Measuring Memory and Metamemory:
Theoretical and Statistical Problems with Assessing Learning (in General) and Using Gamma (in Particular) to Do So

Barbara A. Spellman, Aaron Bloomfield, and Robert A. Bjork

Introduction

This chapter addresses the interrelated problems of assessing learning in general and using γ (the Goodman-Kruskal γ correlation), in particular, to do so. We carry out our analysis in the context of the metamemory literature on judgments of learning (JOLs), but we believe that the lessons learned are widely applicable.

Consequences of Assessing Learning

In what has become a classic metamemory paper, Dunlosky and Nelson (1992) had participants study paired associates such as ocean–tree. Later, the experimenters represented the same items and asked participants to judge how likely they would be to remember the second word if shown the first word 10 minutes later (i.e., they were asked to make JOLs in the form of predicting their future recall performance). There were two independent variables of interest. The first was delay: JOLs were made either immediately (i.e., the next trial after the words were presented) or after some number of intervening (presentation or JOL) trials. The second was type of presentation at the time of making the JOL: Participants saw either the intact cue–target pair (i.e., ocean–tree) or the cue alone (i.e., ocean–?). As measured by γ, JOLs were far more accurate in the delayed cue-only condition than any other condition. The superiority of the delayed cue-only condition is an important effect (e.g., for evaluating whether one has studied enough) and has been replicated many times (see Narens, Nelson, & Scheck; Weaver, Terrell, Krug, & Kelemen, this volume, for a review).

In a similar study, Nelson and Dunlosky (1991) noted that most of their participants reported trying to silently recall the target word when given a delayed cue-only JOL; that is, they made "covert retrieval attempts." In our comment on that article, we (Spellman & Bjork, 1992) argued that some of the superiority of the delayed cue-only condition might be due to a self-fulfilling prophecy — because covert retrieval attempts could have two important, if unintended, consequences (see Figure 1).

The first consequence is strategic: Participants use the outcome of the covert retrieval as a basis to predict future recall on the final test. That is, if they fail at covert retrieval on the JOL trial, they are likely to assume that they will fail again on the

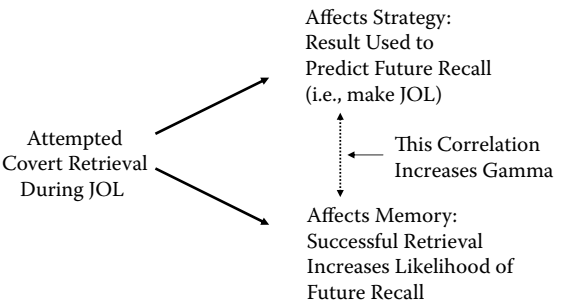

Figure 1 The hypothesized consequences of making a delayed cue-only JOL (Spellman & Bjork, 1992).

distant final recall test; thus, they will give those items a low JOL rating. If they succeed at the covert retrieval, they are likely to assume that they will succeed again on the final recall test, so they will give those items a much higher JOL rating. Evidence for this consequence comes from a different pattern of use of the JOL scale in the delayed cue-only condition (see, e.g., Dunlosky & Nelson, 1992; Kelemen & Weaver, 1997; Kimball & Metcalfe, 2003; Nelson & Dunlosky, 1991; Weaver & Kelemen, 1997). Evidence also comes from studies in which participants are asked to *explicitly* recall the target item when presented with the cue item immediately *before* making the JOL (the PRAM method—pre-judgment recall and monitoring—for studying JOLs developed by Nelson, Narens, & Dunlosky, 2004). When participants make such explicit pre-JOL retrievals they (1) give much higher JOLs to retrieved items than to nonretrieved items (Koriat & Ma'ayan, 2005) and (2) show the same overall pattern of use of the JOL scale as participants who are not instructed to make the explicit retrieval attempts (Nelson et al., 2004).

The second consequence of a covert retrieval is memorial. The act of retrieval is itself a learning event in the sense that the retrieved information becomes more recallable in the future than it would have been otherwise (e.g., Bjork, 1975). A successful retrieval attempt on a JOL trial, therefore, will increase the probability that the judged item is indeed recalled on the later test (Dougherty, Scheck, Nelson, & Narens, 2005; Kelemen & Weaver, 1997; Kimball & Metcalfe, 2003). In other words, by the very act of trying to assess memory, we have changed memory. We argued that those two consequences, and the correlation between them, could account for the superior JOLs in the delayed cue-only condition.

Using Gamma to Assess Learning

We asserted that JOLs in the delayed cue-only condition are far superior to those in the other conditions. But, what do we mean by superior? One way in which judgments could be superior is measured by *calibration*, which is an absolute measure of accuracy. A perfectly calibrated person would, for example, recall none of the items to which she gave a JOL of 0; 20% of the items she gave a JOL of 20; and so forth. In fact, participants in the delayed cue-only condition are better calibrated than in the

other conditions (e.g., Nelson & Dunlosky, 1991). However, most JOL studies have focused on relative accuracy (or resolution), as measured by the Goodman-Kruskal γ correlation (or just γ).

The Goodman-Kruskal γ correlation provides a measure of participants' ability to detect which items are more likely to be remembered than which other items. The γ correlation has become the standard index of JOL accuracy, due in large part to Nelson's (1984) extensive review and analysis of the potentially useful statistics and his ultimate endorsement of γ. He wrote: "Of these measures … the Goodman-Kruskal γ correlation seems best" (p. 124).[1]

Note that γ correlates two observables: JOL ratings and memory performance. Ideally, however, researchers are interested in something unobservable: how well an item was learned in the first place.[2] The problem, as we mentioned, is that in trying to measure learning we might change learning. In fact, we believe that the relatedness of the strategic and memorial consequences of covert retrieval can inflate γ for people who are *not* perfect judges of what they know above what it would be for people who *are* perfect judges of what they know.

Consider, for example, a participant who has learned two pairs of words, with pair A–A′ having been learned slightly better than pair B–B′. When making delayed cue-only JOLs, the participant covertly attempts to retrieve the target word from each pair. Assume, given the probabilistic nature of recall, that the person succeeds at retrieving B′ but not A′ and so, incorrectly, gives B–B′ a higher JOL rating. The successful retrieval of B′ (at a delay) increases the strength of B–B′ in memory, and B′ becomes not only more likely to be recalled on the final test than it was before, but also probably more likely to be recalled than is A′. At final test, B′ might be recalled when A′ is not. Thus, even though the participant was incorrect at assessing the initial relative learning of A–A′ and B–B′, it can appear as if the participant's relative JOLs were accurate. Therefore, as Spellman and Bjork (1992) argued, delayed cue-only JOLs are "predictions [that] create reality."

Chapter Outline

In this chapter we present a mathematical simulation of (what we believe to be) the effects of making a JOL. We show that participants who are less accurate at judging their true state of learning could appear to be more accurate at making JOLs when they base their JOLs on the success or failure of their covert retrieval attempt at the time of the JOL. We examine how much of the improvement in JOL accuracy might be due to the changed use of the JOL scale at a delay and how much might be due to the benefits of successful retrieval. We also use the simulation to illustrate some unsavory properties of the γ statistic and describe experimental design techniques that can help get the most stable γs.

First, we describe a hypothetical participant called the *perfectly insightful participant* — that is, someone who knows exactly what he or she knows — and we illustrate why γ is not "perfect" (i.e., does not equal 1) for such a participant. Second, we introduce our simulation in general terms and describe its assumptions and

implementation. Finally, we present the results of hundreds of simulation runs relevant to the issues mentioned.

Evaluation of the Perfectly Insightful Participant Using Gamma

Someone who is perfect at judging his or her initial learning will not generally obtain a γ of 1. Gamma is calculated by comparing performance for each item to performance for each other item and counting up concordances and discordances. A *concordance* occurs when an item with a JOL that is higher than that of another item is recalled while that second item is not recalled. A *discordance* occurs when an item with a JOL that is higher than that of another item is *not* recalled while that second item *is* recalled. Thus, there is no reference to absolute performance; γ is all about judging relative performance.

The γ correlation is computed as follows:

$$(\text{Concordances} - \text{Discordances})/(\text{Concordances} + \text{Discordances})$$

Note a very important consequence of the definition: Pairs of items that are given identical JOLs and pairs of items that are either both recalled or both not recalled do not contribute to this statistic.[3] Many, sometimes even most, potential comparisons can therefore be irrelevant to the computation of γ.

Consider someone who is perfectly calibrated. Assume further that such a person has learned a list of 60 words with 10 each having a probability of recall of 0, .20, .40, .60, .80, and 1, and that there are not any consequences of making a JOL. In a JOL experiment, then, such a perfect person would then assign JOLs of 0%, 20%, 40%, 60%, 80%, and 100% to the items of each kind, respectively, and at the time of the final test, this person will also recall 0, 2, 4, 6, 8, and 10 items in each JOL category. What is γ for such a "perfect" performance? Because this perfect person sometimes assigns a low JOL to an item that does get recalled (e.g., two of the JOL = 20 items) and a high JOL to an item that does not get recalled (e.g., two of the JOL = 80 items), there are some discordances, and γ is not a perfect 1. For the perfectly calibrated person in this example, γ is .84 — high, but certainly not perfect.

Simulation Overview

The simulation is designed to model participants in an experiment in which they make delayed cue-only JOLs. Readers are encouraged to use the simulation as they read the chapter. (It can be found at http://people.virginia.edu/~bas6g/metamemory. To view all the features described in this chapter, use the "verbose" setting.)

The simulation first generates an initial learning distribution for the items in the study based on a mean, a standard deviation (*SD*), and the number of items entered by the user. During each run, the program simulates two different types of participants.

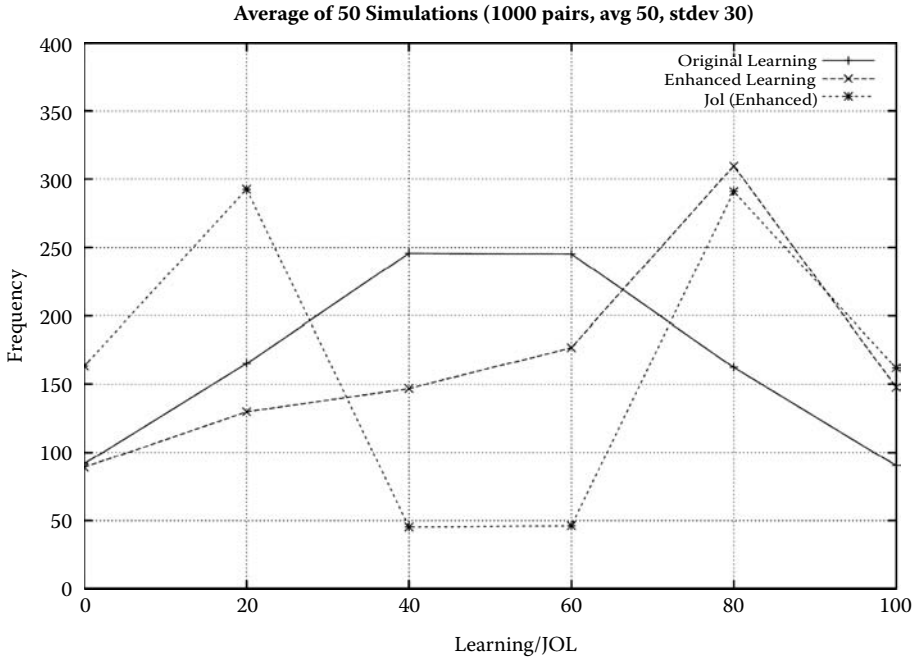

Figure 2 A graph taken from the simulation Web site. The solid line shows initial learning (identical to "perfect" JOLs) and is shown in the Web site in red. For this simulation, the mean is 50, and the standard deviation is 30. The long dashed line (on Web site in green) shows enhanced learning as a result of successful covert retrieval with $d1 = 2$ (moderate learning). The short dashed line (on Web site in blue) shows enhanced JOLs with $d2 = 1.8$ (medium-size scale shift).

- *Perfectly insightful participants.* The JOLs for such participants are exactly equal to the original learning. That is, such participants are assumed to be perfectly accurate assessors of what they know. In addition, the act of making a JOL is assumed to have no consequences for either their actual judgment (i.e., the JOL equals the learning) or the learning of the items.
- *Enhanced participants.* The JOLs are *not* exactly equal to the initial learning. Rather, the act of making a JOL is assumed to have two consequences: (1) a strategic consequence in which such participants draw on the success or failure of covert retrieval attempts to revise their JOLs up or down with respect to their original learning; and (2) a memorial consequence via which the learning of items that were successfully retrieved increases, resulting in such items becoming more likely to be recalled at final test. Simulation users have some control over the functions that modify the shift in JOLs and the learning consequences of successful retrieval.

The simulation presents graphs of the initial learning (red), enhanced learning (green), and enhanced JOLs (blue) (see Figure 2). It computes γs for the perfectly insightful condition and for the enhanced condition (plus two other γs described here). Finally, it gives averages over repeated runs.

Simulation Assumptions and Implementation

Original Learning

The simulation generates a normal distribution for original learning with a mean and standard deviation set by the user. For each simulated participant, the program can simulate the learning of up to 1,000 paired associates. Each pair is represented by a pair number (Simulation Column 1) and has an original learning "strength" from 0 to 100 (Simulation Column 2). This simulation treats recall as probabilistic and an item's strength as reflecting its probability of recall (times 100 for convenience). Items from the generated normal distribution with values greater than 100 are set equal to 100, and those with values less than 0 are set equal to 0. The user can enter a mean (from 0 to 100), a standard deviation, and the number of pairs learned.

For purposes of graphing the original learning (red line), the learning values are placed into six bins: 0–10, 10–30, 30–50, 50–70, 70–90, and 90–100. We selected six to correspond to the number of judgments allowed in most of the early JOL experiments (i.e., participants could make JOLs of 0, 20, 40, 60, 80, or 100; see, e.g., Dunlosky & Nelson, 1992; Kelemen & Weaver, 1997). In some studies, participants, when asked to make a JOL, can respond with any number from 0 to 100 inclusive to represent their estimated probability of recall (Koriat and colleagues tended to use that technique; see, e.g., Koriat & Bjork, 2005; Koriat & Ma'ayan, 2005). In still other studies, the choices were limited to the range of a rating scale (e.g., 0–10, as in Son & Metcalfe, 2005; we address the effects of the choice of JOL scale in Simulations 3 and 4).

Note that all conditions begin with the identical learning strength distribution; that is, initial learning is equated across conditions.

JOLs from Perfect Participants

For participants with perfect insight, JOLs for each item are exact matches to their initial learning. For these participants, the act of making the JOL has no consequences for the JOL or for learning, meaning that their JOLs have the exact same distribution as the initial learning (red line). Thus, the JOLs will be normally distributed because the initial learning is normally distributed. Unlike learning, however, JOLs are observable. Several experiments demonstrated that immediate JOLs are more or less normally distributed (Dunlosky & Nelson, 1994, Experiment 1; Nelson et al., 2004; Weaver & Kelemen, 1997). For purposes of computing γ in most of our simulations, we left the JOLs at their original values (that is, any rational number from 0 to 100 inclusive).

JOLs from Enhanced Participants

Enhanced participants are assumed to make a covert retrieval attempt at the time of JOL. The simulation determines whether that retrieval attempt succeeds and then

TABLE 1 Examples of the Calculations for Revising Strength and Judgments of Learning (JOLs) as a Function of Initial Strength and JOL Retrieval Success (Assuming Default Values $d1 = 2$ and $d2 = 1.8$)

Word Pair (Column 1)	Original Learning (Column 2)	JOL Success? (Column 4)	Enhanced Learning (Column 5)	Enhanced JOL (Column 10)
Pair 1	38	No	38	17
Pair 2	38	Yes	69	72
Pair 3	52	No	52	23
Pair 4	52	Yes	76	79
Pair 5	62	No	62	28
Pair 6	62	Yes	81	83
Pair 7	76	No	76	34
Pair 8	76	Yes	88	89

Note: Column numbers in parenthesis refer to the Web simulation (use the "verbose" setting to view them there). Note that although Pair 2 is learned worse than Pair 3, it is covertly retrieved at JOL, whereas Pair 3 is not. Pair 2 therefore is (incorrectly) given a higher JOL. Because successful covert retrieval also increases the item's learning, Pair 2 is more likely to be recalled than Pair 3 at final test. If that happens, the participant looks correct (i.e., rated Pair 2 higher than Pair 3 and recalled the former but not the latter) but was actually incorrect in judging learning. In the simulation, column 4 reads 0 or 1 which means "no" or "yes," respectively.

modifies the learning and the JOL accordingly. Table 1 gives examples of how the modification works.

Random Value 1 (Simulation Column 3) and Recall at JOL (Simulation Column 4) For the covert retrieval at JOL, a word pair with an original learning strength of, say, 28, will be retrieved 28% of the time; one with a strength of 57, 57% of the time; and so forth. To implement that probabilistic retrieval, for each word pair a random number from 0 to 100, inclusive, is generated from a flat distribution. This random number is compared to the original learning: If the random number is smaller than the original number, the word is assumed to be retrieved at JOL (and gets a 1 in Column 4); if the random number is larger, then it is assumed not to be retrieved at JOL (and gets a 0 in Column 4).

Enhanced Learning (Simulation Column 5) One of the consequences of making a JOL is to increase the strength of a successfully retrieved target above its original learning. It has been shown that making a delayed cue-only JOL has consequences for the memorability of the items; we have unpublished data showing that JOLs are like tests in that they (1) enhance recall above that for pairs given only a single study opportunity and (2) mitigate forgetting over time (see Roediger & Karpicke, 2006, for a review of testing effects). The mitigation effect has been seen in both cued recall and recognition measures (see also Dougherty et al., 2005; Kelemen & Weaver, 1997; Kimball & Metcalfe, 2003).

In the simulation, the form of the increase for successfully retrieved items is

$$\text{Enhanced learning} = \text{Original learning} + (100 - \text{Original learning})/d1$$

If items are not successfully retrieved, then original learning is unchanged. Using this type of function (a delta learning rule function), weak items that are successfully retrieved benefit more than do strong items that are successfully retrieved. The minimum *d*1 is 1, which would set learning of all retrieved items to 100. The default is set at 2 because at typical delays between JOL and final recall, the benefit of a successful JOL is only moderate.[4] The effect of enhanced learning can be seen in the Enhanced Learning column of Table 1 and in Figure 2.

Enhanced JOL (Simulation Column 10) Another consequence of making a delayed cue-only JOL, compared to an immediate one, is a shift in the use of the JOL scale. When participants make immediate JOLs, they tend to use the middle of the JOL scale; when they make delayed JOLs, they more often use the ends of the JOL scale (see Dunlosky & Nelson, 1994; Kimball & Metcalfe, 2003; Nelson et al., 2004; Weaver & Kelemen, 1997). Using a Monte Carlo simulation, Weaver and Kelemen showed that some of the improvement in γ for delayed JOLs is a consequence of that shift in distribution.

In our simulation, the JOL increases if the target was recalled and decreases if it was not. The form of the function is

If recalled: Revised JOL = Original learning + (100 − Original learning)/*d*2

If not recalled: Revised JOL = Original learning − Original learning/*d*2

These functions are presented in the same form as the one for enhancing learning, but there is a more intuitive way of thinking about the JOL functions. Suppose that if an item is retrieved at JOL, the participant first considers giving a JOL of 100 but then modifies that extreme JOL downward by a sense of how well the item had been originally learned. Similarly, suppose that if an item is *not* retrieved at JOL, the participant first considers giving a JOL of 0 but then modifies that extreme JOL upward by a sense of how well the item had been originally learned. Consistent with the notion of adjusting JOLs based on more than just retrieval success or failure, there is evidence that the reaction times for very low and very high JOLs are made fastest, and those in the middle are made slowest (Son & Metcalfe, 2005; but see Kelemen & Weaver, 1996). In that case, the revised JOLs would look like

If recalled: Revised JOL = 100 − Some fraction of (100 − Original learning)

If not recalled: Revised JOL = 0 + Some fraction of original learning

To use the same *d*2 parameter as above, the equations (which now look less intuitive) would be

If recalled: Revised JOL = 100 − (*d*2 − 1)/*d*2 * (100 − Original learning)

If not recalled: Revised JOL = 0 + (*d*2 − 1)/*d*2 * Original learning

In general, these functions give a U-shape pattern to the JOLs, which is consistent with data for delayed JOLs (see Dunlosky & Nelson, 1994; Nelson et al., 2004; Weaver & Kelemen, 1997). The default is set at 1.8 because it tends to give a U shape over a range of learning parameters. It would, of course, be possible to have asymmetric revisions up and down after covert retrieval success and failure, respectively, by using two different *d*2s.

The effect of enhanced JOLs can be seen in the Enhanced JOL column of Table 1 and in Figure 2.

Final Recall

To determine whether final recall succeeds, each pair's strength is compared against a random number.

Random Value 2 (Simulation Column 6) As for Random Value 1, for each word pair, a random number from 0 to 100, inclusive, is generated from a flat distribution. This random value is used to determine recall for both conditions, thus matching them on "memory ability."

Final Recall Perfect Condition (Simulation Column 9) Random Value 2 is compared to original learning (Column 2): If the random number is smaller than the original learning, the word is recalled (and gets a 1 in Column 9); if the random number is larger than the original learning, then it is not recalled (and gets a 0 in Column 9).

Final Recall Enhanced Condition (Simulation Column 12) Random Value 2 is compared to enhanced learning (Column 5): If the random number is smaller than the enhanced learning, the word is recalled (and gets a 1 in Column 12); if the random number is larger than the enhanced learning, then it is not recalled (and gets a 0 in Column 12).

Note that because some pairs in the enhanced condition were strengthened by the covert retrieval practice at JOL, recall in the enhanced condition must be greater than or equal to recall in the perfect condition.

Computing Gamma

The simulation computes four different γs; the two of major interest are the perfect and enhanced conditions (see Table 2).

Perfect Condition To compute γ for the perfect condition, the simulation uses the perfect JOL (which was equal to the original learning) and the outcome of the final recall. This γ and this JOL are for perfectly insightful participants.

TABLE 2 Four Different Gammas Computed by the Simulation

	Learning/Recall	
JOL	Original (Columns 2 and 6)	Enhanced (Columns 4 and 8)
Perfect (Column 5)	Perfect condition	Learning-only condition
Enhanced (Column 7)	Shift-only condition	Enhanced condition

Note: Column numbers in parenthesis refer to the Web simulation.

Enhanced Condition To compute γ for the enhanced condition, the simulation uses the enhanced JOL and the outcome of the enhanced final recall. Note that for each pair, if the covert recall at JOL was successful, both of these numbers are above those in the perfect condition; however, if the covert recall was not successful, learning is the same, but the JOL is lower than in the perfect condition. The two other γs of interest represent conditions in which the covert retrieval at the time of JOL has only one of the two hypothesized effects.

Learning-Only Condition The learning-only condition assumes that in response to covert retrieval attempts at the time of JOL, participants do *not* revise their JOLs but *do* increase the strength of successfully retrieved items. Although we know that JOLs are in fact shifted at a delay, this condition allows us to examine the contribution of the (hypothesized) strength increase alone.

Shift-Only Condition The shift-only condition is the "opposite" of the learning-only condition: It assumes that in response to covert retrieval attempts at the time of JOL, participants *do* revise their JOLs but do *not* also increase the strength of successfully retrieved items. Weaver and Kelemen (1997) demonstrated that some of the increase in γ in the delayed cue-only condition is due solely to the change in use of the JOL scale from a somewhat normal distribution to a U-shape distribution.

Simulations

Simulation 1: Varying the Mean and Standard Deviation of Original Learning

Simulation 1 varies the two parameters of the original learning (normal) distribution: the mean and the standard deviation. One desirable property of a metacognitive measure is insensitivity to level of memory performance (Nelson, 1984); this insensitivity allows comparison of *metacognitive* performance across groups with a *memory* performance that might differ (e.g., young and elderly; see Schwartz & Metcalfe, 1994). We chose means of 50 (the center of the distribution) and 20 and 80 (representing difficult and easy items, respectively). Although 20 and 80 are symmetrical about 50 and therefore it seems as if they should show equal effects, the function for increasing strength after a successful covert retrieval makes them differ. For standard deviations, we chose 10 (a narrow distribution) and 30 (a wide distribution somewhat mirroring immediate JOL use).

Note that when discussing differences across simulations, standard inferential tests do not make sense because we could easily run large numbers of simulated participants, get very small standard errors, and find significant results.

Effect of Varying the Standard Deviation of the Learning Distribution Varying the standard deviation of the learning distribution has a huge effect on γ (see Figure 3). In going from a standard deviation of 10 (top panel) to one of 30 (bottom panel), γ substantially increased; bigger standard deviations lead to bigger γs. In addition, standard deviations of γ across simulations (i.e., the equivalent of experiments) were bigger for the narrow learning distribution than for the wide one. Both of these effects point to the importance of having not only study items that vary in difficulty but also sets of items with equal variability if comparing across different stimuli. Thus, the range of item difficulty can have effects both for estimating the calibration of individual participants and for comparing across participants, conditions, or experiments (Schwartz & Metcalfe, 1994).

Effect of Varying the Mean of the Learning Distribution Varying the learning mean affected γ, although less so than varying the standard deviation. The learning mean of 50 had the lowest γs; changing the mean to 20 or 80 increased γ between .12 and .15, with the one exception described here. Why should the middle of the scale have the lowest γ? We suspect it is because when there are lots of items at the extremes (very poorly or very well learned), those items will behave as expected at final recall — and hence contribute a substantial number of concordances to the γ equation. Items in the middle are less predictable regarding whether they will or will not be recalled at final test and therefore create more discordances, decreasing γ. Note that if γ starts out positive, adding an equal number of concordances and discordances *decreases* γ. For example, suppose that there are 6 concordances and 4 discordances; γ is then

$$\frac{\text{concordances} - \text{discordances}}{\text{concordances} + \text{discordances}} = \frac{6-4}{6+4} = \frac{2}{10} = .20$$

However, if an item or items then contribute both one more concordance and one more discordance, γ becomes

$$\frac{7-5}{7+5} = \frac{2}{12} = .17$$

The exception to the general effect of varying the mean is going from a mean of 50 (medium) to 80 (easy) in the enhanced condition. For that condition, when the mean is 20 or 50, a successful covert retrieval results in a lot of learning, spreading out the learning distribution substantially. However, with a learning mean of 80, there is not much "spreading" left to be done; therefore, the enhanced condition looks like some of the other conditions.

Comparing Conditions Across all parameters, JOLs are better in the enhanced condition than all three other conditions — including the perfect condition. Thus,

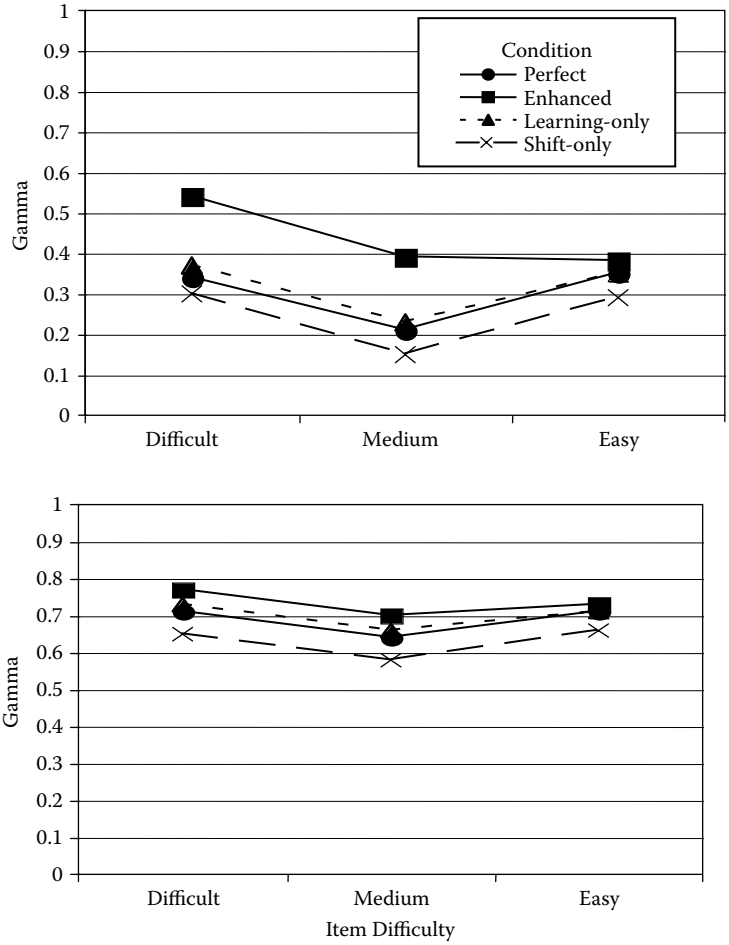

Figure 3 The γs change with changes in mean and variability of learning distribution. Item difficulty refers to means of learning distribution: Difficult = 20; Medium = 50; Easy = 80. Item variability is low in the top panel (standard deviation [*SD*] = 10) and high in the bottom panel (*SD* = 30). (For 50 simulated runs with 100 items each. Learning parameter = 2 [moderate]; JOL-shift parameter = 1.8 [medium]).

revising the JOL and learning in tandem causes an increase in γ. Participants who are worse judges of initial learning (because their JOLs do not equal their initial learning) are better predictors of what they will remember in the future than are the perfectly insightful participants — and therefore have higher γs.

What of the conditions in which the covert retrieval at JOL has only one consequence? When only learning changes, γs are nearly the same as in the perfect condition. When only the JOL distribution changes, γ decreases. The latter effect is surprising and a contrast to the simulation results of Weaver and Kelemen (1997). Our main hypothesis for this result has to do with two differences between the simulations. The first is the simulation of the use of the JOL rating scale: In our simulation, JOLs were rational numbers from 0 to 100, whereas in Weaver and Kelemen's study the JOLs

were the same as used by the participants (0, 20, 40, 60, 80, 100). In Simulations 3 and 4, we demonstrate how restricting the number of JOLs can artificially inflate γ.

The second difference has to do with the way items are given JOLs. In our simulation, JOL assignment depends on the item's original learning strength. In the perfect and learning-only conditions, the JOL is equal to the original learning; in the shift-only and enhanced conditions, the JOL is revised based on whether the item was retrieved during the covert retrieval at the time of JOL. Thus, an item with an original learning of 20, that randomly is covertly retrieved at JOL, is given a JOL of about 60. If such an item is not recalled at final test (as it probably would not be in the shift-only condition because it still has only a 20% chance of being recalled), many discordances result, reducing γ.

Weaver and Kelemen's approach was quite different. First, they assigned JOLs to items by using the JOL distributions generated by participants in an experiment. So, for example, if participants used a particular JOL rating 20% of the time, then .2 of the items were randomly assigned to that JOL. To determine whether an item was recalled, they used the participants' conditional probability of recall for each JOL. So, for example, if 52% of items with a JOL rating of 40 were recalled by participants at final test, then 52% of the items with JOLs of 40 were randomly assigned to be recalled in the simulation. They could then compare what happens to γ when using the conditional probabilities of either immediate or delayed JOLs and crossing that with the JOL rating distribution of either the immediate or delayed JOLs. Using the probabilities from the delayed JOL condition, they found an increase from .73 to .93 in γ when moving from the immediate to delayed JOL distribution. Of course, those conditional probabilities already have built in (we would argue) the enhanced learning as the result of covert retrieval in the delayed condition.

Simulation 2: Varying the Size of the Consequences of Covert Retrievals at JOL

In our second simulation, we vary the consequences of the covert retrievals for both learning and JOLs (see Table 3).

Effects of Changing the Learning Parameter (d1) Changing the learning parameter *d*1 affects only the learning-only and enhanced conditions, that is, only the conditions in which original learning is modified by successful covert retrieval at JOL. When *d*1 = 1, a successful covert retrieval changes learning to 100, which guarantees recall on the final test; that is, *d*1 = 1 simulates maximal learning. A *d*1 of 2 simulates moderate learning and of 4 simulates minimal learning. When *d*1 and *d*2 each equal 1, which makes JOLs either 0 or 100, items successfully covertly retrieved will get JOLs of 100 and will definitely be recalled at final test, thus creating a γ of 1.

Effects of Changing the JOL Shift Parameter (d2) The JOL shift parameter (*d*2) defaults to 1.8, which indicates a moderate shift in JOL use. If *d*2 is set to 1, JOLs become extreme (either 0 or 100); if *d*2 is set to 2.5, JOLs are shifted only slightly as a result of covert retrieval success or failure. In this simulation, if the JOL distribution is shifted, it does not matter how much it is shifted because (1) items are shifted as a

TABLE 3 Mean (and Standard Deviation [*SD*]) of Gammas for 50 Simulated Runs With 100 Items Each and Varying Size of Consequences of Covert Retrievals at Judgment of Learning (JOL)

Parameters				Condition			
Learning Mean	Learning SD	*d1* (Learning)	*d2* (JOL)	Perfect	Enhanced	Learning Only	Shift Only
50	30	1	1.0	.64 (.08)	1.00 (0)	.73 (.06)	.57 (.13)
50	30	1	1.8	.63 (.08)	.89 (.03)	.72 (.07)	.56 (.08)
50	30	1	2.5	.62 (.10)	.89 (.04)	.72 (.07)	.57 (.10)
50	30	2	1.0	.63 (.08)	.78 (.10)	.64 (.07)	.53 (.14)
50	30	2	1.8	.63 (.08)	.69 (.07)	.65 (.09)	.56 (.09)
50	30	2	2.5	.63 (.07)	.69 (.07)	.65 (.08)	.57 (.08)
50	30	4	1.0	.64 (.09)	.68 (.11)	.64 (.09)	.56 (.15)
50	30	4	1.8	.62 (.09)	.63 (.09)	.63 (.08)	.56 (.10)
50	30	4	2.5	.64 (.07)	.64 (.08)	.65 (.07)	.58 (.08)

Note: When $d1 = 1$ a successful covert retrieval changes learning to 100, thus guaranteeing recall at final test (maximal learning); $d1 = 2$ simulates moderate learning (simulation default value); $d1 = 4$ simulates minimal learning. When $d2 = 1.0$, JOLs become extreme (either 0 or 100); if $d2 = 1.8$, JOLs shift as in many delayed JOL studies (simulation default value); if $d2 = 2.5$, JOLs shift only slightly.

function of their current strength, and (2) γ measures relative accuracy. So, if Item Q is recalled at final test and Item R is not, it does not matter whether their JOLs are 57 and 36, respectively, or 81 and 74, respectively; they will still produce a concordance.

Comparing Conditions Of course, changing these parameters has no effect on the perfect condition because that condition enjoys neither of the consequences of covert retrievals at JOL. The enhanced condition has the highest γ when learning is more than minimal (when $d1 = 4$, the enhanced learning distribution moves very little). The shift-only condition again has the lowest γ.

Simulation 3: Varying the Number of JOL Ratings

Varying the number of JOL ratings that participants can use affects γ in several ways (see Figure 4). First, in almost all conditions, reducing the number of JOL ratings increases γ. The effect was particularly strong in the mean = 20, standard deviation = 10 condition (top left panel), in which, for example, the γ in the perfect condition increased by .16. Second, reducing the number of JOL ratings increases the variability of γ, particularly when the standard deviation is small (top panels).

These effects occur because of how γ deals with "ties." Ties occur when two items are given identical JOL ratings or have the same recall status.

Ties reduce the stability of γ in the following way: Suppose participants study N word pairs. When each pair (its JOL and its recall) is compared to every other pair, there are $(N * (N - 1))/2$ comparisons. However, not every comparison results in a concordance or discordance. If two items are both recalled, they produce neither; if

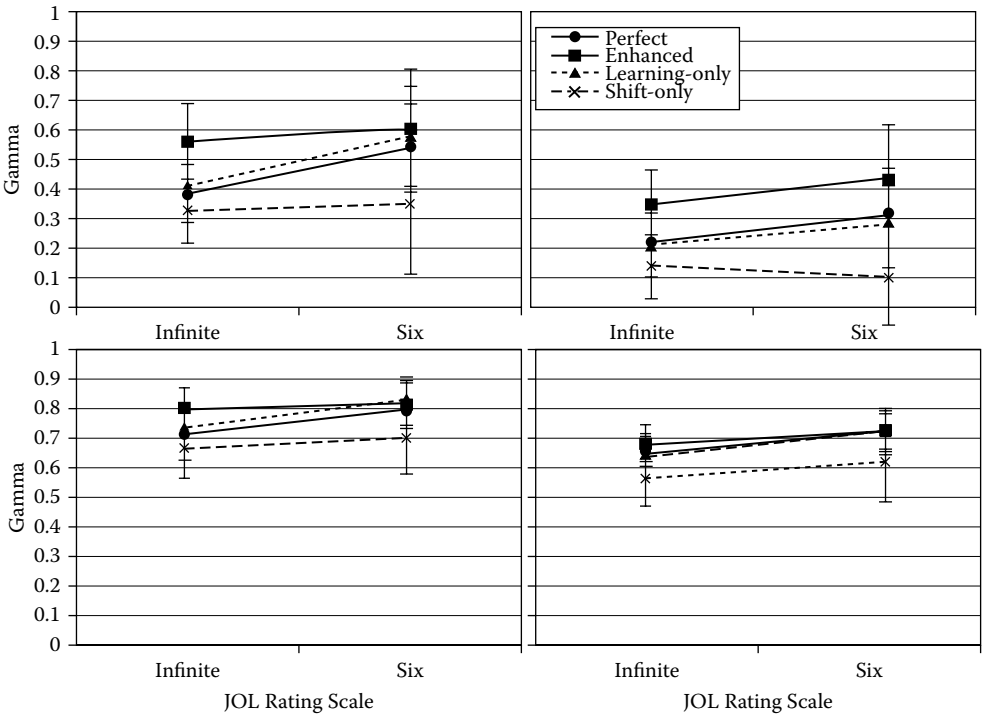

Figure 4 The γs change with different numbers of possible JOL ratings. The mean of the learning distribution is 20 (difficult) in the left panels and 50 (medium) in the right panels. The variability is low in the top panels (standard deviation [*SD*] = 10) and high in the bottom panels (*SD* = 30). Infinite is any rational number from 0 to 100; six places each JOL into a bin of 0–10, 10–30, 30–50, 50–70, 70–90, or 90–100.

two items are both not recalled, they produce neither; if two items are given the same JOL rating, they produce neither. Suppose that half of the items are recalled at final test. Now, the maximum total number of comparisons that could result in a concordance or discordance is

$$(½N * (½N - 1))/2 + (½N * (½N - 1))/2$$

a much smaller number. (For example, if $N = 10$, the equation on top yields 45; the equation on the bottom yields 20.)

When JOL ratings are rational numbers (as generated in our simulations), ties in ratings are unlikely or uncommon. When the JOL scale is limited to 0, 20, 40, and so on or to a 0–10 rating scale, ties are frequent.[5] Increasing the number of options on a scale should decrease the number of ties.

With a limited JOL scale (especially when the standard deviation of learning is small) γ becomes more variable because there are many "tied" JOLs, so γ is based on fewer concordances and discordances and is therefore less stable. With a limited JOL scale, γ becomes inflated because, with a larger scale, items that are close in JOL rating but differ in recall will produce many discordances; however, when the scale is

limited, those items will receive the same JOL rating and will not contribute discordances. (And, consistent with these remarks, reducing the number of discordances causes a bigger increase in γ than increasing concordances by the same number.)

Simulation 4: Varying the Number of Study Items

It is, of course, a general rule in experiments to try to get as many observations as possible from each participant. This advice is particularly important when computing γ because, as described, so many potential comparisons are thrown away due to ties in JOL ratings or recall status. Table 4 shows the effects of varying the number of items studied by each participant. Note the huge standard deviations with only 15 observations, especially with a narrow learning distribution (e.g., $SD = 10$). Remember that in a within-subject design, if a participant studies 60 words but the pairs are in four conditions, γ is being computed on (at best) only 15 observations per cell. Note also that, as in Simulation 3, γ generally continues to be higher when the number of ratings is limited.

Discussion

Across variations in many parameters, the enhanced condition, in which covert retrieval at the time of JOL affects both learning and JOL, produces the highest γ, even higher than those for our hypothetical perfectly insightful participant. These high γs do not result when only learning is enhanced or when only JOLs are shifted; rather, they result from the correlation between the two consequences of successful covert retrieval.

Other Factors

Our simulation, of course, does not take into account all factors that could affect γ. For example, we have intentionally left out forgetting from the simulation. How forgetting is modeled could affect the different γs in different ways. One way to model forgetting would be to decrease the learning of all items by the same amount; another would be to decrease the learning of all items by the same percentage. As long as the relative probability of recall of different items does not change, γ should not change (except at very low recall rates in which γ relies on very few concordances and discordances). Another way to model forgetting would be to have some probabilistic forgetting function. Again, however, if that function only inverted learning strengths of a few items, γs might decrease and become more variable, but the conditions should remain relatively the same. Finally, any of those could be implemented but with the addition of different forgetting rates for items that were or were not successfully retrieved at JOL. We believe that successful covert retrievals, like successful tests, slow the rate of forgetting. Therefore, JOLs for items that were enhanced based on

TABLE 4 Mean (and Standard Deviation [*SD*]) of Gammas for 50 Simulated Runs Varying Number of Items and Number of Judgment of Learning [JOL] Ratings (Learning Mean = 50; *d*1 = 2; *d*2 = 1.8)

		Condition							
Number of Items		Perfect		Enhanced		Learning Only		Shift Only	
Learning	St. Dev	Infinite	Six	Infinite	Six	Infinite	Six	Infinite	Six
15	30	.63 (.24)	.71 (.23)	.66 (.22)	.73 (.22)	.64 (.23)	.74 (.24)	.53 (.27)	.58 (.28)
60	30	.61 (.12)	.70 (.13)	.68 (.11)	.73 (.11)	.63 (.11)	.72 (.12)	.55 (.12)	.62 (.13)
100	30	.61 (.08)	.69 (.08)	.67 (.08)	.71 (.09)	.63 (.08)	.71 (.09)	.55 (.08)	.59 (.09)
1,000	30	.63 (.02)	.72 (.03)	.69 (.02)	.74 (.02)	.65 (.02)	.74 (.02)	.57 (.02)	.62 (.02)
15	10	.17 (.32)	.19 (.50)	.35 (.28)	.46 (.41)	.19 (.34)	.24 (.55)	.09 (.31)	.06 (.42)
60	10	.20 (.15)	.30 (.25)	.35 (.13)	.44 (.18)	.23 (.15)	.34 (.25)	.11 (.13)	.03 (.24)
100	10	.25 (.10)	.34 (.15)	.40 (.10)	.51 (.15)	.26 (.09)	.36 (.15)	.15 (.11)	.12 (.14)
1,000	10	.24 (.04)	.33 (.05)	.39 (.03)	.48 (.05)	.25 (.03)	.34 (.05)	.15 (.04)	.10 (.06)

successful retrievals will remain more accurate over time because those items will be less affected by the forgetting function.

In some recent studies, participants have been asked to make JOLs over longer intervals, ranging from a day to a week (e.g., Koriat, Bjork, Sheffer, & Bar, 2004). Over such long intervals, forgetting would not be the only function to be modeled; there is also the question of whether and how participants strategically factor in the long delay when making JOLs.

The Trouble With Gamma and Finding Relief

We have seen that γ is sensitive to various parameters, sometimes in expected ways and sometimes in unexpected ways. Because γ is a correlation, it is sensitive to the standard deviation of the learning distribution; small standard deviations (i.e., a "restricted range") reduce γ and increase its variability. Also, γ is very variable when there are a small number of items (e.g., 15) going into its computation. The γ correlation does turn out to be sensitive to the mean of original learning. And, reducing the number of possible JOL ratings participants can potentially make (from 101 to 6) can significantly increase γ. All of these consequences occur, at least in part, because in computing γ ties are not counted.

These problems can be ameliorated to some extent through careful experimental design. Study items should have a range of difficulty within conditions and should be

equally difficult across conditions. As many observations as possible should go into each computation of γ. And, participants should be allowed to use as wide a JOL rating scale as can be practically and sensibly used in the study.

Conclusions

The results of our simulations demonstrate that the superior γs in the delayed cue-only JOL condition need not reflect more accurate assessments of original learning. Rather, inaccurate assessments might lead to accurate predictions when those assessments and actual recall performance are correlated by virtue of both being based on the outcome of covert retrievals at the time of JOL. We believe that such JOLs irretrievably alter the state of learning, thus making accurate assessments of original learning permanently unrecoverable. But, delayed cue-only JOLs do make people much better at something different and, in fact, something more useful — predicting what they will recall in the future.

The γ correlation has flaws. It is important to recognize those flaws and to try to design studies to minimize their effects. At times, it may be important to use other measures, such as measures of absolute accuracy, along with γ's measure of relative accuracy (see also Masson & Rotello, 2008). Despite the troubles with γ, however, we are not convinced it should be discarded. Perhaps Tom Nelson's (1984) true opinion of γ was similar to that of Winston Churchill's opinion of democracy: "Democracy," said Sir Winston, "is the worst form of government except all those other forms that have been tried from time to time."

Notes

1. Note, however, that he compared it to other statistics useful for analyzing 2×2 feeling-of-knowing data. One of γ's good properties, he noted, is that it could be used for tables larger than 2×2, as is done in JOL studies. However, he did not compare γ to the other statistics for larger tables.
2. Although "judgment of learning" does sound as if it should judge the unobservable learning, many have noted that, "Judgments of learning … are predictions about future test performance" (Nelson & Narens, 1994, p. 16).
3. "Gamma was designed to be unaffected by ties" (Nelson, 1984, p. 116; see Gonzalez & Nelson, 1996, for an explanation). Note, however, as we show below, manipulations that affect the proportion of ties will affect γ.
4. Note that the memorial benefits of delayed cue-only JOLs need not show up when compared to delayed cue-targets JOL (which are, in effect, re-presentations). Cue-only JOLs can only help items that can be successfully retrieved at the time of JOL, but as the time from initial presentation to JOL gets longer, that proportion of items decreases. Cue-target JOLs can help all items at all times. The relevant comparisons to see the benefits of delayed cue-only JOLs are (1) items with single presentations (which will be remembered less frequently) and (2) items that are explicitly recalled at delays matching that of the JOLs (which will be remembered more frequently than single presentation items and as frequently as JOL items).

5. Gonzalez and Nelson (1996, p. 162) noted that such ties are ambiguous — they might be intended (the participant might have wanted to give two items ratings of 20), or they might be limited by the (in)sensitivity of the procedure (the participant might have wanted to give the items ratings of, e.g., 18 and 22 but could not because of the scale).

References

Bjork, R. A. (1975). Retrieval as a memory modifier. In R. Solso (Ed.), *Information processing and cognition: The Loyola Symposium* (pp. 123–144). Hillsdale, NJ: Erlbaum.

Dougherty, M. R., Scheck, P., Nelson, T. O., & Narens, L. (2005). Using the past to predict the future. *Memory & Cognition, 33*, 1096–1115.

Dunlosky, J., & Nelson, T. O. (1992). Importance of the kind of cue for judgments of learning (JOL) and the delayed-JOL effect. *Memory & Cognition, 20*, 374–380.

Dunlosky, J., & Nelson, T. O. (1994). Does the sensitivity of judgments of learning (JOLs) to the effects of various study activities depend on when the JOLs occur? *Journal of Memory and Language, 33*, 545–565.

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

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

Kimball, D. R., & Metcalfe, J. (2003). Delaying judgment of learning affects memory, not metamemory. *Memory & Cognition, 32*, 918–929.

Koriat, A., & Bjork, R. A. (2005). Illusions of competence in monitoring one's knowledge during study. *Journal of Experimental Psychology: Learning, Memory, Cognition, 31,* 187–194.

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., & Ma'ayan, H. (2005). The effects of encoding fluency and retrieval fluency on judgments of learning. *Journal of Memory and Language, 52*, 478–492.

Masson, M. E. J., & Rotello, C. M. (2008). Bias in the Goodman-Kruskal Gamma coefficient measure of discrimination accuracy. Unpublished manuscript.

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

Nelson, T. O., & Dunlosky, J. (1991). When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: The "delayed-JOL effect." *Psychological Science, 2*, 267–270.

Nelson, T. O., & Narens, L. (1994). Why investigate metacognition? In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 1–26). Cambridge, MA: MIT Press.

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.

Roediger, H. L., III, & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1*, 181–210.

Schwartz, B. L., & Metcalfe, J. (1994). Methodological problems and pitfalls in the study of human metacognition. In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 93–114). Cambridge, MA: MIT Press.

Son, L. K., & Metcalfe, J. (2005). Judgments of learning: Evidence for a two-stage process. *Memory & Cognition, 33*, 116–1129.

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.

Weaver, C. A., III, & Kelemen, W. L. (1997). Judgments of learning at delays: Shifts in response patterns or increased metamemory accuracy? *Psychological Science, 8,* 318–321.