Memory & Cognition
2006, 34 (7), 1527-1538

Delayed judgments of learning cause both a decrease in absolute accuracy (calibration) and an increase in relative accuracy (resolution)

JAMES P. VAN OVERSCHELDE and THOMAS O. NELSON
University of Maryland, College Park, Maryland

A version of the PRAM methodology that permits an analytical evaluation of judgment of learning (JOL) accuracy was used for the first time to assess absolute accuracy (specifically, calibration). Results are reported from a new experiment in which Swahili–English translation equivalents were studied, followed sometime later (either immediately, ~1 min, or ~8 min) by pre-JOL recall and JOLs, and followed eventually by final recall. The calibration accuracy for predicting final recall decreased as the delay between study and JOL increased, with the decrease being most dramatic when only items that were recalled at the time of the JOL were considered. In contrast, relative accuracy (as measured by an overall gamma) improved as the delay between study and JOL increased. Participants appear insensitive to the combined effects of the recallability of the items at the time of the JOLs and of the delay between JOL and testing on the accuracy of JOLs.

Judgments of learning (JOLs) are defined as judgments that "occur during or after acquisition and are predictions about future test performance on recently studied items" (Nelson & Narens, 1994, p. 16).

There are two methods for evaluating the accuracy of JOLs. The most frequently used method involves calculating a measure of relative accuracy (a.k.a. resolution), usually as a gamma correlation computed in terms of performance on one item relative to performance on another item (e.g., if item *a* received a higher JOL than item *b*, then a person's relative accuracy would be perfect if, when the two items differ in subsequent performance, the likelihood of recall at test is greater for item *a* than for item *b*). One of the most robust findings regarding the relative accuracy of JOLs is that when JOLs are generated immediately after study, the JOL ratings are positively, but only moderately, correlated with eventual recall performance (Nelson & Dunlosky, 1991). However, when the JOLs are delayed briefly (Nelson & Dunlosky, 1991) or for several minutes (Kelemen & Weaver, 1997), then the JOL ratings are positively, and almost perfectly, correlated with eventual recall performance. The increase in the relative accuracy of JOL ratings with a short delay has been termed the *delayed-JOL effect* (for a review, see Schwartz, 1994).

This research was funded, in part, by National Institute of Mental Health Ruth L. Kirschstein NRSA Grant F32 MH070964-01 to J.P.V.O. and by Grant R305H030283 from the Cognition and Student Learning (CASL) research program at the Institute of Education Sciences of the U.S. Department of Education to T.O.N. Correspondence concerning this article should be addressed to J. P. Van Overschelde, 201 Stoney Creek Vista, Wimberley, TX 78676 (e-mail: jimvano@psyc.umd.edu).

Note—This article was accepted by the previous editorial team, when Colin M. MacLeod was Editor.

A second method for evaluating the accuracy of JOLs involves calculating a measure of absolute accuracy (hereafter, we will use the term *calibration*), which is a comparison of the magnitude of all items receiving a particular JOL with the percentage of items recalled correctly at test (e.g., perfect calibration is said to occur when 80% of the items are recalled correctly that had received JOLs of 80%). By comparison with the findings of resolution, the findings regarding the calibration of JOLs are much less consistent. In fact, much research by Nelson and his colleagues has resulted in interesting, but conflicting, findings. Nelson and Dunlosky (1991) found that the calibration of JOLs was more accurate when JOLs were generated after a brief delay than when they were generated immediately after study. More recently, Nelson and his colleagues (Scheck, Meeter, & Nelson, 2004; Scheck & Nelson, 2005) have found that, with difficult items, the calibration of JOLs was less accurate with delayed JOLs than with immediate JOLs, whereas, for easy items, the calibration of JOLs was more accurate with delayed JOLs than with immediate JOLs. In tangentially related research, Koriat and his colleagues have also found several conditions under which the JOLs become less accurate, although not as a function of JOL delay: The calibration of JOLs can decrease as the delay between JOL and test increases (Koriat, Bjork, Sheffer, & Bar, 2004) and as the amount of practice with items increases (Koriat, 1997; Koriat, Sheffer, & Ma'ayan, 2002).

Understanding the factors that lead to accurate calibration of JOL is therefore important for theory. Furthermore, understanding the factors underlying the calibration of JOLs is also important for pedagogical reasons. For example, if students' judgments about the degree of their learning on various items are not accurate, their decisions about which items should receive additional study will be flawed, and test performance may suffer.

Unfortunately, researchers have been hampered in their understanding of the factors underlying the calibration of JOLs in general and of the effect of delaying JOLs in particular, because early experimental methodologies required researchers to infer the recall status of items at the time of the JOLs (Kelemen & Weaver, 1997; Nelson & Dunlosky, 1991; Spellman & Bjork, 1992). In a typical JOL experiment, word pairs were presented for study (e.g., OCEAN–TREE), and, at the time of the JOL, the cue item was presented (e.g., OCEAN–?), to which the participant was assumed to try to covertly recall the target item (e.g., TREE) before making a JOL rating. To eliminate this recall assumption, Nelson, Narens, and Dunlosky (2004) modified the methodology. Called *prejudgment recall and monitoring*, or PRAM, their methodology involves inserting a test trial (typically, cued recall) immediately prior to a JOL trial. By making this methodological change, it is no longer necessary to assume the recall status of items at the time of the JOL. Therefore, items can now be partitioned into subgroups of items that were either recalled or not recalled at the time of the JOL, and relative and absolute accuracies of JOLs can be calculated separately for these two kinds of items.

Using PRAM, Nelson et al. (2004) discovered that relative accuracy of JOLs, calculated for only the items that were recalled correctly at the time of the JOL, was only slightly greater for delayed JOLs than for immediate JOLs. This finding is in stark contrast to the substantial increase in relative accuracy observed when recalled and nonrecalled items are pooled together.

By contrast, analyses using the PRAM methodology (and its decomposition of items into subgroups that are versus are not recalled at the time of the JOLs) have never been attempted for calibration. Accordingly, the primary purpose of the present research was to use PRAM to analytically explore changes in the components of calibration (e.g., calibration for only items recalled at the time of the JOL and calibration for all items combined) so as to better understand the nature of JOLs. Furthermore, because there is no method that researchers agree on for analyzing calibration curves, the secondary purposes of the present research were (1) to examine calibration not only for the mean percentage of items recalled on the criterion test but also for the median percentage of items recalled on the criterion test (which turned out to yield a substantially different pattern than mean recall) and (2) to explore a new analysis of calibration accuracy in which interval-scale assumptions on the criterion variable (number or percentage of items recalled) are not required for the assessment of statistical reliability. Finally, we wanted to replicate and extend Nelson et al.'s (2004) finding of a delayed-JOL effect on relative accuracy (i.e., gamma) by incorporating three different delay intervals between study and JOL instead of the two intervals used previously.

METHOD

Design and Participants

A within-participants design was used with one independent variable (composed of three levels of JOL delay, wherein pre-JOL recall and JOL occurred after 0, 5, or 50 intervening trials on other items, where the trials were either study trials or pre-JOL recall and JOL on other items and where the three conditions were designated JOL0, JOL5, and JOL50, respectively). The mean elapsed time between the offset of a study trial and the onset of a pre-JOL cued recall trial was 0 sec for JOL0, 46 sec for JOL5, and 467 sec for JOL50. A total of 62 undergraduates from the University of Maryland at College Park volunteered to participate in partial fulfillment of a course requirement.

Materials and Procedure

A total of 66 Swahili–English translation equivalents (e.g., ARDHI–SOIL) were drawn from the Nelson and Dunlosky norms (1994). The first 6 items for every participant constituted a primacy buffer and were not evaluated in any analyses. The experiment was composed of three phases: During the first phase, all 66 Swahili–English translation equivalents (hereafter, items) were presented individually in a random order (randomized anew for each participant) at a 7-sec rate to familiarize the participants with the items (Thiede & Dunlosky, 1994). During the second phase, the 60 postprimacy Swahili–English items were presented in a new random order for study at a 7-sec rate. However, during this phase, the participants were instructed to also expect cued recall trials followed by JOL trials for the items and were told to try to learn the items for a final cued recall test during the third phase of the experiment.

A pre-JOL cued recall test occurred immediately prior to the JOL for each item. The cued recall test consisted of the Swahili word as a prompt (e.g., ARDHI–??), and the participants were instructed to type the English translation equivalent (e.g., SOIL). The test was self-paced, and a response was needed to proceed to the next trial. The JOL task was described to the participants as a rating task during which they were to "indicate how confident [they were] that in about 10 minutes [they] would be able to recall the English word when prompted with the Swahili word." The participants were instructed to indicate their confidence by giving ratings of 0, 20, 40, 60, 80, or 100,[1] for which 0 meant *definitely won't recall*, 20 meant *20% likely to recall*, . . . 80 meant *80% likely to recall*, and 100 meant *definitely will recall*. The JOLs were self-paced.

After all 66 items had been studied and had received JOLs, the participants were given a final cued recall test. The items were presented in a new random order, and the same test procedure was used as had been used for the pre-JOL recall.

RESULTS AND DISCUSSION

First, we present the recall data followed by a replication and extension of the effects of JOL delay on the overall and component measures of the relative accuracy of JOLs. Next, we present the most important new findings, which pertain to the effects of the delay between study and JOL on the calibration of JOLs and on the component measures of calibration arising from the use of the PRAM methodology. Finally, we present some parametric data about the effects of the JOL delay on both the magnitude of JOL and the amount of forgetting during this delay and during the delay between JOL and final testing.

To minimize errors due to incorrect spelling, a response was scored as correct if the first four letters were correct (see Nelson et al., 2004). An alpha of .01 was used for all reported analyses, unless noted differently.

Recall

By using the PRAM methodology, we were able to analyze more than just the final test recall data. We also analyzed the percentage of items recalled correctly at the

time of the JOL (really, immediately prior to the JOL) and the percentage of items recalled correctly at the time of the final test conditionalized on correct recall at the time of the pre-JOL recall. The three sets of data are summarized in Figure 1.

Percentage of correct recall during final recall. The mean percentages of items recalled during the final test yielded a significant effect of JOL delay. A repeated measures ANOVA revealed that final recall differed nonmonotonically across the three JOL delays [$F(2,122) = 13.34$]. Post hoc tests showed that final recall was greater for the JOL5 items ($M = 53.6\%$) than for both the JOL0 items ($M = 46.5\%$) [$t(61) = 3.65$] and the JOL50 items ($M = 44.5\%$) [$t(61) = 5.05$], and there was no significant difference between recall of the JOL0 and JOL50 items [$t(61) = 1.12$]. Previous literature shows that the magnitude of the JOL delay produces inconsistent effects on final recall, and the present results might have been due to trade-offs between (1) the potentiating benefits of a successful retrieval at the time of the JOL (Kelemen & Weaver, 1997; Spellman & Bjork, 1992; Whitten & Bjork, 1977) and (2) the likelihood of successful retrieval at the time of the JOL. We assess these two factors in the next paragraphs.

Forgetting during the delay between study and pre-JOL recall. The mean percentages of items correctly recalled during pre-JOL recall were analyzed by a one-way (3 levels of JOL delay) repeated measures ANOVA. As expected, the lengths of JOL delay that we investigated were different enough from each other to produce substantially different levels of recall at the time of the JOLs and reflect a negatively decelerating forgetting function. The analyses revealed a significant effect of JOL delay [$F(2,122) = 244.58$], with greater recall for JOL0 items ($M = 95\%$) than for JOL5 items ($M = 67\%$) [$t(61) = 12.13$] and for JOL5 items than for JOL50 items ($M = 45\%$) [$t(61) = 11.70$].

Forgetting during the delay between pre-JOL recall and final test. The mean percentages of items recalled during the final test conditionalized on correct pre-JOL recall were significantly different across the three JOL delay conditions by a repeated measures ANOVA [$F(2,122) = 65.22$], with final recall increasing monotonically across the three JOL delays. Post hoc tests revealed that the mean conditional percent recalled was significantly smaller for JOL0 items ($M = 48.2\%$) than for JOL5 items ($M = 76.4\%$) [$t(61) = 11.41$] and significantly smaller for JOL5 items than for JOL50 items [$t(61) = 4.33$]. Not surprisingly, the mean percentages of items recalled during the final test conditionalized on incorrect pre-JOL recall were at the floor ($M = 0\%$ for JOL0 items, $M = 4\%$ for JOL5 items, and $M = 2\%$ for JOL50 items) and were not significantly different across the three JOL delay conditions [$F(2,46) = 2.33$, $p > .10$].

Resolution Accuracy

Overall gamma. The Goodman–Kruskal gamma correlation, designated as G, is the most frequently used measure of the resolution of metacognitive judgments (for rationale, see Gonzalez & Nelson, 1996; Nelson, 1984). Gamma (computed on only those dyads of items for which a given dyad contains no tied JOLs and no ties in final recall; e.g., JOL = 80% for one item and 60% for the other item with only one of the two items being correct in final recall) is defined as follows:

$$G = (C - D) / (C + D), \quad (1)$$

where C is the number of concordant dyads of items, and D is the number of discordant dyads of items. Concordant dyads are dyads in which the person predicted greater performance on item a than on item b, and final memory performance was greater for item a than for item b (i.e., item a was recalled, but item b was not recalled). Discor-

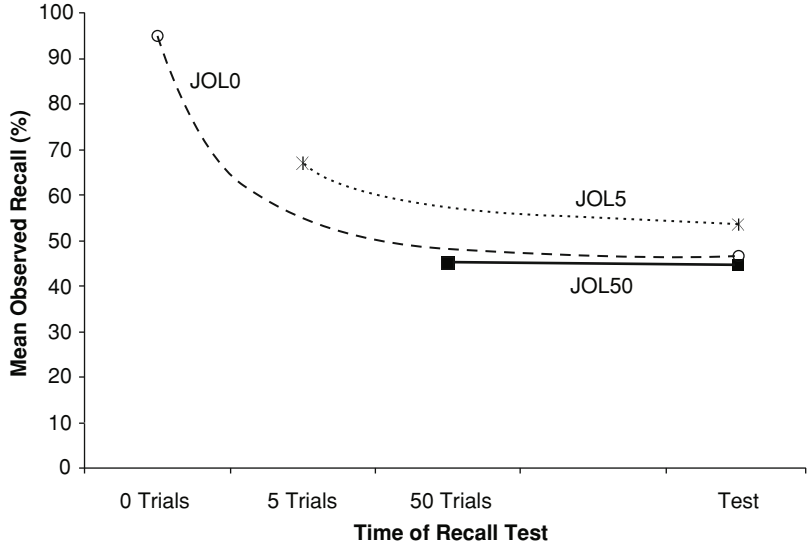


Figure 1. Mean percentage of items recalled during the prejudgment recall test and at final test as a function of the JOL delay, with hypothetical forgetting curves connecting data points within each JOL delay condition.

dant dyads are dyads in which the person predicted greater performance on item *a* than on item *b*, but final memory performance was greater for item *b* than for item *a*.

The mean gamma on all items, regardless of whether they were recalled during pre-JOL recall, is designated *G*.. (following Nelson et al., 2004) and is shown for each of the three levels of JOL delay in the first column of Table 1. A repeated measures ANOVA on the mean gamma correlations yielded a significant effect of JOL delay [$F(2,122) = 23.15$]. Post hoc analyses revealed that the *G*.. for JOL50 items was significantly greater than for JOL5 items [$t(61) = 4.01$], and it was significantly greater for JOL5 items than for JOL0 items [$t(61) = 3.82$]. It is important to note that this increase in *G*.. with the increase in JOL delay replicates Nelson and Dunlosky (1991) and Kelemen and Weaver (1997), who used the traditional (nonPRAM) methodology.

Decomposition of overall gamma into component gammas. By using the PRAM methodology, items can be partitioned into items that were recalled versus those that were not recalled at the time of the JOL. This decomposition allows for the formation of three kinds of dyads: (1) both items were recalled at the time of the JOL (designated as *RR dyads*), (2) one item was recalled and one was not recalled at the time of the JOL (designated as *RN dyads*), and (3) neither item was recalled at the time of the JOL (designated as *NN dyads*). Then, a gamma can be computed for each of the three kinds of dyads (Nelson et al., 2004); these component gammas are designated G_{RR}, G_{RN}, and G_{NN}, respectively. Also, the proportions of all dyads that are of each kind can be computed and are designated as P_{RR}, P_{RN}, and P_{NN}, respectively. For example, P_{RR} would represent the number of RR dyads divided by the total number of dyads, and $P_{RR} + P_{RN} + P_{NN} = 1$ (Nelson et al., 2004). The reason that the three component gammas and the three aforementioned proportions are important is because they can be combined into the following combinatorial rule (from Nelson et al., 2004, Equation 3) that gives rise to the overall gamma (*G*..):

$$G.. = \left(P_{RR} \times G_{RR}\right) + \left(P_{RN} \times G_{RN}\right) + \left(P_{NN} \times G_{NN}\right). \quad (2)$$

The mean component gammas and the corresponding proportions of dyads of each kind, as a function of JOL delay, are shown in the final six columns of Table 1.

A repeated measures ANOVA was computed both on G_{RN} and on G_{RR} as a function of JOL delay. (Note the G_{NN} was indeterminate for JOL0 because there were no non-tied dyads—i.e., final recall was nil for items that had not been recalled during pre-JOL recall—and, therefore, no statistical analysis was computed on G_{NN}.) Because a number of participants had indeterminate G_{RN} and G_{RR}, we replaced missing values with the mean G_{RN} and G_{RR}, respectively.

The first finding of interest is that the main effect of JOL delay on G_{RN} was significant [$F(2,122) = 21.05$]. Post hoc analyses revealed that G_{RN} was significantly greater for JOL50 items than for JOL0 items [$t(61) = 4.51$] and significantly greater for JOL5 items than for JOL0 items [$t(61) = 7.35$]. However, the mean G_{RN} for JOL5 items and JOL50 items did not differ significantly ($t < 1$), probably because of a ceiling effect due to near-perfect JOL accuracy in those two conditions. These results replicate and extend the results from Nelson et al. (2004), who found that G_{RN} was greater for delayed JOLs than for immediate JOLs.

Table 1
Mean Overall and Component Gammas and the Proportion of All Dyads Entering Into Each Component Gamma, as a Function of JOL Delay

	All Dyads	RR Dyads		RN Dyads		NN Dyads	
JOL Delay	*G*..	P_{RR}	G_{RR}	P_{RN}	G_{RN}	P_{NN}	G_{NN}
0	.62	.92	.61	.08	.76	.00	Ind
5	.84	.30	.51	.69	.96	.01	.61
50	.94	.05	.48	.93	.95	.02	.18

Note—Ind, indeterminate.

The second finding of interest is that the main effect of JOL delay on G_{RR} was not significant [$F(2,122) = 1.52$, $p > .10$]. This finding represents a failure to replicate Nelson et al. (2004), who observed a small but significant increase in G_{RR} as JOL delay increased, whereas the trend we observed went in the opposite direction. Some possibilities for this failure to replicate include, but are not limited to, the following: (1) We used Swahili–English translation equivalents, whereas Nelson et al. used familiar words in unrelated, noun–noun pairs; (2) our participants had familiarization trials prior to the study–JOL trials, whereas Nelson et al.'s participants did not; (3) we controlled the number of intervening trials between study and JOL, whereas Nelson et al. used a randomization procedure in which the number of intervening trials ranged from 5 to 30; and (4) Nelson et al.'s list contained 126 items, whereas ours contained 66, which may have resulted in insufficient power in our experiment (e.g., fewer items per participant). We leave this issue as a topic for future research.

The third finding of interest is that G_{RN} was significantly and substantially greater than G_{RR} for all three JOL delays (all ts > 2.56, ps $< .05$). This finding replicates Nelson et al. (2004) and again demonstrates that the accuracy of participants' judgments of future performance on one item relative to another item is greater when, at the time of the JOL, one item is recalled and the other item is not recalled (vs. when both items are recalled).

Calibration Accuracy

Unfortunately, there is no well-established technique for analyzing the reliability of differences between calibration curves that does not make unjustifiably strong scaling assumptions. As a result, meaningful (in the measurement sense; see, e.g., Coombs, Dawes, & Tversky, 1970; Townsend & Ashby, 1983) conclusions about better calibration in one condition than in another condition are difficult to attain.[2] For instance, different conclusions about whether calibration is better in one condition than in another condition can arise depending on whether one analyzes the absolute deviations (vs., say, squared deviations)

between observed performance and perfect calibration. Therefore, in our treatment of calibration accuracy, we report several different descriptive and inferential statistics, as well as a new technique for analyzing the reliability of differences in calibration accuracy between conditions that makes weaker assumptions than other techniques that rely on absolute or squared deviations.

We also apply the PRAM decomposition methodology (which was developed by Nelson et al., 2004, specifically to analyze the relative accuracy of metacognitive judgments) to analyze the calibration accuracy of JOLs. This is the first application of the PRAM methodology to calibration accuracy, and, as a result, it allowed us to evaluate the calibration accuracy separately for (1) all items, (2) only items recalled at the time of the JOL, and (3) only items not recalled at the time of the JOL.

Thus, the organization below unfolds in terms of first reporting several descriptive and inferential statistics on all items, followed next by the corresponding statistics on only items recalled at the time of the JOL, followed finally by the corresponding statistics on only items not recalled at the time of the JOL.

Calibration accuracy on all items. One of the most common methods for analyzing the accuracy of JOLs is to determine the overall mean JOL rating and the mean percentage of items recalled at final test for each participant. A 2 (measure; JOL vs. recall) × 3 (JOL delay) repeated measures ANOVA was computed on these data. There was a significant effect of measure [$F(1,61) = 10.67$, $SEM = 458$], with the mean percentage of items recalled ($M = 48.2$) being greater than the mean JOL rating ($M = 41.0$), which reflected that the participants were underconfident overall (actual > predicted). The interaction was also significant [$F(2,122) = 8.50$, $SEM = 81.1$] (see Figure 2). Post hoc tests of the simple effects revealed that recall and JOL did not differ significantly for JOL0 items ($t < 1$). However, recall was significantly greater than JOL for JOL5 items [$t(61) = 4.38$] and for JOL50 items [$t(61) = 6.65$]. Taken together, these results indicate that, on average, the JOL ratings were similar to actual recall when the JOLs occurred immediately after study, but the JOL ratings were below actual recall when the JOLs occurred after both short and long delays. In other words, the participants were better calibrated when JOLs were generated immediately after study than when JOLs were generated after a delay.

Calibration accuracy on all items: Mean observed recall versus predicted recall. Another common method for evaluating calibration is to generate calibration curves. We did so by computing the mean observed percentage of all items recalled during the final test as a function of the magnitude of JOL (i.e., as a function of the predicted percentage of recall). The resulting calibration curves for JOL0, JOL5, and JOL50 items are shown in the top panel of Figure 3, and the numbers of items at each JOL level on which the calibration curves were based are shown in the Appendix. Unlike Nelson and Dunlosky (1991), who reported that calibration was closer to perfect for delayed-JOL items than for immediate-JOL items, we found that, at intermediate magnitudes of JOL ratings (i.e., 20, 40, and 60), the calibration curve was closer to perfect for JOL0 items than for either JOL5 or JOL50 items (see Figure 3). Furthermore, the calibration curve for JOL50 items was farthest from perfect at those magnitudes of JOL ratings.

By using the PRAM methodology, we decomposed calibration into the component parts that usually are aggregated together—namely, only items that are recalled during pre-JOL recall and only items that are not recalled during pre-JOL recall. To our knowledge, no previous research has evaluated these underlying components of overall calibration and the role they may play in producing the overall calibration curves shown in Figure 3. We report these data next.

Calibration accuracy only on items recalled during pre-JOL recall: Mean observed recall versus predicted recall. For items that were recalled only at the time of the JOL, we computed the mean percentage of items re-

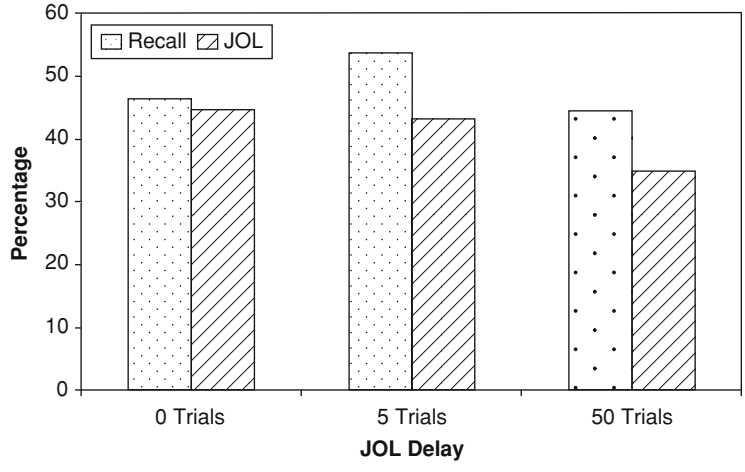
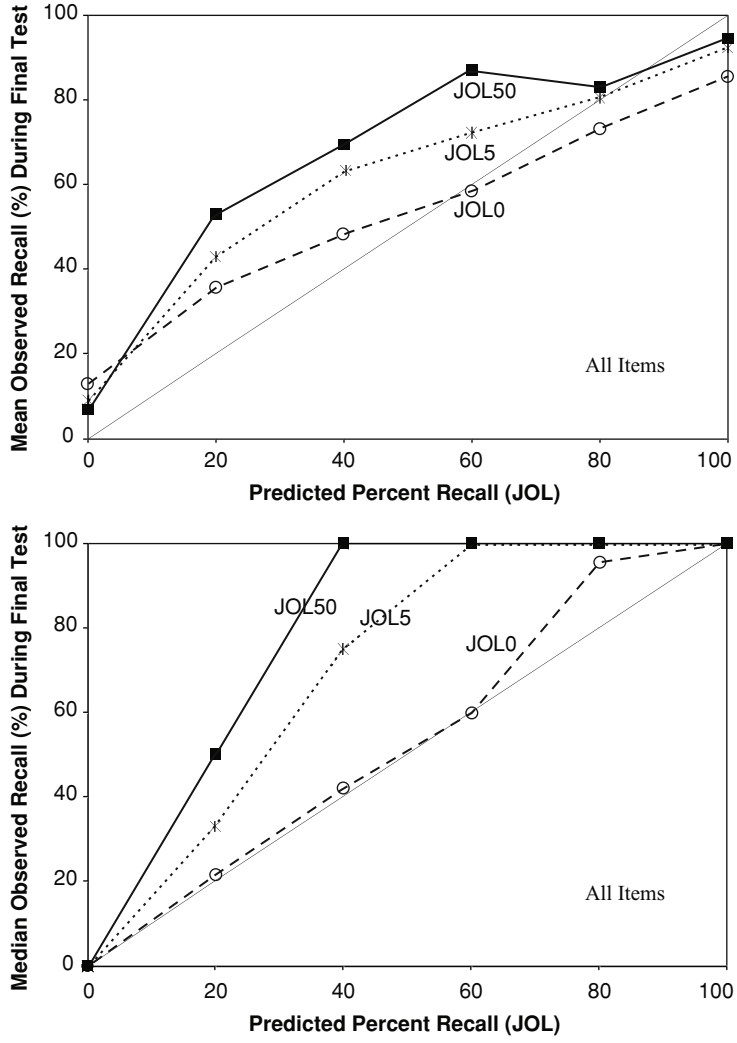


Figure 2. Mean percentage predicted recall (JOL) and mean percentage actual recall as a function of the JOL delay.



	JOL Magnitude					
	0	20	40	60	80	100
A: JOL0 > JOL5	9	22	20	18	13	9
B: JOL0 < JOL5	10	12	10	8	7	9
A/(A+B)	47.4%	64.7%	66.7%	69.2%	65.0%	50.0%
	n.s.	$p < .05$	$p < .05$	$p < .05$	n.s.	n.s.
C: JOL5 > JOL50	7	20	13	13	12	8
D: JOL5 < JOL50	13	10	5	3	6	4
C/(C+D)	35.0%	66.7%	72.2%	81.3%	66.7%	66.7%
	n.s.	$p < .05$	$p < .05$	$p < .05$	n.s.	n.s.

Figure 3. For all items recalled and nonrecalled during pre-JOL recall: Mean percentage of items recalled during final test (top panel) and median percentage of items recalled during final test (middle panel) as a function of the magnitude of JOL (i.e., predicted percentage likelihood of recall) for items in each of the three conditions (JOL0, JOL5, JOL50). The bottom panel contains the comparison of JOL0 versus JOL5 in terms of the frequencies of participants whose percentage of final recall on JOL0 items was closer to perfect calibration than their percentage of final recall on JOL5 items at a given magnitude of JOL (top row), and vice versa for the second row; the third row expresses the entry in the first row as a percentage of all participants who contributed to the first and second rows; the fourth row contains the statistical reliability of the difference between the entry in the third row versus 50% (the null hypothesis) based on a sign test on the frequencies in the first and second rows; and the fifth through eighth rows contain the corresponding entries for the comparison of JOL5 versus JOL50.

called during the final test as a function of the magnitude of JOL. The resulting calibration curves for JOL0, JOL5, and JOL50 items are shown in the top panel of Figure 4, and the numbers of items at each JOL level on which the calibration curves were based are shown in the Appendix. At all magnitudes of JOL ratings from 0 through 60 inclusive, the calibration curve for JOL0 items was closer to perfect than were the corresponding calibration curves for JOL5 items and JOL50 items. Also, the calibration curve for JOL50 items was farthest from perfect at these points. This is roughly in accord with the above-mentioned findings about calibration for all items.

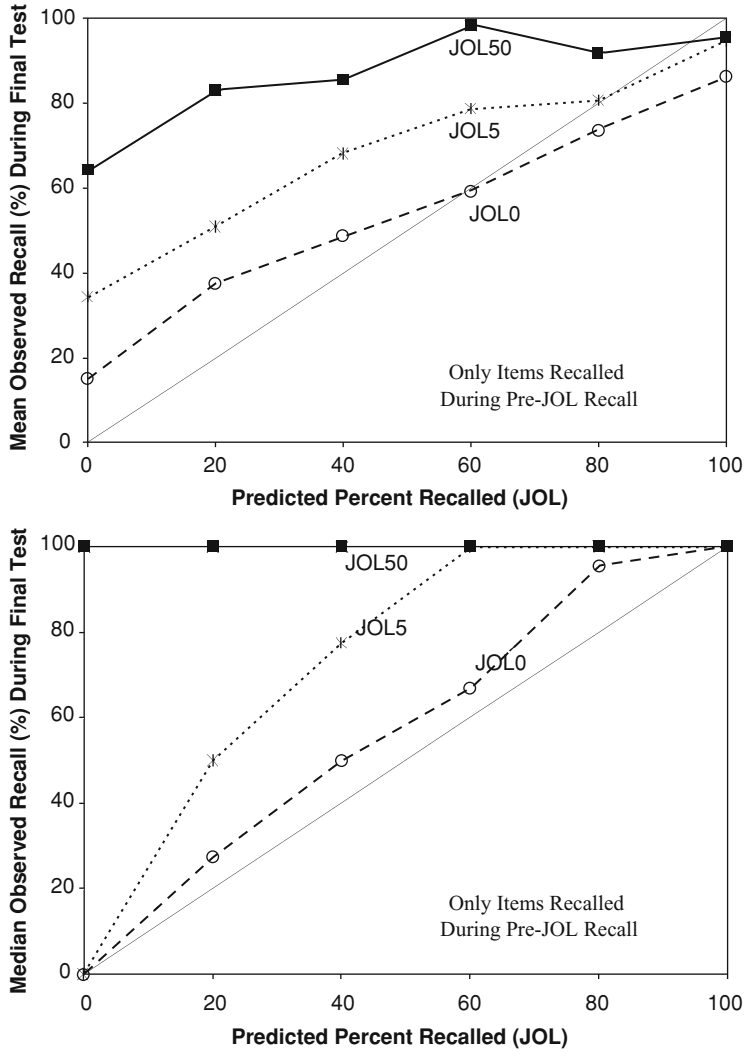
Calibration accuracy on all items: Median observed recall versus predicted recall. Realizing that the mean observed recall could be greatly affected by extreme scores, we evaluated the median (across participants) observed percentage of recall as a function of the predicted percentage of recall. The resulting calibration curves for JOL0, JOL5, and JOL50 items are shown in the middle panel of Figure 3. Several noteworthy findings are evident. First, for JOL0 items, the calibration was nearly perfect across almost all levels of JOL, with the only obvious deviation occurring at JOL = 80 (where the median observed recall was nonetheless closer to perfect calibration for JOL0 items than for JOL5 or JOL50 items). Second, for JOL5 items, at least 50% (i.e., the median) of the participants displayed no differences in observed final recall for items receiving JOLs of 60 versus 80 versus 100, with final recall being at the ceiling for those items. Third, the situation was even more extreme for JOL50 items, where at least 50% (i.e., the median) of the participants displayed no differences in observed recall for items receiving JOLs of 40 versus 60 versus 80 versus 100.

The above patterns for mean observed recall and median observed recall have several implications for investigators of metacognitive accuracy. First, we should be wary of comparing calibration curves derived only from mean observed recall because the nonceiling performance evident in mean observed recall (e.g., the top panel of Figure 3) could have been due to only a small subset of poor-performing participants whose recall was lower than the recall of the majority of participants. For instance, 74% of the participants achieved 100% recall on items that had a JOL rating of 80% in the JOL50 condition (e.g., median = 100% recall for those items; see the middle panel of Figure 3), whereas the remaining 26% of the participants who recalled fewer than 100% of those items caused the mean value to appear as almost perfect calibration (see the top panel of Figure 3). Second, investigators should attempt to evaluate the statistical reliability of differences in calibration between conditions by using inferential tests that do not depend on mean observed performance. Our attempt to accomplish this goal involved comparing each participant's performance in each of the to-be-compared conditions, using only the ordinal aspects of their recall performance, as described after the next section.

Calibration accuracy only on items recalled during pre-JOL recall: Median observed recall versus predicted recall. We evaluated the median (across participants) observed percentage of recall on only items recalled during pre-JOL recall for each magnitude of JOL. The resulting calibration curves for JOL0, JOL5, and JOL50 items are shown in the middle panel of Figure 4. Several findings are noteworthy. First, across all magnitudes of JOL except for 0 and 100 (where median final recall performance was at the floor and ceiling, respectively), the calibration curve was closer to perfect calibration for JOL0 items than for JOL5 items or JOL50 items. Second, in the JOL5 condition, at least 50% (i.e., the median) of the participants displayed no differences in observed final recall for items receiving JOLs of 60 versus 80 versus 100, with final recall being at the ceiling for those items. Third, at least 50% of the participants displayed no differences in final recall (with final recall being at the ceiling for those items) for any magnitude of JOL versus any other when the JOLs had been delayed until 50 trials after study.

New Analysis of Calibration Accuracy

Calibration accuracy on all items: Frequencies of participants who are better calibrated in one of two conditions and statistical reliability of comparisons between two conditions. To avoid making overly strong scaling assumptions about the intervals between a given participant's observed recall performance for items that had received a given magnitude of JOL, but nonetheless to allow an evaluation between two conditions in terms of the accuracy of calibration, we first determined for each participant the observed percentage of recall in each condition for items that had received a given magnitude of JOL rating (e.g., for the items that had received a JOL of 60%, Participant K might have recalled 70% of the items in the JOL0 condition and 80% of the items in the JOL5 condition). Of crucial importance, we included only the participants for whom both of the recall percentages being compared were either greater than or equal to the magnitude of the JOL for those items (e.g., as in the aforementioned example in which both 70% and 80% were greater than 60%) or less than or equal to the magnitude of the JOL for those items (e.g., Participant L's observed percentages of recall of 50% in the JOL5 condition and 40% in the JOL50 condition for items that had received a JOL of 60%). Utilizing only the ordinal aspects of those recall percentages, we could then deem one of those two conditions as being better calibrated than the other (e.g., the JOL0 condition was better calibrated than the JOL5 condition for Participant K above, and the JOL5 condition was better calibrated than the JOL50 condition for Participant L above); importantly, we did not include in this analysis a participant's percentages whenever one of the two percentages was greater than the magnitude of JOL while the other of the two was less than the magnitude of the JOL, because such an inclusion would require assumptions of an interval scale of the deviation of the observed percentage from the percentage predicted by the JOL (e.g., if the observed percentages of recall for Participant M on items that had received a JOL of 80% were 90% for the JOL0 condition and 30% for the JOL50



	JOL Magnitude					
	0	20	40	60	80	100
A: JOL0 > JOL5	10	20	20	18	13	9
B: JOL0 < JOL5	7	10	8	6	7	8
A/(A+B)	58.8%	66.7%	71.4%	75.0%	65.0%	52.9%
	n.s.	$p < .05$	$p < .05$	$p < .05$	n.s.	n.s.
C: JOL5 > JOL50	9	18	13	13	14	8
D: JOL5 < JOL50	3	4	3	0	4	2
C/(C+D)	75.0%	81.8%	81.3%	100.0%	77.8%	80.0%
	n.s.	$p < .05$	$p < .05$	$p < .05$	$p < .05$	$p < .05$

Figure 4. For only items recalled during pre-JOL recall: Mean percentage of items recalled during final test (top panel) and median percentage of items recalled during final test (middle panel) as a function of the magnitude of JOL (i.e., predicted percentage likelihood of recall) for items in each of the three conditions (JOL0, JOL5, JOL50). The bottom panel contains the comparison of JOL0 versus JOL5 in terms of the frequencies of participants whose percentage of final recall on JOL0 items was closer to perfect calibration than their percentage of final recall on JOL5 items at a given magnitude of JOL (top row), and vice versa for the second row; the third row expresses the entry in the first row as a percentage of all participants who contributed to the first and second rows; the fourth row contains the statistical reliability of the difference between the entry in the third row versus 50% (the null hypothesis) based on a sign test on the frequencies in the first and second rows; and the fifth through eighth rows contain the corresponding entries for the comparison of JOL5 versus JOL50.

condition, a conclusion about which of those two conditions was better calibrated would require making an assumption about the interval between 90% and 80% vs. the interval between 30% and 80%, which is an assumption we wished to avoid).[3]

Next, for a given magnitude of JOL, we tallied the frequency of participants whose observed percentage of recall was closer to that magnitude of JOL for the items in Condition I versus Condition J (e.g., Participant K above had greater calibration—i.e., the observed percentage was closer to the magnitude of the JOL—for the JOL0 items than for the JOL5 items) and also the frequency of participants whose observed percentage of recall was closer to that magnitude of JOL for the items in Condition J versus Condition I.

Finally, we compared those two frequencies of participants via a sign test to determine whether significantly more participants were better calibrated in Condition I than in Condition J for a given magnitude of JOL. For instance, a joint outcome of 20 participants being better calibrated in Condition I than in Condition J and 2 participants being better calibrated in Condition J than in Condition I [i.e., $20/(20 + 2) \times 100 = 91\%$ of those participants being better calibrated in Condition I than in Condition J vs. vice versa] would yield a significant ($p < .05$) difference between those two conditions by a sign test.

The outcome of the above analysis for each of the six magnitudes of JOL is summarized in the bottom panel of Figure 3. The entries in each row are the descriptive statistics of (i) the frequency of participants whose observed percentage of recall was closer to the percentage predicted by the magnitude of JOL (given by the column heading) for the first condition than for the second condition, (ii) the corresponding frequency whose observed percentage was farther for the first condition than for the second condition, (iii) the percentage of participants who were closer versus farther [i.e., "i"/("i" + "ii") × 100], and (iv) the statistical reliability (by a sign test) of the deviation of that percentage from the null hypothesis of 50% of the participants being better calibrated on one condition than on the other condition.

As indicated in the bottom panel of Figure 3, more participants were better calibrated in the JOL0 condition than in the JOL5 condition for magnitudes of JOL of 20, 40, and 60 (with no significant differences, or even strong trends, in the direction of JOL5 > JOL0 for the remaining JOL magnitudes of 0, 80, or 100). Put differently, contrary to the conclusion reported by Nelson and Dunlosky (1991, with no replications that we are aware of),[4] delayed JOLs did not have better calibration accuracy than immediate JOLs here. Instead, our findings showed significantly better JOL calibration for immediate JOLs than for JOLs delayed for approximately 1 min, which in turn had better calibration than JOLs delayed for approximately 8 min (i.e., decreasing calibration accuracy with increases in the delay between study and JOL). Thus, when evaluating all items (regardless of whether or not they are recalled at the time of the JOL), we found that calibration accuracy decreased significantly as the JOL delay increased.

Calibration accuracy only on items recalled during pre-JOL recall: Frequencies of participants who are better calibrated in one of two conditions and statistical reliability of comparisons between two conditions. Descriptive statistics in terms of the frequencies of participants whose final-recall calibration on only items recalled during pre-JOL recall was better in one of the two conditions are shown in the bottom panel of Figure 4. We performed the same sign tests as reported above for all items, but this time for only the items recalled during pre-JOL recall. The participants showed better calibration accuracy for JOL0 items than for JOL5 items at magnitudes of JOL of 20, 40, and 60 (all ps $< .05$), with the calibration trends for the remaining magnitudes of JOL also favoring JOL0 items over JOL5 items. Similarly, the participants showed better calibration accuracy for JOL5 items than for JOL50 items for all magnitudes of JOL from 20 through 100 inclusive (all ps $< .05$). Thus, for only items recalled during pre-JOL recall, the calibration accuracy decreased as the delay between study and JOL increased.

Conclusion

The most important new finding of the present research was that as the delay between study and JOL increases (from approximately 0 sec to 8 min), JOLs become less well calibrated (as shown in Figures 2, 3, and 4). This is true when one analyzes both mean recall performance and median recall performance (shown in Figure 3). In fact, the decrease in calibration is more dramatic when median recall performance is considered, with more than 50% of the participants showing no calibration accuracy at several levels of JOL. Furthermore, using the PRAM methodology (Nelson et al., 2004) to analytically explore changes in the components of calibration, we found that the decrease in calibration accuracy of JOLs was most dramatic when only items that were recalled at the time of the JOL were considered (shown in Figure 4), and these findings have broad theoretical and pedagogical implications.

In contrast to the results for the calibration accuracy of JOLs, another important finding of the present research was that the relative accuracy of JOLs, as measured by gamma correlations, increased as JOL delay increased. Furthermore, using the PRAM methodology, gamma was decomposed into its component gammas. Consistent with previous research, we showed that G_{RN} and P_{RN} increased monotonically and significantly, and P_{RR} decreased monotonically and significantly as JOL delay increased. These findings, along with the finding that $G_{RN} >> G_{RR}$, account for the bulk of the increase in the overall gamma correlations observed as the JOL delay increases.

Why Might Delayed JOLs Tend to Underestimate Subsequent Recall?

Why do people (especially in their JOLs for only items recalled at the time of the JOL) tend to underestimate the likelihood of final recall, and why does the magnitude of this underestimation increase with increases in the delay between study and JOL (i.e., greater underestimation for JOL50 items than for JOL5 items and greater underesti-

mation for JOL5 items than for JOL0 items)? One possible explanation, based both on the well-established finding that the probability of recall is a negatively decelerating function (e.g., Bahrick, 1984; Ebbinghaus, 1885/1913; Peterson & Peterson, 1959) and on the assumption that people have some kind of a psychological anchor point for their JOLs (Scheck & Nelson, 2005), is the following. Perhaps, because of their frequent experience at making JOLs immediately after studying a particular item of information (e.g., a phone number or chemical symbol) that is followed by a somewhat lengthy interval during which substantial forgetting occurs, people tend to predict substantial forgetting of the items (e.g., 50%). Because those JOLs are made immediately after study, substantial forgetting does occur soon thereafter (being in the steepest portion of the forgetting curve), consistent with such a prediction. However, when JOLs are delayed, as with the JOL5 and JOL50 items here, people have less knowledge about the amount of forgetting that will occur from that point in time on (being in the shallow portion of the forgetting curve), and if they use the same anchor point as would be appropriate for immediate JOLs (Koriat et al., 2004; Scheck et al., 2004), then they will overestimate the amount of subsequent forgetting (i.e., underestimate the likelihood of subsequent recall, as we found here for JOL5 items and for JOL50 items). Thus, the explanation is that the greater the likelihood of subsequent recall for items that are recalled at the time of the JOL, the greater will be the amount of underestimation, which is in accord with our finding that the greater the likelihood of final recall given correct pre-JOL recall, the greater the magnitude of underestimation in people's JOLs (i.e., the conditional probability of final recall given correct pre-JOL recall increased from JOL0 to JOL5 to JOL50, as did the magnitude of underestimation in the JOLs across those three conditions).

This explanation for JOL calibration accuracy is consistent with Koriat et al.'s (2004) findings that JOLs become increasingly overconfident as the delay between JOL and final test increases. This conclusion follows from the combined influence of two factors: (1) the amount of forgetting occurring during the delay between JOL and final test is likely to increase as the delay between JOL and final test increases, and (2) participants use a single psychological anchor point for estimating JOLs regardless of the actual delay between JOLs and final testing (e.g., approximately 50% in Koriat et al., 2004). Given these two factors, the JOLs will become progressively more overconfident as the delay between JOL and final test increases, which is the pattern observed by Koriat et al. (2004).

Furthermore, this explanation is consistent with data from investigations of the underconfidence with practice effect—the effect whereby JOLs become more underconfident as participants study and are tested on items multiple times (Koriat, 1997; Koriat et al., 2002; Scheck & Nelson, 2005). Not surprisingly, after successfully recalling an item several times, there is little or no forgetting occurring during the delay between JOL and testing. Although participants appear to raise the psychological anchor point on which the JOLs are based, the increase is not enough to compensate for the minimal forgetting that is actually occurring. Hence, the JOLs become more underconfident, on average, with practice.

The present findings are important theoretically for several reasons. First, the kind of analysis used to evaluate the accuracy of JOLs can dramatically alter the results and is likely to affect the subsequent conclusions about the factors underlying JOLs. In the present data, the relative accuracy of JOLs improved significantly as the delay between study and JOL increased, but the calibration of JOLs decreased significantly as the delay increased. Second, because prior research concerning the calibration of JOLs has yielded inconsistent results, the present research demonstrates that the PRAM methodology can yield important new information about factors underlying JOL accuracy, and it is likely to aid researchers who are attempting to better describe metacognitive monitoring processes. Third, the present results were analyzed using both the mean likelihood of recall as a function of the magnitude of JOL and the median likelihood of recall versus tallies of individual participants' likelihood of recall (e.g., as shown in Figures 3 and 4). Different conclusions can arise when calibration is assessed in these different ways, and we suggest that future researchers investigating the absolute accuracy of JOLs should consider examining these assessments of calibration. Fourth, we present a new analysis of calibration in which interval-scale assumptions on the criterion variable (number or percentage of items recalled) are not required for the assessment of statistical reliability.

The present findings are important pedagogically because if students' JOLs about the degree of their learning on various items are not accurate, their decisions about which of the various items they have mastered, which items should receive the highest priority for additional study, or when studying an item (or set of items) should be terminated will be flawed, as a result, and test performance may suffer. Therefore, recommendations based on the present findings would have to consider the students' goals. For example, when students are studying for a test, their judgments will be more accurate in an absolute sense when the judgments are generated immediately after study than when they are generated after a short delay. In other words, immediate judgments will more accurately reflect how well a particular item is known or if additional study time should be allocated to a recallable item than will delayed judgments. If delayed judgments are used, then students are likely to spend much more time studying recallable items than is necessary, when time is likely to be better spent studying nonrecallable items. However, when students are studying for a test, their judgments will be more accurate in a relative sense when the judgments are generated after a long delay than after a short delay or immediately after study. In other words, with delayed JOLs, if item *a* is given a higher JOL than item *b*, then the student knows that item *a* is more likely to be recalled than item *b* at test, but the same is much less true for immediate JOLs. Therefore, students are more accurate at assessing

which items should receive the highest priority for additional study after a delay than immediately after study.

The present research has demonstrated the importance of the PRAM methodology for evaluating JOLs—namely, because it allows researchers to evaluate only items that are recalled at the time of the JOL. This ability to partial out and evaluate only recalled items is important because, unlike JOLs for nonrecalled items, JOLs for recalled items are accurate in an absolute sense only when they reflect the amount of forgetting actually occurring between the time of the JOL and final testing. This methodology allows us to switch from investigating the effect of JOL delay on the accuracy of JOLs to the more pragmatic investigation of the combined effects of the recallability of the items at the time of the JOLs and of the delay between JOL and testing on the accuracy of JOLs.

REFERENCES

Bahrick, H. P. (1984). Semantic memory content in permastore: Fifty years of memory for Spanish learned in school. *Journal of Experimental Psychology: General*, **113**, 1-29.

Coombs, C. H., Dawes, R. M., & Tversky, A. (1970). *Mathematical psychology: An elementary introduction*. Englewood Cliffs, NJ: Prentice-Hall.

Ebbinghaus, H. (1913). *Memory: A contribution to experimental psychology*. New York: Columbia Teachers College. (Original work published 1885)

Gonzalez, R., & Nelson, T. O. (1996). Measuring ordinal association in situations that contain tied scores. *Psychological Bulletin*, **119**, 159-165.

Kelemen, W. L., & Weaver, C. A. (1997). Enhanced metamemory at delays: Why do judgments of learning improve over time? *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **23**, 1394-1409.

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General*, **126**, 349-370.

Koriat, A., Bjork, R. A., Sheffer, L., & Bar, S. K. (2004). Predicting one's own forgetting: The role of experience-based and theory-based processes. *Journal of Experimental Psychology: General*, **133**, 643-656.

Koriat, A., Sheffer, L. & Ma'ayan, H. (2002). Comparing objective and subjective learning curves: Judgments of learning exhibit increased underconfidence with practice. *Journal of Experimental Psychology: General*, **131**, 124-128.

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

Nelson, T. O., & Dunlosky, J. (1991). The delayed-JOL effect: When delaying your judgments of learning can improve the accuracy of your metacognitive monitoring. *Psychological Science*, **2**, 267-270.

Nelson, T. O., & Dunlosky, J. (1994). Norms of paired-associate recall during multitrial learning of Swahili–English translation equivalents. *Memory*, **2**, 325-335.

Nelson, T. O., & Narens, L. (1994). Why investigate metacognition? In J. Metcalfe & A. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 1-25). Cambridge, MA: MIT Press, Bradford Books.

Nelson, T. O., Narens, L., & Dunlosky, J. (2004). A revised methodology for research on metamemory: Pre-judgment recall and monitoring (PRAM). *Psychological Methods*, **9**, 53-69.

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

Scheck, P., Meeter, M., & Nelson, T. O. (2004). Anchoring effects in the absolute accuracy of immediate versus delayed judgments of learning. *Journal of Memory & Language*, **51**, 71-79.

Scheck, P., & Nelson, T. O. (2005). Lack of pervasiveness of the underconfidence-with-practice effect: Boundary conditions and an explanation via anchoring. *Journal of Experimental Psychology: General*, **134**, 124-128.

Schwartz, B. L. (1994). Sources of information in metamemory: Judgments of learning and feelings of knowing. *Psychonomic Bulletin & Review*, **1**, 357-375.

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.

Townsend, J. T., & Ashby, F. G. (1983). *The stochastic modeling of elementary psychological processes*. Cambridge: Cambridge University Press.

Whitten, W. B., & Bjork, R. A. (1977). Learning from tests: Effects of spacing. *Journal of Verbal Learning & Verbal Behavior*, **16**, 465-478.

NOTES

1. The participants were instructed to provide JOLs of only 0, 20, 40, 60, 80, or 100. However, due to a programming error, when the first few participants entered JOL ratings other than these six, the computer accepted them and transformed them to the closest of the six values above; for the remaining participants, no other values except for those six were accepted.

2. Which may explain why JOL research has focused predominantly on relative accuracy, because most researchers agree that the gamma correlation is an effective method for analyzing the relative accuracy of JOLs.

3. Notice that the interval between an observed percentage of recall for one condition and the magnitude of JOL can be compared meaningfully with the interval between a different observed percentage of recall and that same magnitude of JOL (when both observed percentages are greater than the magnitude of JOL) without interval-scale assumptions, because one of those intervals will be a proper subset of the other, and, therefore, only the ordinal aspects of the observed percentages have to be utilized (e.g., any monotonic transformation of the percentages of recall would leave the conclusion unchanged about whether the first or second observed percentage of recall was closer to the magnitude of JOL). This can become important when the scale of a given variable is bounded (e.g., percentage of correct recall is bounded by 0 and 100); for instance, the amount of overestimation for Participant M, who had 30% recall on items that had received JOLs of 80%, could never be matched by an equivalent amount of underestimation for those items, because it is impossible to have an interval of 50% extending above the predicted percentage of 80% (i.e., it is impossible to observe a percentage of recall = 130%). Similarly, the corresponding comparison can be meaningful when both of the observed percentages are less than the magnitude of JOL.

4. Note that Nelson and Dunlosky (1991) assessed the reliability of the difference in calibration accuracy by a sign test on the mean proportion of correct recall for immediate versus delayed JOLs. Such a test ignores individual participants (unlike in the present technique) and was based on only six data points (the means). If their test were conducted on the means in the top panel of Figure 3, no significant differences would emerge; if their test were conducted on the medians in the middle panel of Figure 3, four out of four differences would be in the same direction of JOL0 > JOL5 for calibration accuracy as we observed when the participant was used as the unit of analysis in the sign test.

(*Continued on next page*)

APPENDIX
Total Number of Items at Each Judgment of Learning (JOL) Level on Which the Calibration Curves Are Based

	JOL Rating (%)					
JOL Delay	0	20	40	60	80	100
All Items						
JOL0	194	334	325	160	137	189
JOL5	382	169	142	127	154	245
JOL50	605	105	78	67	129	235
Recalled Items						
JOL0	156	285	237	158	134	185
JOL5	51	127	129	119	150	241
JOL50	25	48	64	61	122	230

(Manuscript received December 21, 2004;
revision accepted for publication July 8, 2005.)