge about the way in which memory is affected by various study activities (Begg et al., 1989, 1991; King et al., 1980). Prior research has already established that people have metacognitive knowledge about the effects of various study activities (see Kreutzer, Leonard, & Flavell, 1975, and Seamon & Virostek, 1978). Although immediate JOLs and delayed JOLs may be a function of both *a priori* metacognitive knowledge and the on-line retrievability of responses, the timing of JOLs may affect the degree to which people base JOLs on each of those kinds of information.

For instance, one possibility is that *a priori* metacognitive knowledge is given greater weight when people make immediate JOLs than when they make delayed JOLs. This might occur for either (or both) of two possible reasons: (1) because people may be less likely during delayed JOLs (than during immediate JOLs) to remember how the to-be-judged item was studied, such that the delayed JOLs might encourage people to rely more heavily on the on-line retrieval of the response, and (2) because people may retrieve almost every response during immediate JOLs but may adjust their average JOLs by their metacognitive knowledge about imperfect eventual recall in prior situations. Also, recent research on the "overshadowing effect" (Price & Yates, 1993) suggests that at some point, the shift from basing JOLs on *a priori* metacognitive knowledge to basing JOLs more on on-line retrieval could become abrupt, such that the delayed JOLs may become based almost completely on on-line retrieval. This shift in the basis for the JOLs may also underlie the greater polarization (and greater confidence) shown in the delayed JOLs than in the immediate JOLs (see Discussion in Experiment 1). Also, item-by-item JOL accuracy may be affected by the degree to which people base their JOLs on *a priori* metacognitive knowledge versus on-line retrieval, especially if one of those two bases for JOLs is less predictive of eventual item-by-item recall.

The monitoring-retrieval hypothesis is a general hypothesis that can be instantiated in several ways. One version that seems especially promising is the *monitoring-dual-memories hypothesis* (Nelson & Dunlosky, 1991), wherein a person is assumed to retrieve information about the to-be-judged item from both short-term memory and long-term memory. However, the information retrieved from short-term memory will function as noise (via interference) for the monitoring of information retrieved from long-term memory, because only the latter is relevant for eventual recall. Moreover, the person may be unable to differentiate between the retrieval of information from long-term memory versus short-term memory. By contrast, when the JOL is delayed until long enough after the study of an item that the information in short-term memory about the item has been forgotten (completed after 30 sec of filled activity; Peterson & Peterson, 1959), then less interference will occur in the monitoring of information retrieved from long-term memory about that item. This may be another reason for the greater polarization (and greater confidence) shown in the delayed JOLs than in the immediate JOLs.

The aforementioned notion of interference between the information retrieved from short-term memory and the information retrieved from long-term memory is assumed to follow the laws of interference that were developed several decades ago and that are not controversial. That is, the amount of interference affecting the person's monitoring of the retrieval of information about the item in long-term memory should be an increasing function of the similarity between whatever is retrieved from short-term memory and whatever is retrieved from long-term memory. Accordingly, when a memory trace of the to-be-judged item is present in short-term memory, as in immediate JOLs or in delayed JOLs with a stimulus-response cue for the JOL, the interference will be maximal to the person's monitoring of that item in long-term memory; when that item is displaced from short-term memory by new items, then because the similarity of those new items in short-term memory to the to-be-judged item is reduced, the amount of interference for monitoring the relevant information from long-term memory will also be reduced. A testable prediction from this notion is that the increase in JOL accuracy as the JOL is delayed (during a filled time interval, so as to have forgetting of the item from short-term memory[3]) should be a mirror image of the rate of forgetting of the potentially interfering information about that item in short-term memory. Other testable predictions can be derived from the degree of similarity between the item being judged and other recently presented items (cf. interference and the transfer surface in Osgood, 1949), but in this case the interference is on the metacognitive monitoring of information retrieved from memory.

The monitoring-retrieval hypothesis focusses on how JOLs are based on the online retrieval of whatever information in memory underlies recall (e.g., some current theories of memory would call this "memory strength"; Gillund & Shiffrin, 1984). The following question then immediately arises and is fundamental for theory: When someone's JOLs accurately predict that eventual memory performance will be better for items that had one kind of study activity rather than another, is this due only to the monitoring of the differences in memory strength produced by the two kinds of activity (i.e., an indirect effect of study activity on the JOLs that is mediated entirely by the memory strength in long-term memory), or is it also due to some direct influence of the study activities per se (i.e., an effect of the study activities on the JOLs that occurs in addition to any indirect effects from the differences in memory strength produced by the various study activities)? This is an important theoretical question for future research to investigate.

How Should Students Monitor Their Memories?

The present research extends the conclusions from previous research in which JOL accuracy was greater for delayed than immediate JOLs (Dunlosky & Nelson, 1992;

[3] The importance of a filled interval is well-known in experiments on forgetting from short-term memory. In accordance with (a) the proposition that the delayed-JOL effect is due to the to-be-judged item being forgotten from short-term memory and (b) the empirical fact that such forgetting requires a filled interval, we note that recent research failed to find a delayed-JOL effect even with stimulus-alone cues when the interval between the study of the item and the delayed JOL was unfilled by any other items; however, when the interval was filled, the usual delayed-JOL effect did occur (Narens, personal communication, 1992).

Kroll et al., 1992; Nelson & Dunlosky, 1991). We found a substantial delayed-JOL effect in three kinds of situations: (a) when items are presented only once, (b) when items are learned via imagery or via rote rehearsal, and (c) when items have repeated presentations. This, combined with our finding that delayed JOLs yield greater sensitivity than immediate JOLs for assessing the effectiveness of different kinds of study activities, suggests that when students have the option of monitoring their memories either immediately after study or after a brief delay, they should delay making their JOLs until a short time after study. Such delayed JOLs should yield both a more accurate prediction of eventual recall and a better informed choice of the kind of study activity that will be most effective for learning those items.

REFERENCES

ARBUCKLE, T. Y., & CUDDY, L. L. (1969). Discrimination of item strength at time of presentation. *Journal of Experimental Psychology*, **81**, 126–131.

BAUER, R. H., KYAW, D., & KILBEY, M. M. (1984). Metamemory of alcoholic Korsakoff patients. *Society for Neurosciences Abstracts*, **10**, 318.

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.

BEGG, I., MARTIN, L., & NEEDHAM, D. (1992). Memory monitoring: How useful is self-knowledge about memory? *European Journal of Cognitive Psychology*, **4**, 195–218.

BOWER, G. H., & WINZENZ, D. (1970). Comparison of associative learning strategies. *Psychonomic Science*, **20**, 119–120.

CARROLL, M. & NELSON, T. O. (1993). Overlearning has a greater influence on the feeling of knowing in with-subjects designs than in between-subjects designs. *American Journal of Psychology*, **106**, 227–235.

COSTERMANS, J., LORIES, G., & ANSAY, C. (1992). Confidence level and feeling of knowing in question answering: The weight of inferential processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **18**, 142–150.

CRAIK, F. I. M. (1970). The fate of primary memory items in free recall. *Journal of Verbal Learning and Verbal Behavior*, **9**, 142–148.

DUNLOSKY, J., & NELSON, T. O. (1992). Importance of the kind of cue for judgments of learning (JOL) and the delayed-JOL effect. *Memory & Cognition*, **20**, 373–380.

FLAVELL, J. H. (1979). Metacognition and cognitive monitoring: a new area of cognitive-developmental inquiry. *American Psychologist*, **34**, 906–911.

GILLUND, G., & SHIFFRIN, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review*, **91**, 1–67.

GRIFFIN, D., & TVERSKY, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive Psychology*, **24**, 411–435.

GRONINGER, L. D. (1976). Predicting recognition during storage: The capacity of the memory system to evaluate itself. *Bulletin of the Psychonomic Society*, **7**, 425–428.

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

HAYS, W. L. (1988). *Statistics* (4th ed.). New York: Holt, Rinehart, Winston.

HINTZMAN, D. L., & BLOCK, R. A. (1971). Repetition and memory: Evidence for a multiple-trace hypothesis. *Journal of Experimental Psychology*, **88**, 297–306.

JACOBY, L. L., & KELLEY, C. M. (1987). Unconscious influences of memory for a prior event. *Personality and Social Psychology Bulletin*, **13**, 464–470.

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.

KORIAT, A. (1993). How do we know that we know? The accessibility model of the feeling of knowing. *Psychological Review*, in press.

KORIAT, A. (1994). Memory's knowledge of its own knowledge: the accessibility account of the feeling of knowing. In J. Metcalfe & A. Shimamura (Eds.), *Metacognition*. MIT press, in press.

KREUTZER, M. A., LEONARD, C., & FLAVELL, J. H. (1975). An interview study of children's knowledge about memory. *Monographs of the Society for Research in Child Development*, **40**, 1–57.

KROLL, N. E. A., JAEGER, G., & DORNFEST, R. (1992). Metamemory for the bizarre. *Journal of Mental Imagery*, **16**, 173–190.

LEONESIO, R. J., & NELSON, T. O. (1990). Do different measures of metamemory 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.

LOVELACE, E. A., & MARSH, G. A. (1985). Prediction and evaluation of memory performance by young and old adults. *Journal of Gerontology*, **40**, 192–197.

Maki, R. H., & Berry, S. L. (1984). Metacomprehension of text material. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **10**, 663–679.

Mazzoni, G., & Cornoldi, C. (1993). Strategies in study time allocation: why is study time sometimes not effective? *Journal of Experimental Psychology*, **122**, 47–60.

Metcalfe, J., Schwartz, B. L., & Joaquim, S. G. (1993). The cue-familiarity heuristic in metacognition. *Journal of Experimental Psychology*, **19**, 851–861.

Murphy, M. D., Sanders, R. E., Gabriesheski, A. S., & Schmitt, F. A. (1981). Metamemory in the aged. *Journal of Gerontology*, **36**, 185–193.

Myers, J. L., & Well, A. D. (1991). *Research design and statistical analysis*. New York: HarperCollins.

Narens, L., Jameson, K. A., & Lee, V. A. (1994). Subthreshold priming and memory monitoring. In J. Metcalfe & A. Shimamura (Eds.), *Metacognition*, MIT press, in press.

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. (Ed.) (1992). *Metacognition: core readings*. Boston: Allyn and Bacon.

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., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. H. Bower (Ed.), *The Psychology of Learning and Motivation*, vol. 26 (pp. 125–173). New York: Academic Press.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: verbal reports on mental processes. *Psychological Review*, **84**, 231–259.

Osgood, C. E. (1949). The similarity paradox in human learning: a resolution. *Psychological Review*, **56**, 132–143.

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).

Peterson, L. R., Hillner, K., & Saltzman, D. (1962). Supplementary report: Time between pairings and short-term retention. *Journal of Experimental Psychology*, **64**, 550–551.

Peterson, L. R., & Peterson, M. J. (1959). Short-term retention of individual verbal items. *Journal of Experimental Psychology*, **58**, 193–198.

Pressley, M., Levin J. R., & Ghatala, E. S. (1984). Memory strategy monitoring in adults and children. *Journal of Verbal Learning and Verbal Behavior*, **23**, 270–288.

Price, P. C., & Yates, J. F. (1993). Judgmental overshadowing: Further evidence of cue interaction in contingency judgments. *Memory & Cognition*, **21**, 561–572.

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.

Reder, L. M., & Ritter, F. E. (1992). What determines initial feeling of knowing? Familiarity with question terms, not with the answer. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **18**, 435–452.

Schwartz, B. L., & Metcalfe, J. (1992). Cue familiarity but not target retrievability enhances feeling-of-knowing judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **18**, 1074–1083.

Seamon, J. G., & Virostek, S. (1978). Memory performance and subject-defined depth of processing. *Memory & Cognition*, **6**, 283–287.

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

Shaughnessy, J. J., & Zechmeister, E. G. (1992). Memory-monitoring accuracy as influenced by the distribution of retrieval practice. *Bulletin of the Psychonomic Society*, **30**, 125–128.

Shaughnessy, J. J., Zimmerman, J., & Underwood, B. J. (1972). Further evidence on the MP-DP effect in free-recall learning. *Journal of Verbal Learning and Verbal Behavior*, **11**, 1–12.

Smith, E. E., Baressi, J., & Gross, A. E. (1971). Imaginal versus verbal coding and the primary-secondary memory distinction. *Journal of Verbal Learning and Verbal Behavior*, **10**, 597–603.

Toothaker, L. E. (1993). *Multiple comparison procedures*. London: Sage Publications.

Townsend, J. T., & Ashby, G. A. (1984). Measurement scales and statistics: The misconception misconceived. *Psychological Bulletin*, **96**, 394–401.

Vesonder, G. T., & Voss, J. F. (1985). On the ability to predict one's own responses while learning. *Journal of Memory and Language*, **24**, 363–376.

Westcourt, K. T., & Atkinson, R. C. (1973). Scanning for information in long- and short-term memory. *Journal of Experimental Psychology*, **98**, 95–101.

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.

(Received May 21, 1993)
(Revision received November 14, 1993)