



ELSEVIER

Acta Psychologica xxx (2003) xxx–xxx

**acta
psychologica**

www.elsevier.com/locate/actpsy

Multiple study trials and judgments of learning

Martijn Meeter^{a,*}, Thomas O. Nelson^{b,*}

^a Department of Psychonomics, University of Amsterdam, Roetersstraat 15, 1018 WB Amsterdam, The Netherlands

^b Department of Psychology, University of Maryland, College Park, MD 20742, USA

Received 5 February 2002; received in revised form 9 December 2002; accepted 27 January 2003

Abstract

We compared judgments of learning (JOLs) that were made either (a) after 1 study trial, (b) 2 study trials, or (c) in-between the 1st and 2nd study trials. In regard to the absolute accuracy of JOLs at predicting subsequent recall, we replicated previous findings of an underconfidence-with-practice effect for immediate JOLs and report for the first time a new finding of an underconfidence-with-practice effect for delayed JOLs (i.e., delayed JOLs after one trial overestimated the likelihood of subsequent recall, whereas delayed JOLs after two trials underestimated that likelihood). Also, although delayed JOLs always had a greater relative accuracy than did immediate JOLs, the relative accuracy of immediate and delayed JOLs was approximately the same after 1 versus 2 study trials. These results demonstrate that additional study trials affect the absolute accuracy of all JOLs but not the relative accuracy of any JOLs. Thus an increase in the number of study trials produced an increasing bias to be underconfident about the subsequent likelihood of recall but did not affect people's ordering of which items had been more (versus less) well-learned.

© 2003 Published by Elsevier Science B.V.

PsycINFO classification: 2343

Keywords: Metacognition; Confidence; Memory

1. Introduction

A judgment of learning (JOL) is a prediction of memory performance after acquisition of an item. Subjects are typically asked: “Will you be able to remember this item in about 10 min?” Subsequently a memory test occurs, and the predictions of

* Corresponding authors. Tel.: +31-20-5256724 (M. Meeter).

E-mail addresses: mmeeter@fmg.uva.nl (M. Meeter), tnelson@glue.umd.edu (T.O. Nelson).

subjects are compared to recall performance (e.g., Arbuckle & Cuddy, 1969; Dunlosky & Nelson, 1994; Koriat, 1997; Lovelace, 1984; Nelson & Dunlosky, 1991; Shaughnessy, 1981; Tabor Connor, Dunlosky, & Hertzog, 1997) or occasionally to recognition performance (Thiede & Dunlosky, 1994) or motor learning (Simon & Bjork, 2001).

According to the inferential hypothesis of metamemory, learners who make JOLs do not have access to the strength of the memory they have to judge. Instead, they have to make an inference about future recall on the basis of cues available to them (Koriat, 1997; Schwartz, Benjamin, & Bjork, 1997). Koriat (1997) has divided cues used into internal cues (pertaining to the characteristics of the item such as its perceived difficulty), external cues (pertaining to the study regime), and mnemonic cues (pertaining to experiences gathered during learning or during retrieval). Cue use is not always perfect: subjects may ignore certain kinds of cues (Dunlosky & Nelson, 1994; Koriat, 1997), or mistakenly use cues that do not have predictive value (Benjamin, Bjork, & Schwartz, 1998). Especially, external cues often do not receive as great a weight as they should (Koriat, 1997).

A case in point is repetition of items. Koriat, Sheffer, and Ma'ayan (2002), reanalyzing several earlier studies as well as two new ones, showed that subjects tend to underestimate the increase in recall probability associated with item repetitions. Though JOLs increased with the number of presentations, recall probability increased even more. While after one repetition JOLs typically expressed a slight overconfidence of the subject, this changed to an increasing underconfidence after one or more item repetitions. Koriat et al. (2002) called this the underconfidence-with-practice effect.

An important distinction for the present article is between absolute accuracy and relative accuracy. *Relative accuracy* pertains to the accuracy of distinguishing between one object (or item) relative to another object (or item) without regard to the units of measurement and without regard to the assignment of magnitudes on the variable of interest. *Absolute accuracy* pertains to the accuracy of assigning numbers to the objects (or items) in terms of the judged magnitudes on the variable of interest. An example from the measurement of weight may help clarify this distinction. Relative accuracy pertains to determining which of two weights is heavier (e.g., by using a pan balance with one weight on each pan and observing which pan goes down, so as to isolate the *heavier* of the two weights) without regard to assigning numbers that correspond to the magnitude of each weight, whereas absolute accuracy pertains to the assignment of numbers (e.g., by using a bathroom scale to determine how *heavy* a given weight is in terms of the assignment of a magnitude such as “5 kg”) without any direct comparison of one weight with another weight. Thus errors of relative accuracy pertain to judging one weight as being heavier than another weight (whereas in fact the judgment is reversed because the second weight is heavier than the first weight), errors of absolute accuracy pertain to *bias* wherein the number assigned to a given weight is too large or too small (e.g., assigning a number such as “5 kg” to a weight that in fact is only 4 kg). Notice that although relative accuracy (aka “resolution” in the literature on judgment and decision making) is neither more nor less sensitive than

absolute accuracy (aka “calibration” in the literature on judgment and decision making), in some ways relative accuracy is more fundamental than absolute accuracy (Lichtenstein & Fischhoff, 1977, p. 181) such as an error of relative accuracy (e.g., judging weight *J* as being heavier than weight *K* when in fact the reverse is the case) will necessarily lead to an error of absolute accuracy (e.g., at least one of the numbers assigned to those two weights must be wrong) whereas an error of absolute accuracy does not necessarily lead to an error of relative accuracy (e.g., judging weight *J* to be “5 kg” and weight *K* to be “3 kg” would be biased if the actual magnitudes were 4 and 2 kg, respectively, but the relative ordering of weight *J* being heavier than weight *K* would be correct).

Applied to the accuracy of metacognitive JOLs, the above distinction is that relative accuracy pertains to judgments in which the ordering of a given pair of items is correct (e.g., the judgment that item *J* has been learned better than item *K*, and a subsequent retention test yields an outcome of item *J* being recalled but item *K* not being recalled), whereas absolute accuracy pertains to judgments in which the predicted likelihood of subsequent recall corresponds to the obtained likelihood of subsequent recall (e.g., for all items to which a predicted percentage of correct recall was “60%”, the obtained percentage of correct recall on the eventual test is 60%, indicating no bias, or is 70%, indicating underconfidence of the judgments, or is 40%, indicating overconfidence of the judgments).

Accordingly, in the abovementioned study by Koriat et al. (2002), the cited finding of underconfidence reflects bias in absolute accuracy wherein the JOLs predicting the likelihood of subsequent recall were systematically lower than the obtained likelihood of subsequent recall. Practice thus affects the absolute accuracy of JOLs. Effects on the relative accuracy of JOLs have also been investigated, with conflicting results. Some studies found an increase in relative accuracy with practice (Lovelace, 1984; unpublished experiments reported in Koriat et al., 2002), while Koriat (1997, experiment 1) found no improvements if study trials were not interleaved with test trials.

One factor having a strong effect on relative accuracy is the delay between study and JOL. When a JOL is made immediately after a single study trial on an item, as was the case in all studies reviewed above, relative accuracy tends to be low. When the JOL is delayed for awhile, however, relative accuracy is typically very high. For example, Nelson and Dunlosky (1991) found that mean accuracy (as measured by the Goodman–Kruskal gamma correlation between JOLs and recall) increased from +0.38 to +0.90 (maximum possible = +1.0) when JOLs were delayed approximately 1 min after study. This delayed-JOL effect has subsequently been replicated under different conditions and in different subject populations (e.g., Dunlosky & Nelson, 1992, 1994; Kelemen & Weaver, 1997; Tabor Connor et al., 1997; Weaver & Kelemen, 1997).

Although the theoretical mechanism underlying the delayed JOL effect remains controversial, it is generally agreed that subjects engage in covert retrieval attempts when making delayed JOLs (Kelemen & Weaver, 1997; Nelson & Dunlosky, 1991; Spellman & Bjork, 1992). With delayed JOLs, subjects may rely mostly on what Koriat (1997) called mnemonic cues.

Delayed JOLs have been shown to be sensitive to several manipulations to which immediate JOLs are not sensitive. For example, imagery leads to better recall than rote rehearsal; however, although the magnitude of immediate JOLs does not differ across the two kinds of rehearsal, the magnitude of delayed JOLs is greater for imagery than for rote rehearsal (Dunlosky & Nelson, 1994). It is thus possible that delayed JOLs track the effects of repetitions on recall likelihood accurately. This idea gives rise to an hypothesis tested here: delayed JOLs will not show the underconfidence-with-practice effect.

The delayed JOL effect may also explain at least part of the increase in relative accuracy that occurs in immediate JOLs after multiple study trials in some studies (Koriat et al., 2002; Lovelace, 1984). A JOL placed immediately after the second study trial is necessarily delayed with respect to the first study trial, and therefore it may incorporate information about the first study trial. Koriat (1997) offered a similar explanation, suggesting that with more study trials JOLs come to rely more on mnemonic cues.

To test the second hypothesis, we investigated delayed JOLs that occurred after the first study trial but before the second study trial. This *in-between delayed JOL* may accurately assess the contribution of the first study trial but not of the second (that has not yet occurred). If immediate JOLs after two study trials are indeed more accurate because of the delay with respect to the first study trial, they may have the same or better accuracy as in-between delayed JOLs.

Thus the current study was designed in part to replicate the findings of Koriat et al. (2002) with immediate JOLs and to determine empirically whether or not they can be extended to delayed JOLs, as well as exploring whether increases in the relative accuracy of immediate JOLs elicited after multiple study trials may be due to their functioning somewhat as delayed JOLs.

2. Method

2.1. Design

The within-subject design was comprised of five conditions. The conditions differed in whether a given item received one or two study trials, whether the JOL was immediate or delayed, and whether there was an extra study trial *after* the JOL. In all conditions, only one JOL was elicited per item, and the items were evenly distributed across the conditions.

The five conditions are summarized in Fig. 1, which shows that in the first condition, a JOL was elicited immediately after one study trial (*one-trial immediate*). In the second condition, a delayed JOL was elicited after one study trial (*one-trial delayed*). In both of these two conditions there was no second study trial; these conditions are the usual immediate and delayed JOLs (e.g., Nelson & Dunlosky, 1991). In the third condition a delayed JOL was elicited after the first study trial, but immediately before the second study trial on that item. This is the in-between delayed JOL (*in-between delayed*). In the fourth condition, the JOL followed immediately after the

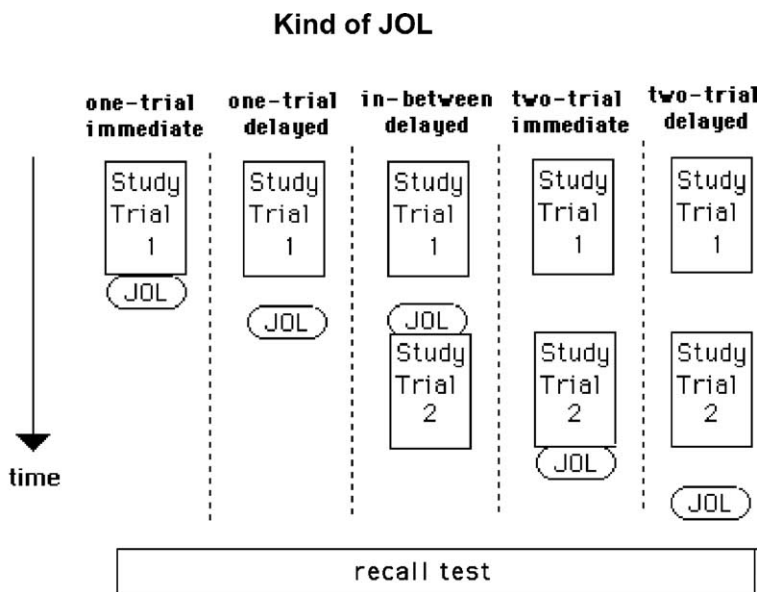


Fig. 1. Design of the experiment for the five within-subject conditions.

second study trial (*two-trials immediate*), and in the fifth condition a delayed JOL was elicited after the second study trial (*two-trials delayed*). Notice that at the time of the JOL the in-between delayed JOL and the one-trial delayed JOL are identical. The subject was not informed in advance about the condition a particular item was in.

2.2. Subjects and materials

The subjects were 54 University of Maryland undergraduates enrolled in introductory and advanced psychology courses, who participated for extra credit. Their mean age was 21. Thirty-three were female, and twenty-two were male.

Items were 60 pairs of unrelated concrete nouns (e.g., OCEAN–TREE) taken from those used by Nelson and Dunlosky (1991). Data collection occurred on Apple computers.

2.3. List construction

Of the 60 pairs, five were randomly designated to a buffer to eliminate primacy effects. The 55 remaining pairs were randomly distributed over five groups of 11 pairs, corresponding to the five conditions in the experiment. For each subject, pairs were randomly reassigned anew to the five conditions. The 55 non-practice items were divided into two blocks that retained their order throughout the experiment (i.e., the first block of items presented for the first and second trials was also the first

block of items during the recall test). The 11 pairs in each condition were evenly divided over two blocks, with the eleventh being a toss-up. Order of the pairs within a block was randomized anew for the second study trial and for the recall test.

2.4. Procedure

All instructions were presented by the computer. Subjects were informed about the structure of the experiment, the list length, and the format of the final test.

When a pair was presented for study, it appeared on the screen for 8 s. The JOLs were elicited as follows. The stimulus word appeared on the screen, and the subject was asked to rate how likely he or she was “to recall the response word when cued with the stimulus word in a test 10 min from now.” The rating occurred on a percentage scale from 0 to 100, in steps of 20 (i.e., 0%, 20%, . . . , 100%).

The time between study and test was approximately 12 min. The time between study and a delayed JOL was approximately 4 min. After all study trials and JOLs had occurred, a distraction task occurred (in which the subjects were asked about their memory in everyday life). Next the self-paced cued-recall test was administered.

Following the test on the final item, subjects received an opportunity to see their recall performance and were asked a final question. They were shown, in a diagram similar to Fig. 1, three kinds of JOLs they had received: the in-between-delayed JOL, the two-trial immediate JOL and the two-trial delayed JOL. They were instructed to rank-order these three kinds of JOL in regard to how accurate the subject believed they had been as predictors of recall (i.e. the subject was asked to judge the accuracy of his or her own JOLs).

3. Results

3.1. Recall performance and magnitude of JOLs

Table 1 shows the proportion of response words recalled in each condition. Recall was well off the floor in all conditions. The differences between the means of the con-

Table 1

Proportion of items recalled, mean magnitude of JOL, difference between mean magnitude of JOL and recall, and mean gamma correlation between recall performance and JOLs for each type of JOL

Type of JOL	One-trial immediate	Two-trial immediate	In between delayed	One-trial delayed	Two-trial delayed
Recall	0.38 (.03)	0.56 (.04)	0.55 (.04)	0.29 (.04)	0.55 (.04)
Magnitude of JOL	0.43 (.03)	0.49 (.03)	0.31 (.03)	0.35 (.03)	0.54 (.04)
Difference	0.05 (.04)	-0.08 (.04)	-0.24 (.03)	0.06 (.02)	-0.01 (.03)
Gamma correlations	0.49 (.07)	0.50 (.08)	0.60 (.08)	0.85 (.07)	0.82 (.08)

Note: mean values (with SEM inside parentheses).

ditions were significant [$F(4, 212) = 41.89; p < 0.001$]. Student Newman Keuls (SNK) post-hoc tests for multiple ranges showed that recall was reliably greater in the conditions with two study trials (two-trial immediate, two-trial delayed, and in-between delayed JOLs) than in the conditions with only one study trial (one-trial immediate and one-trial delayed JOLs). Recall was significantly greater in the condition with one-trial immediate than in the one-trial delayed JOL condition, which in other studies (e.g., Nelson & Dunlosky, 1991) was not the case. As both the presence or absence and the direction of this difference seem to have varied unsystematically across previous experiments, it will not be discussed further.

3.2. *Magnitude of JOLs and absolute accuracy*

All JOLs were made on a percentile scale, but are reported here as proportions to facilitate comparisons with the proportion of correct recall (see Table 1). Mean magnitude of JOL differed reliably between the five conditions [$F(4, 212) = 23.43; p < 0.001$]. SNK post-hoc tests revealed that in the two conditions in which a JOL was elicited after two learning trials had a reliably greater magnitude than the conditions with one learning trial, and the condition with in-between-delayed JOLs.

A repeated-measures *t*-test comparing the magnitude of JOLs (second row of Table 1) with the proportions of items recalled (first row of Table 1) showed two significant differences. JOLs underestimated subsequent recall in the in-between delayed-JOL condition [$t(53) = 7, 78; p < 0.001$], while for the one-trial delayed JOLs there was an overestimation of recall [$t(53) = 2, 65; p < 0.05$].

To test for the underconfidence-with-practice effect (stronger underconfidence with item repetition; Koriat et al., 2002), the difference between JOLs and recall for conditions after a single presentation was compared with the same difference after two presentations. As the third line in Table 1 shows, JOLs in the two conditions with one study trial exhibited overconfidence (though only significant in the delayed JOL condition), while JOLs in the two conditions with two study trials tended toward underconfidence. This shift from over-to-underconfidence was significant both for immediate JOLs [$t(53) = 4.56; p < 0.001$] and for delayed JOLs [$t(53) = 3.65; p < 0.005$].

3.3. *Relative accuracy of JOLs*

As in previous research (e.g., Nelson & Dunlosky, 1991; for rationale, see Nelson, 1984), the relative accuracy of JOLs as predictions of recall was measured by the Goodman–Kruskal gamma correlation (hereafter, gamma correlation) between JOLs and recall. The fourth line of Table 1 gives the mean gamma correlation for the five conditions.

The difference between the conditions was highly significant [$F(4, 96) = 6.56; p < 0.001$]. Differences between individual conditions were analyzed with SNK post-hoc tests. Two clusters of conditions differed reliably from each other. The one-trial immediate JOLs and two-trial immediate JOLs were reliably less accurate than the one-trial delayed JOLs and two-trial delayed JOLs. However, within

Table 2

Percentages of subjects who judged that in-between delayed JOLs, two-trial immediate JOLs, or two-trial delayed JOLs, were most accurate

Type of JOL	In-between delayed	Two-trial immediate	Two-trial delayed
Most accurate according to subjects	18%	55%	27%

each of these clusters, the accuracy of one-trial immediate JOLs did not differ from that of two-trial immediate JOLs, and the accuracy of one-trial delayed JOLs did not differ from that of two-trial delayed JOLs. The accuracy of in-between delayed JOLs was intermediate between these clusters and did not differ reliably from the accuracy in these conditions. Delayed JOLs were thus more accurate than immediate JOLs, with in-between delayed JOLs forming an exception: these were not reliably more accurate than the immediate JOLs.

3.4. Subjects' judgments about their own JOL accuracy

Table 2 shows that the subjects judged the accuracy of the JOL immediately after the second trial to be most accurate of the three kinds of JOLs being judged [$\chi^2(2) = 10.778, p < 0.005$] even though it was in fact the least accurate of these conditions (see third row of Table 1). Two-trial immediate JOLs were judged as being more accurate than both the in-between delayed JOLs [$\chi^2(1) = 9.256, p < 0.002$] and the two-trial delayed JOLs [$\chi^2(1) = 4.455, p < 0.035$]. By contrast, the two-trial delayed JOLs were the most accurate of these conditions (see third row of Table 1) but were judged to be the most accurate by only 27% of the subjects.

4. Discussion

Presenting items only once is a legitimate beginning for investigations of metamemory. In naturalistic learning situations, however, decisions about learning presumed to involve a JOL are often made after multiple study trials and therefore need to be encompassed by theories of metacognition (cf. Nelson & Narens, 1994). For example, a student usually does not stop studying upon going through the items only once. Consequently, it is important to understand how JOLs and their accuracy are affected by multiple study trials.

Analyses showed that multiple study trials increased both recall performance, and the magnitude of JOLs. These increases were not equal in size, however: the present study replicates the underconfidence-with-practice effect for immediate JOLs (Koriat et al., 2002). Of particular importance, the present research shows for the first time that the underconfidence-with-practice effect also occurs for delayed JOLs. This disconfirms the hypothesis described in Section 1 that delayed JOLs would not be subject to the underconfidence-with-practice effect. One possibility for the mechanism underlying the underconfidence-with-practice effect for delayed JOLs can be derived

from Koriat's (1997) distinction between external versus internal cues, namely, perhaps when making JOLs people do not give as much weight to external cues (such as the amount of study time) as objectively should have been given.

With respect to delayed JOLs, other studies have found them to distinguish well between items learned under different study conditions (Dunlosky & Nelson, 1994). Dunlosky and Nelson (1994) did not examine whether the difference in magnitude of delayed JOLs between the two conditions was as large as that in recall, however. From the present findings, one might expect this not to be the case.

With relative accuracy, we did not find the improvement with added study trials reported by Lovelace (1984) and Koriat et al. (2002). An increase in relative accuracy with practice was also absent in the one experiment by Koriat (1997) without test trials in between study trials. Together with the present findings, this suggests that increases in relative accuracy with additional study trials are either very small or only appears under specific circumstances. By and large, additional study trials thus seem to affect absolute accuracy of JOLs, but not relative accuracy. In terms of the distinction drawn above in Section 1, this finding implies that the additional study trials affected the assignment of numbers (i.e., absolute accuracy in which there was bias in the direction of the predicted likelihood of recall underestimating the obtained likelihood of recall) without affecting the more fundamental attribute of judging which items were more well-learned (i.e., the relative accuracy in which the predicted ordering of recall of the two items was roughly as accurate after several study trials as after only one study trial). Thus an increase in the number of study trials produced an increasing bias to be underconfident (when people had to assign numbers reflecting the predicted magnitude of recall) but did not affect people's ordering of which items were more (versus less) well-learned.

One last finding concerned participants' perception of the accuracy of different kinds of JOL. Although the delayed JOL effect is very substantial and was replicated again in this study, more participants thought immediate JOLs were the most accurate than vice versa. This shows that a sizeable portion of participants were not aware of which JOLs were accurate and which were not.

Acknowledgements

This research was partially supported by NIMH grant K05-MH1075 to the second author. We thank Petra Scheck for her help in data collection.

References

- Arbuckle, T., & Cuddy, L. (1969). Discrimination of item strength at time of presentation. *Journal of Experimental Psychology*, *81*, 126–131.
- Benjamin, A. S., Bjork, R. A., & Schwartz, B. L. (1998). The mismeasure of memory: when retrieval fluency is misleading as a metacognitive index. *Journal of Experimental Psychology: General*, *127*, 55–68.

- Dunlosky, J., & Nelson, T. O. (1992). Importance of the kind of cue for judgments of learning and the delayed JOL-effect. *Memory and Cognition*, *20*, 373–380.
- Dunlosky, J., & Nelson, T. O. (1994). Does the sensitivity of JOLs to various study activities depend on when the JOLs occur? *Journal of Memory and Language*, *33*, 545–565.
- 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 and 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., Sheffer, L., & Ma'ayan, H. (2002). Comparing objective and subjective learning curves: judgment of learning exhibit increased underconfidence-with-practice. *Journal of Experimental Psychology: General*, *131*, 147–162.
- Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know? *Organizational Behavior and Human Performance*, *20*, 159–183.
- Lovelace, E. A. (1984). Metamemory: monitoring future recallability during study. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *10*, 756–766.
- Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling of knowing predictions. *Psychological Bulletin*, *95*, 109–133.
- Nelson, T. O., & Dunlosky, J. (1991). When people's judgments of learning are extremely accurate at predicting subsequent recall: the delayed JOL effect. *Psychological Science*, *2*, 267–270.
- Nelson, T. O., & Narens, L. (1994). Why investigate metacognition?. In J. Metcalfe, & A. Shimamura (Eds.), *Metacognition, knowing about knowing*. Cambridge (MA): Bradford books.
- Schwartz, B. L., Benjamin, A. S., & Bjork, R. A. (1997). The inferential and experiential bases of metamemory. *Current Directions in Psychological Science*, *6*, 132–137.
- Shaughnessy, J. (1981). Memory monitoring accuracy and modification of rehearsal strategies. *Journal of Memory and Language*, *20*, 216–230.
- Simon, D. A., & Bjork, R. A. (2001). Metacognition in motor learning. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *27*, 907–912.
- Spellman, B. A., & Bjork, R. A. (1992). When predictions create reality: judgments of learning may alter what they are intended to assess. *Psychological Science*, *5*, 315–316.
- Tabor Connor, L., Dunlosky, J., & Hertzog, C. (1997). Agerelated differences in absolute but not relative metamemory accuracy. *Psychology and Aging*, *12*, 5071.
- 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.
- Weaver, C. A., & Keleman, W. L. (1997). Judgments of learning at delays: shifts in response patterns or increased metamemory accuracy? *Psychological Science*, *8*, 318–321.