

Research Report

WHEN PEOPLE'S JUDGMENTS OF LEARNING (JOLs) ARE EXTREMELY ACCURATE AT PREDICTING SUBSEQUENT RECALL: The "Delayed-JOL Effect"

Thomas O. Nelson and John Dunlosky

Department of Psychology, University of Washington

Abstract—*Judgments of learning (JOLs), which pertain to knowing what one knows and which help to guide self-paced study during acquisition, have almost never been very accurate at predicting subsequent recall. We recently discovered a situation in which the JOLs can be made to be extremely accurate. Here we report the conditions under which such high accuracy occurs, namely, when the JOL made on the stimulus cue is delayed until shortly after study rather than being made immediately after study. Discussion is focused both on theoretical explanations (to be explored in future research) and on potential applications of the delayed-JOL effect.*

Imagine a student who is studying foreign-language vocabulary (e.g., items such as *chateau-castle*) for an upcoming examination. Two metacognitive components that are fundamental during such self-paced learning are: (1) the *monitoring* of what the student knows, providing a basis for predicting subsequent retention; and (2) the *control* of his or her subsequent allocation of study time (Nelson & Narens, 1990). The interplay between these two components allows for an efficient use of study time. For instance, the learner should allocate more study time to items not yet well-learned and less to items that are already well-learned (for data about the interplay between metacognitive monitoring and metacognitive control of self-paced study time in the adult literature, see Nelson & Leonesio, 1988; for developmental literature, see the "discrimination-utilization hypothesis" in Bisanz, Vesonder, & Voss, 1978).

Thus the self-monitoring that occurs

Send correspondence and reprint requests to Thomas O. Nelson, Psychology (NI-25), University of Washington, Seattle, WA 98195.

during learning, called "judgments of knowing" or "judgments of learning" (JOL), has a guiding role in the self-paced acquisition of new information. Correspondingly, the *accuracy* of JOLs is critical because if the JOLs are inaccurate, the allocation of subsequent study time will correspondingly be less than optimal. Unfortunately, the nearly universal finding in the literature (in the normal adult literature, in the developmental literature, and in the neuropsychological literature) is that the accuracy of JOLs is far from perfect—in fact usually closer to nil than to perfect!

As one extreme example, Vesonder and Voss (1985) had their college students make a JOL immediately after studying each item that had not previously been learned (paired-associate items in Experiments 1 and 2, and sentences in Experiment 3), and the JOLs were not very accurate at predicting subjects' recall performance a few minutes later. In particular, the average Goodman-Kruskal gamma correlation, G , between the JOLs and subsequent recall, ranged across conditions from $+ .09$ to $+ .48$ (see their Tables 1 through 4)¹. Other laboratories have also found JOL

1. G has a possible range from -1.0 to $+1.0$, with 0 representing nil predictive accuracy and with $+1.0$ representing perfect predictive accuracy, and is the best of the available measures for summarizing the accuracy of metacognitive monitoring judgments (for reasons, see Nelson, 1984). Also, G has a probabilistic interpretation (rather than an interpretation in terms of the percentage of variance accounted for), which when applied to JOL accuracy is as follows: If the person gives one item a higher JOL than another item, and one of those two items is eventually recalled whereas the other is not, then what is the probability (P) that the recalled item was the one that had originally received the higher JOL? Thus $P = .50$ when the JOLs are completely uninformative (i.e., nil predictive accuracy, as if the JOLs were assigned by

accuracy in that same range, e.g., $G = +.33$ from Bower and Winchester (1970).

Why is the predictive accuracy of these JOLs so low? We wondered if the low accuracy might hinge on the fact that people were monitoring their memories immediately after study, whereas the criterion test did not occur until several minutes later. That is, given the traditional distinction between short-term memory (STM) and long-term memory (LTM), perhaps: (1) immediately after study, information from STM about the to-be-judged item is adding noise to the person's monitoring of whatever information (including no information!) about that item is retrievable from LTM; and (2) the eventual test performance will be based on only the information retrievable from LTM. This idea that *people who are making JOLs will monitor both their STM and LTM* we call the Monitoring-Dual-Memories (MDM) principle. Given the MDM principle and the fact that the aforementioned experiments yielding low JOL accuracy had requested the JOLs immediately but did not assess recall until several minutes later, the hypothesis we tested was this: Delaying the person's JOL long enough to exceed the duration of information in STM (estimated at no longer than 30 s; e.g., Peterson & Peterson, 1959) will increase the accuracy of the JOL for predicting eventual recall from LTM, because there will be no information in STM about the to-be-judged item to add noise to the delayed JOL's assessment of LTM.

chance), and $P = 1.0$ when the JOLs are perfectly accurate. Most important, we can determine P from G via the equation $P = .5 + .5G$ (derived in Nelson, 1984, Eq. 7). Accordingly, for the aforementioned JOLs the P ranges from $.55$ to $.74$, which is closer to the chance value of $.50$ than to the perfect-accuracy value of 1.0 .

METHOD

Subjects and Tasks

Subjects were 30 undergraduates from the University of Washington who volunteered for extra course credit. The task involved one study-test trial of paired-associate learning, with the subject being instructed to study every pair so that later he or she could recall the second word (response word) when cued with the first word (stimulus word). The pairs were presented for study at an 8-s rate; these items and this presentation rate were based on earlier research that had yielded intermediate levels of eventual recall (i.e., far from the performance ceiling or floor). The subject's other task was to make a self-paced JOL for each item, with the cue for JOL being the stimulus word from a given pair 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)?

The experiment was controlled by an Apple II computer, and the subject typed all responses on the computer keyboard.

Items and Design

The items were 66 pairs of unrelated concrete nouns (e.g., OCEAN-TREE). The list was constructed so that the first six items were practice items not tested for recall. The remaining 60 items were divided into two blocks of 30 items per block, with 15 items in each block randomly assigned to the immediate-JOL condition and 15 to the delayed-JOL condition. The order of the items slated to receive immediate versus delayed JOL was random, except for the restriction that no more than three items from the same JOL condition could occur consecutively. For the immediate JOLs, the cue for JOL occurred immediately after the offset of the item. For the 15 delayed JOLs in a given block, the following algorithm was used: Following both the 15th immediate JOL and the study of all items slated to receive delayed JOL, the

JOL cues for the first five items that had been slated to receive delayed JOL occurred in random order, followed by the JOL cues for the second five items slated to receive delayed JOL, followed by the JOL cues for the final five items slated to receive delayed JOL. Thus the delayed JOLs for Block 1 occurred after all of the immediate JOLs in Block 1 but before any of the immediate JOLs in Block 2, and the number of other items (either being studied or receiving JOL) between the study and JOL for a given delayed-JOL item was a computer-controlled random variable whose values were never less than 10 nor greater than 33. Immediately after the final delayed-JOL in Block 2, the self-paced recall test occurred such that all of the items in Block 1 were tested before any items in Block 2 were tested. The order of the items in each block was randomized anew from study to test.

Thus a within-subject design was used in which the JOL for a given item occurred either immediately after studying the item or was delayed for a short while after study (as operationalized above). This allowed for a within-subject comparison of the kind of self-monitoring that potentially could be used by a student studying foreign-language vocabulary, either by making a JOL immediately after studying a given item or by delaying the JOLs until after studying a block of items.

RESULTS

Before presenting the data concerning JOL accuracy, we should mention that the procedure was successful at yielding intermediate levels of recall: The mean percentage of correct recall was 45.5% (SEM = 5.1) for items that had immediate JOLs and 45.9% (SEM = 5.3) for items that had delayed JOLs.

Relative Accuracy of the JOLs

In accord with past research, a Goodman-Kruskal gamma correlation (G) was computed between JOL and subsequent recall, with one G computed for a given subject's items that had immediate JOL and another G computed for the items that had delayed JOL. JOL accuracy was much greater for delayed-JOL items

(mean $G = +.90$, SEM = .02) than for immediate-JOL items (mean $G = +.38$, SEM = .08), and this difference is highly significant, $p < .001$ by a sign test.²

To make obvious the prevalence of this extreme accuracy for delayed JOLs, Figure 1 shows a frequency distribution

2. Five subjects had an indeterminate G in one of the conditions and therefore are excluded from this analysis. Also, it is worth noting that a particular item with a delayed JOL (versus immediate JOL) will tend to have a slightly smaller amount of intervening activity from other items' study between that particular item's JOL and test, which could alone be responsible for the greater JOL accuracy of delayed JOLs. We assessed this possibility by comparing JOL accuracy for (1) the delayed JOLs in Block 1 versus (2) the immediate JOLs in Block 2. That is, the former have the study and/or JOL of all of the items from Block 2 (and usually several items from Block 1) intervening between their JOLs and recall, whereas the latter have the study and/or JOL of only a subset of the items from Block 2 intervening between their JOLs and recall. The mean gammas for JOL accuracy were +.90 and +.36, respectively, for those two subsets of items. This difference is highly significant ($p < .001$ by a sign test), demonstrating that the increased accuracy of delayed JOL occurs even when the amount of study for other items intervening between the JOL and recall of a given item is slightly less for immediate JOL than for delayed JOL.

We also ran another experiment in which 20 new subjects went through the identical procedure as in the experiment described in the text, except that the recall test occurred for all the items in Block 2 before it occurred for any items in Block 1 (i.e., study and make either immediate or delayed JOLs on items in Block 1, followed by study and make either immediate or delayed JOLs on items in Block 2, followed by the recall test for all items in Block 2, followed by the recall test for all items in Block 1). As in the previous paragraph, the critical comparison is between the delayed-JOL items in Block 1 vs. the immediate-JOL items in Block 2; in this new experiment, the retention interval from the time of the JOL to the recall test was always longer for the delayed-JOL items than for the immediate-JOL items. Accordingly, the eventual recall was significantly lower for the former items than for the latter items: The mean percentages of correct recall were 29% and 40%, respectively, $t(19) = 2.78$, $p < .05$. Of particular interest, the mean G for JOL accuracy was +.78 and +.47, respectively (the median G was +.98 and +.54, respectively), and this advantage in predictive accuracy of delayed

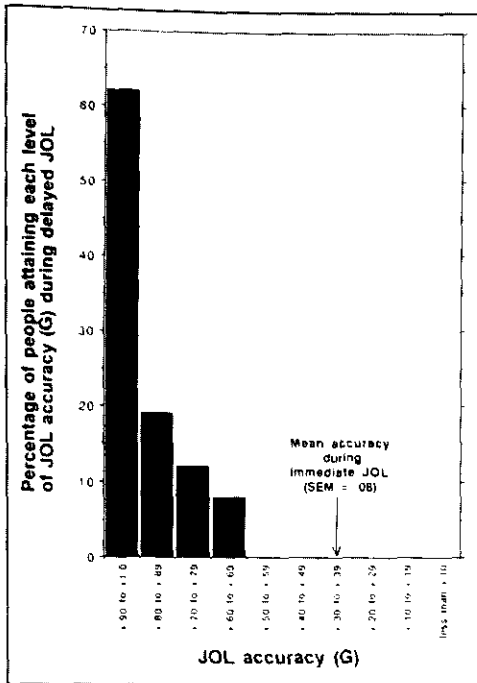


Fig. 1. Percentage of people who attained each level of JOL accuracy as operationalized by the value of the Goodman-Kruskal gamma correlation (G) on delayed-JOL items. For reference, those people's mean G on immediate-JOL items is also shown.

of the individual subjects' Gs for delayed JOL. Every subject's accuracy on delayed JOL was greater than the mean of those same subjects' accuracy on immediate JOL! Moreover, if one operationalizes "extreme JOL accuracy" conservatively as a G of +.90 or greater, then Figure 1 shows that the delayed JOLs yielded extreme JOL accuracy for more than 60% of the subjects; if one operationalizes "extreme JOL accuracy" more liberally as a G of +.80 or greater, then the delayed JOLs were extremely accurate for 80% of the subjects! No subject's delayed JOLs yielded a G less than +.63.

Using the probabilistic interpretation of G described in footnote 1, here the mean P for delayed JOLs is $P = .5 +$

JOL over immediate JOL was significant by a sign test ($p < .05$). Thus the delayed-JOL effect is not contingent upon having the retention interval between the JOLs and recall be shorter for the delayed JOLs than for the immediate JOLs; the delayed-JOL effect occurs even when the retention interval is longer (and recall is lower) for delayed-JOL items than for immediate-JOL items.

$(.5)(.90) = .95$. That is, for items receiving different ratings during delayed JOL and eventuating in different recall, the average probability of accurate metacognitive monitoring for one item relative to another item is .95, which is very close to being perfect.

Absolute Accuracy of the JOLs

As in the previous literature, our primary focus on JOL accuracy was in terms of relative accuracy (i.e., one item relative to another). However, we also explored another analysis of JOL accuracy in terms of the absolute (cardinal) accuracy of these JOLs. (This analysis cannot be done meaningfully when JOLs are made via a Likert scale—e.g., "extremely sure I will recall," "probably will recall," etc.—but may have some worth when the JOLs are made via the numeric likelihood of recall, e.g., "80% chance that I will recall.") In contrast to the evaluation of relative accuracy that showed how accurate people are at monitoring which items they know more about, the focus of this evaluation is on how accurate people are at predicting the likelihood of correct recall on a given item. As a functional example of this distinction, above we showed that the items to which people assign JOLs of, say, 80% are recalled with a greater probability than the items to which they assign JOLs of 20% (i.e., relative JOL accuracy), whereas here we are asking *what percentage of the former items are recalled, and how close is that percentage to the 80% that would be expected if people had perfect absolute JOL accuracy?*

This question can be addressed by constructing a calibration curve (e.g., as used to evaluate the accuracy of confidence judgments about the correctness of previous responses; Lichtenstein, Fischhoff, & Phillips, 1982). Applying this to the JOL situation, we aggregated items receiving the same predicted percentage of recall, computed the actual percentage of those items that were recalled, and plotted the actual percentage against the predicted percentage. The resulting calibration curves for immediate JOLs and for delayed JOLs are shown in Figure 2. The curve for delayed JOLs is very close (and closer than the curve for

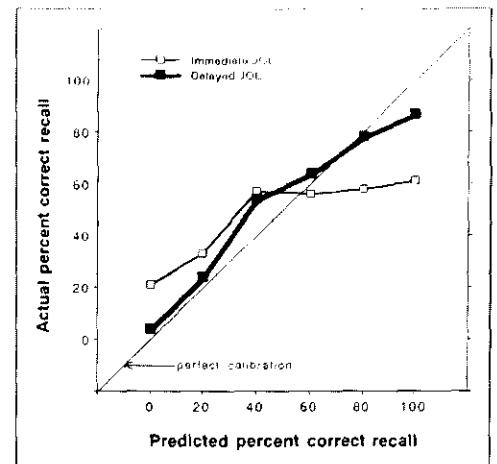


Fig. 2. Mean percentage of recall at each level of predicted recall for immediate JOL and for delayed JOL.

immediate JOLs) to the main diagonal of perfect calibration.³

DISCUSSION

Isolating the conditions under which metacognitive monitoring can be improved is important not just for methodology. It is also important for theories of metacognition, as illustrated by the following three questions (and the beginnings of answers given inside braces): (1) How accurate are people *capable of being* as measuring devices of what they've learned? {Much more accurate than in previous investigations, and almost perfectly accurate!}; (2) Which factors affect that accuracy? {One such factor is the delay between study and JOL for a given item.}; (3) What are the implications for how we should conceptualize the metacognitive monitoring system? {See next paragraph.}

One possible mechanism for the *delayed-JOL effect*—i.e., delayed JOLs being more accurate than immediate JOLs for predicting eventual retention—stems from the aforementioned MDM

3. This is also evident if the data points in Figure 2 are fit by a least-squares linear equation of best fit and the parameter values are compared for immediate vs delayed JOLs: With actual recall designated by Y and predicted recall designated by X, the perfect-calibration equation would be $Y = X$ (or $Y = 1.0X + 0$), and the obtained equation for delayed JOLs is $Y = .84X + 10$ whereas the obtained equation for immediate JOLs is $Y = .39X + 28$.

principle. When people assess what they've learned by querying themselves after study, they may *simultaneously* examine both STM and LTM (relevant data are reported in Wescourt & Atkinson, 1973). The notion that people monitor STM while introspecting was emphasized by Ericsson and Simon (1980). Moreover, the information from STM not only might be adding noise to the evaluation of LTM, but might even dominate the information from LTM (e.g., information is retrieved more quickly from STM than from LTM; Wescourt & Atkinson, 1973). Another possible explanation for the delayed-JOL effect is that the context of a delayed JOL is less like the context existing at study and more like the context existing at recall (cf. "transfer appropriate processing" in Begg et al., 1989). Still other explanations are also possible.⁴

We are currently conducting empirical investigations of both the MDM principle and the boundary conditions for the

4. In end-of-session interviews for the experiment described in the second paragraph of footnote 2, 19 of 20 subjects reported having attempted to recall the correct response to the stimulus cue during delayed JOL, but only half of the subjects reported using that strategy during immediate JOL. Therefore, another potential mechanism for the delayed-JOL effect is that the immediate JOLs (in contrast to the delayed JOLs) may be based less on evidence about recall/nonrecall during JOL than on an inference stemming from other currently unspecified aspects of the situation (i.e., different bases for immediate vs. delayed JOL).

delayed-JOL effect (e.g., preliminary data suggest the possibility of different outcomes from initiating the JOL via a stimulus-only cue vs. via a stimulus-response cue). We have already replicated the delayed-JOL effect in several other experiments that will be reported in a longer article elsewhere (e.g., when the lag between study and JOL for the delayed-JOL condition is always six intervening items, the median G is still +1.0 for delayed JOL, vs. +.43 for immediate JOL).

The delayed-JOL effect is also important for practical applications. For instance, one implication is that the student we considered at the outset of this article should assess how well he or she has learned a given foreign-language vocabulary item by querying after a brief delay rather than immediately after study. Delaying the stimulus cue for JOL (e.g., "How well do I know the English translation of *chateau*?") should both increase the accuracy of the student's self-monitoring as a predictor of future test performance and increase the efficiency of cognitive activities for which self-monitoring is a key component (e.g., decisions about the allocation of subsequent self-paced study time).

Acknowledgment—This research was supported by NIMH grant MH32205 to the first author.

REFERENCES

- 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.
- Bisanz, G.L., Vesonder, G.T., & Voss, J.F. (1978). Knowledge of one's own responding and the relation of such knowledge to learning. *Journal of Experimental Child Psychology*, 25, 116-128.
- Bower, G., & Winchester, P. (1970). [Metamemory judgments.] Stanford University, unpublished data (reported in Leonesio & Nelson, 1990).
- Ericsson, K.A., & Simon, H.A. (1980). Verbal reports as data. *Psychological Review*, 87, 215-251.
- 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.
- Lichtenstein, S., Fischhoff, B., & Phillips, L.D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 306-334). New York: Cambridge University 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., & Leonesio, R.J. (1988). Allocation of self-paced study time and the "labor-in-vain effect." *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 14, 676-686.
- Nelson, T.O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. Bower (Ed.), *The psychology of learning and motivation, Advances in Research and Theory*, Vol. 26. New York: Academic Press, pp. 125-173.
- Peterson, L.R., & Peterson, M.J. (1959). Short-term retention of individual verbal items. *Journal of Experimental Psychology*, 58, 193-198.
- 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.
- Wescourt, K.T., & Atkinson, R.C. (1973). Scanning for information in long- and short-term memory. *Journal of Experimental Psychology*, 98, 95-101.

(RECEIVED 10/11/90; REVISION ACCEPTED 1/21/91)

This document is a scanned copy of a printed document. No warranty is given about the accuracy of the copy. Users should refer to the original published version of the material.