ABSTRACT

Title of dissertation: THE EFFECT OF PRACTICE WITH TEST ON THE RELATIVE ACCURACY OF JUDGMENTS OF LEARNING

Yoonhee Jang, Doctor of Philosophy, 2006

Dissertation directed by: Professor Thomas S. Wallsten
Department of Psychology

To investigate what aspects of practice increase the relative accuracy of judgments of learning (JOLs), this study manipulated both JOL timing and type of practice. Three experiments examined the hypotheses that practice with test but not without test improves the relative accuracy of JOLs, and that a similar process mediates the effects of both delay and practice. The results of the experiments revealed that practice without test does not increase the relative accuracy of JOLs, but practice with test does, and that this advantage is different from the advantage caused by delay. These results are discussed in the context of the retrieval hypothesis of memory as well as theories of JOLs.

THE EFFECT OF PRACTICE WITH TEST ON THE RELATIVE ACCURACY OF JUDGMENTS OF LEARNING

by

Yoonhee Jang

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2006

Advisory Committee:
Professor Thomas S. Wallsten, Chair
Associate Professor Michael R. P. Dougherty
Assistant Professor David E. Huber
Associate Professor Kent L. Norman
Professor Donald R. Perlis

Dedication

To the memory of my grandmother, Jeongdong Han, who provided physical, emotional, and psychological care.

Acknowledgements

It is a pleasure to acknowledge debts incurred over the years in the preparation of this dissertation. Foremost, my most sincere thanks go to my advisor, Dr. Tom Wallsten whose careful guidance, kind encouragement, and willing assistance helped bring my dissertation to a successful conclusion. I would like to thank my dissertation committee members, Drs Mike Dougherty, Dave Huber, Kent Norman, and Don Perlis for their many valuable comments. I remain deeply indebted to Dr. Tom Nelson who guided and stimulated my thinking about this dissertation until he passed away in January 14, 2005.

I also want to take this opportunity to thank Dr. Heungchul Lee, who kindled my interest in cognitive psychology, for providing all sorts of tangible and intangible support, and Sangceup Yune, who encouraged me to re-start my graduate career, for assisting me in innumerable ways.

Finally, my highest appreciation goes to my parents, Jaeyoung Jang and Younghee Lee, brother, Wonseok Jang, sister, Yunjeong Jang, and sister-in-law, Kyungsuk Han for their support, patience, and trust in me. I also send gratitude separately to my nephew, Junsoo Jang.

Whatever I might say here cannot do full justice to the extent and the value of their contributions.

Table of Contents

List of Figures

Chapter 1: Introduction

How does one know one has studied something sufficiently well to remember it later? Under what circumstances is one's confidence in what one has studied a true measure of what he or she later will remember? Answers to these questions are important to help people learn effectively and are broadly applicable to any educational setting (e.g., learning foreign language vocabulary). Effective learning and remembering depend on the learner's metacognitive skills. The current study focuses on a specific kind of metacognitive monitoring called judgments of learning (JOLs), which refers to an individual's judgments of the likelihood of recalling previously studied items. Research on the topic seeks to understand reasons for the accuracy of JOLs in predicting item-by-item memory performance.

To begin, the distinction between absolute accuracy (or calibration) and relative accuracy (or resolution) should be noted. Absolute accuracy refers to the accuracy of assigning probabilities to the items in terms of the judged likelihood of correct recall (e.g., the correspondence between percent of items recalled and the mean of JOLs). Relative accuracy refers to the accuracy of distinguishing between one item relative to another (e.g., the extent to which JOLs discriminate between recalled and unrecalled items). Some variables influence both kinds of accuracy in the same direction. For example, if there is a delay between study trial and judgment trial, not only does the absolute accuracy of JOLs more or less improve, but their relative accuracy also increases, which is called delayed-JOL effect (Nelson & Dunlosky, 1991). However, other variables increase one of them and decrease the other.

Repeated presentation of the list[1], for example, increases the relative accuracy of JOLs and decreases their absolute accuracy; people tend to underestimate their recall performance with practice, which is called underconfidence-with-practice (UWP) effect (Koriat, Sheffer, & Ma'ayan, 2002). In exploring the issue of JOL accuracy, the present research is limited to consideration of the relative accuracy of JOLs[2] (see Koriat et al., 2002, for concerns of both absolute and relative accuracy). Although there is not yet a consensus explanation for delayed-JOL and practice effects, several theories of the underlying process of JOLs have been proposed that bear on these issues.

Theories of JOLs and the Relative Accuracy of JOLs

Three theories of JOLs that account for effects of both delay and practice will be described. The first two of the theories, described below, were originally developed to account for the delayed-JOL effect, and the third focused more on the practice effect. However, all three theories refer to both effects implicitly or explicitly.

The first theory is the monitoring-dual memories hypothesis (Nelson & Dunlosky, 1991). According to this theory, when one assesses the likelihood of memory performance, the individual monitors the information retrieved from memory

[1] Repeated presentation of the list means that a study phase goes through all of the items, and then another study phase occurs for each item. This procedure is in contrast to massed repetition, in which each item is presented multiple times without any other items or events between the repeated presentations in a study phase.

[2] Since Koriat et al. (2002) first brought the UWP effect to the attention of researchers, many studies have discussed JOL accuracy in terms of absolute accuracy. To avoid confusion, the present research focuses only on relative accuracy and does not discuss absolute accuracy. Hereafter, JOL relative accuracy is called JOL accuracy. As seen in Appendixes, however, the UWP effect did occur in the present data, and the complete results of the absolute accuracy of JOLs are available-upon-request from the author.

about the to-be-judged item. If such information at the time of the JOL is not predictive of eventual memory performance, then JOL accuracy will be low. When a JOL is made immediately after item presentation, the information retrieved from short-term memory (STM) functions as noise for the monitoring of information retrieved from long-term memory (LTM), thus reducing JOL accuracy. By contrast, when the JOL is delayed until the to-be-judged item has left STM, less interference occurs in monitoring item information retrieved from LTM, and therefore JOL accuracy increases.

The monitoring-dual memories hypothesis does not explicitly refer to the practice effect. However, it implies that practice can increase JOL accuracy because this hypothesis suggests that JOLs are based on the retrieval of target, and that practice leads to access of the information about the to-be-judged item which is predictive of future recall.

Second, Spellman and Bjork (1992) offered a different explanation of the delayed-JOL effect. They assumed that individuals covertly attempt target retrieval when making JOLs and argued that the delayed-JOL effect occurs because retrieved items become more retrievable due to the spacing of the retrieval practice whereas unretrieved items do not receive such spaced retrieval practice and become less retrievable. Accordingly, a high correlation between delayed JOLs and recall occurs because of the effect of spaced retrieval practice on subsequent recall, referred to as the self-fulfilling-prophecy hypothesis by Nelson, Narens, and Dunlosky (2004) (or the memory hypothesis by Kimball & Metcalfe, 2003).

The self-fulfilling-prophecy hypothesis suggests that JOL accuracy increases

with practice, presumably due to the increased retrievability of recalled items. This hypothesis predicts that the final recall should be greater (1) for items that previously had delayed JOLs than for items that previously had immediate JOLs (cf. Kimball & Metcalfe, 2003) and (2) for items previously repeated than for items not previously repeated. While the latter has been found widely, the former is not the case. The direction of the difference in recall performance has varied unsystematically; recall was reliably greater after delayed JOLs than after immediate JOLs (e.g., Dunlosky & Nelson, 1994; Koriat & Shitzer-Reichert, 2002); recall did not differ after delayed versus immediate JOLs (e.g., Jang & Nelson, 2005; Nelson & Dunlosky, 1991); and recall was reliably greater after immediate JOLs than after delayed JOLs (e.g., Dunlosky & Nelson, 1992; Meeter & Nelson, 2003).

The third theory is the cue-utilization framework (Koriat, 1997). According to this theory, JOLs are based on a variety of cues, which are more or less predictive of memory performance, and will be accurate to the degree that the cues are consistent with the factors underlying recall. There are two modes of influence on JOLs: (1) the theory (rule)-based influence that entails an analytic deduction based on a priori theory, and (2) the experience (heuristic)-based influence underlying the reliance on mnemonic cues that reflect the degree to which the studied items have been mastered (Koriat, 1997; Koriat, Bjork, Sheffer, & Bar, 2004). With practice learning the same items, the basis for JOLs changes from a theory-based inference towards a greater reliance on experience-based mnemonic cues that can serve as valid cues for JOLs resulting in increased JOL accuracy.

Koriat (1997) also proposed that the delayed-JOL effect is due to the reliance

on mnemonic cues pertaining to the ease with which the target can be retrieved, and that the two effects of delay and practice are mediated by a similar process, the function of mnemonic cues. Indeed, Koriat and Shitzer-Reichert (2002) reported results obtained with children aged 7 through 10 that were consistent with this idea; JOL accuracy did not differ after delay versus with practice.

Empirical results on the effect of practice on JOL accuracy are not fully consistent with theories predicting an increase in accuracy. At the outset, it is imperative to clarify how previous experiments manipulated practice. Koriat (1997, Experiments 1 & 2), Koriat et al. (2002), and Koriat and Shitzer-Reichert (2002) compared a condition that provided practice with study, JOL rating, and test to a condition that did not provide any practice (SJTSJT vs. SJT, where S, J, and T represent study, JOL rating, and test, respectively), and Lovelace (1984) compared a condition that included study and test practice to one that did not (STSJT vs. SJT). Both experiments provided study and test practice and found that JOL accuracy increased as a result. By contrast, Jang and Nelson (2005), Koriat (1997, Experiment 3), and Meeter and Nelson (2003) who all found no practice effect, and Dunlosky and Nelson (1994) in which JOL accuracy even decreased with practice, investigated study-alone practice (SSJT vs. SJT). One possible interpretation of the lack of consensus is that test practice is necessary to achieve a practice effect on JOL accuracy. This explanation is plausible because King, Zechmeister, and Shaughnessy (1980) found that prediction of memory performance was more accurate when individuals were given tests prior to the prediction task.

However, the study by King et al. (1980) has two problems. First, because

they did not include the proper control condition, their results could not ascertain whether practice without test also increased accuracy but to a lesser degree than occurred when test practice was included. Second, King et al. (1980) did not use Goodman-Kruskal gamma, a nonparametric correlation coefficient, which is the measure of relative accuracy used in almost all articles published since the 1980s. This measurement problem also occurs in Lovelace (1984), who reported that practice both with and without test increased accuracy. Gamma does not assume interval scales on either of the variables being correlated; the Likert-type scales that both King et al. (1980) and Lovelace (1984) used should not be assumed to be interval scales. Ties should be excluded when the experimental procedure forces ties to occur; ties occur whenever a *j*-place rating scale is used to rate *k* items, with $j < k$, for example, $j = 6$ and $k = 72$ in King et al. (1980), and $j = 5$ and $k = 40$ in Lovelace (1984). Gamma is unaffected by ties either in the ratings or in the eventual memory performance (Gonzalez & Nelson, 1996; Nelson, 1984).

At least a part of the increase of JOL accuracy that occurs as a result of practice may be due to a delayed-JOL effect, because 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. However, that interpretation does not explain why JOL accuracy increased only when items were repeated with test, but not without test.

Retrieval Practice Effect and Retrieval Hypothesis

Features of study-alone practice and practice with test are fundamentally different in the memory literature. Once some to-be-remembered information is

stored in memory, additional test trials tend to enhance performance more than additional study trials do. Dempster (1996) concluded that any effects of test are not due simply to re-presentation of the recalled item but due to the memory retrieval itself, which is called retrieval practice effect (or testing effect; Glover, 1989).

For example[3], Allen, Mahler, and Estes (1969) found that 5 study trials of paired associates followed by 5 cued-recall tests of the pairs led to better final retention on a cued-recall test one day later than did only 10 presentations. This result suggests that the function of practice with test differs from the function of study-alone practice because the number of total trials is the same for each condition.

Carrier and Pashler (1992) compared two methods of learning paired associates. In the pure study trial method, both items of a pair were presented simultaneously for study. In the test trial/study trial method, people attempted to retrieve the response item during a period in which only the stimulus was present (and the response item of the pair was presented after some delay). The results revealed that there was a reliable advantage on the final cued-recall test of the test trial/study trial condition over in the pure study trial condition. The authors concluded that retrieving an item from memory when tested has beneficial effects for final retention beyond the effects due to just studying the item.

During multi-trial learning, people are generally accurate in distinguishing previously recalled and non-recalled items and can monitor their knowledge of the outcomes of previous tests. Consequently, items can be learned efficiently on

[3] Although many studies in the memory literature have reported that practice with test improves final recall performance more strongly than study-alone practice does, I describe here only the results from cued-recall both because cued-recall was used in this research, and because the processes underlying cued-recall and free-recall differ (e.g., Bregman & Wiener, 1970).

subsequent study trials (Bisanz, Vesonder, & Voss, 1978; Gardiner & Klee, 1976; Klee & Gardiner, 1976; Robinson & Kulp, 1970). Such results suggest that retrieval practice can play an important role in metacognitive judgments. Indeed, JOLs are more strongly correlated with recall on the previous test than with recall on the subsequent test (King et al., 1980; Koriat, 1997; Lovelace, 1984), which suggests that JOLs are based on information pertaining to the outcome of the previous recall. Thus, JOLs may in part constitute postdiction based on memory for remembered items.

The retrieval practice effect can be explained by the retrieval hypothesis in that the number of retrieval events, not just the amount of processing, influences final retention (Bjork, 1988; Dempster, 1996; Glover, 1989). Bjork (1988) suggested that an initial retrieval aids a later retrieval to the extent that it constitutes practice for that later retrieval. An act of retrieval does not simply strengthen an item's representation in memory, but rather enhances some aspect of the retrieval process per se.

The retrieval hypothesis also extends to the practice effect on JOL accuracy. When cued recall is tested immediately after presentation, probability of correct recall is near 1.0 (Tulving & Arbuckle, 1963). However, people are able to judge whether a presently studied item has been learned well enough to be recalled on a later test because items differ in associative strength immediately following presentation, and people can discriminate these differences in item difficulty (Arbuckle & Cuddy, 1969; Underwood, 1964). Presumably, if items are repeated without test, people have little information about item retrievability and instead rely heavily on item difficulty when making JOLs, and their accuracy is related to their ability to assess item difficulty (Koriat, 1997; Koriat et al., 2002). Therefore, their accuracy is not

guaranteed but depends on the correlation between item difficulty and recall. By contrast, if items are repeated with test, people can discriminate recalled from non-recalled items through the intervening test. Then, when making JOLs, they may use the information about the items' retrievability, which is highly predictive of eventual memory performance and leads to more accurate JOLs.

Specific Goals of the Present Research

First, this study involves empirical generalizations of the practice effect on JOL accuracy because the cause of the conflicting results summarized above has not yet been resolved. The hypothesis under test is that intervening test trials in practice are critical for enhancing JOL accuracy. This would be true if the intervening tests improve individuals' abilities to discriminate items that can be recalled from items that cannot, and therefore allow more effective assessments of future performance. This research also investigated whether overt responses through test trials but not covert ones are necessary to increase JOL accuracy. Although memory performance improves with practice only when an overt response is involved (Cohen & Johansson, 1967; King et al., 1980), the differentiation between overt retrieval responses and covert retrieval attempts on JOL accuracy is not known.

Second, this research involves theoretical interpretation of the practice effect. All three theories (monitoring-dual memories hypothesis, Nelson & Dunlosky, 1991; self-fulfilling-prophecy hypothesis, Spellman & Bjork, 1992; and cue-utilization framework, Koriat, 1997) agree that practice may increase JOL accuracy by providing valid information retrieved from memory about the to-be-judged item. However, none of the theories discriminates the effects of practice with versus without test, which has

already been discussed in memory research and potentially is critical in metacognition research. Unless the theories differentiate the role of test practice from the role of study practice, the practice effect may be overlooked.

Finally, the present research tested the hypothesis from the cue-utilization framework that the effects of delay and practice are mediated by a similar process, the function of mnemonic cues pertaining to the fluency with which the target can be retrieved (Koriat, 1997). Little research has tested whether this is so. The single exception is the result obtained with children that the correlation between recall and JOLs did not differ after delay versus with practice (Koriat & Shitzer-Reichert, 2002). However, the increase of JOL accuracy after delay and with practice may not derive from the same source for at least two reasons. First, JOL distributions are different between JOLs after delay and with practice. Delayed JOLs are typically associated with a polarized distribution of JOLs ratings (Dunlosky & Nelson, 1994) such that people tend to use extremely low and high values of the scale more frequently than middle values. For immediate JOLs, by contrast, people tend to use middle values more often than extreme values. Koriat and Goldsmith (1996) also noted that in confidence judgments, polarized distributions tend to be associated with better accuracy. With practice, however, the increase in usage of extreme JOL values is found mainly at the higher values of JOLs, but not at the lower values of JOLs (Dunlosky & Nelson, 1994, Koriat et al., 2002). Second, the type of items that contributes to the increase of JOL accuracy may be different between JOLs after delay and with practice. In the multi-trial learning situation, prediction of memory performance is quite accurate for items recalled correctly on the immediately

preceding trial (Bisanz et al., 1978; Robinson & Kulp, 1970). Thus, items recalled correctly on trial *n-1* would be expected to receive "Yes" predictions on trial *n*, and indeed such predictions, with few exceptions, should likely be accurate because items recalled correctly on trial *n-1* should tend to be recalled correctly again on trial *n* (Vesonder & Voss, 1985). If JOLs are based on such knowledge of items recalled correctly from the previous test trial, it is expected that the increase of JOL accuracy after delay and with practice may derive from different sources for each other because a pair of items in which one is recalled and the other is not at the time of JOLs mainly contribute the delayed-JOL effect (Nelson et al., 2004).

Chapter 2: Experiment 1

Experiment 1 manipulated JOL timing (immediate vs. delayed) and type of practice. The type of practice had 5 conditions. Four of them were as follows: no practice (SJT, the control condition); study-alone practice (SSJT: Dunlosky & Nelson, 1994; Jang & Nelson, 2005; Koriat, 1997, Experiment 3; Lovelace, 1984; Meeter & Nelson, 2003); practice with study and test (STSJT: Lovelace, 1984); and practice with study, JOL rating and test (SJTSJT: Koriat, 1997, Experiments 1 & 2; Koriat et al., 2002; Koriat & Shitzer-Reichert, 2002). These conditions were used in the previous studies separately and in part, but no comprehensive research has manipulated them all and compared immediate with delayed JOLs in one study. The fifth type of practice, which included study and JOL rating (SJSJT), tested whether JOL rating repetition (i.e., covert retrieval attempt) influences JOL accuracy.

Method

Participants. Two hundred and twenty-five undergraduate students at the University of Maryland were recruited and received credit for psychology courses in return for their participation. Forty-five participants were assigned to each of the 5 groups by block randomization. Participants in all experiments were treated in accord with the "Ethical Principles of Psychologists and Code of Conduct" (American Psychological Association, 1992).

Materials. Stimuli consisted of 60 concrete (Concreteness $\geq$ 6.10; norms from Paivio, Yuille, & Madigan, 1968), unrelated noun-noun pairs, consistent with most of the previous research. The first 6 pairs constituted practice, and the last 6 pairs were

excluded from recall so as to prevent recency effects. The remaining 48 pairs comprised two blocks of 24 pairs per block and were the only ones that were analyzed.

Design. The experiment design was a 5 × 2 mixed factorial with type of practice (SJT, SSJT, SJSJT, STSJT, vs. SJTSJT) manipulated between subjects and JOL timing (immediate vs. delayed) manipulated within subjects.

SJT was the control condition, having one study and JOL rating phase followed by one test phase, SSJT differed from SJT by having a repeated study-alone phase, SJSJT differed from the control condition by having added study and JOL rating phases, STSJT had additional study and test phases, and SJTSJT had additional study, JOL rating, and test phases.

Procedure. All participants were instructed to study pairs and to indicate their JOL for each pair when the first word in the pair appeared alone as the cue for the JOL. Pairs destined for immediate and delayed JOLs underwent the same study procedure, but differed at JOL timing. During the study phase, each pair was presented in the center of the screen for 5 sec. Pairs were randomly assigned to the immediate- or the delayed-JOL condition, which was self-paced in both cases. Each immediate JOL occurred immediately after the offset of each pair (i.e., right after a pair was presented for 5 sec). It was prompted with the cue word and the question "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?" The participants reported their estimate on a scale of "0 = *definitely will not recall*, 20 = *20% sure*, 40 = *40% sure*, 60 = *60% sure*, 80 = *80% sure*, and 100 = *definitely will recall*". Delayed JOLs

occurred after the final immediate JOL or study trial of a given block. All pairs were randomized anew for each study phase of all conditions and for each participant.

During the test phase, in which recall was also self-paced, the participants were instructed to type the target word when cued by the first word of the pair. If they had no guess, then they typed *NEXT* so as to proceed to the next test trial. All pairs were randomized anew for each test phase of all conditions and for each participant.

Results

For each experiment, I first briefly report the recall and JOL magnitude results for each condition. I then look at relative accuracy assessed with Goodman-Kruskal gamma, followed by the proportion of items receiving each JOL rating. Throughout, all tests of statistically significant differences use $\alpha = .05$, and estimates of effect size (ES) are reported as partial eta squared for significant effects.

The descriptive statistics and the results from separate two-way ANOVAs on percent correct recall and on JOL magnitude are reported in Appendix A. Using Tukey's honestly significant difference (HSD) test to compare the pooled immediate and delayed JOLs across the five practice conditions, the significance pattern of recall can be represented as 'SJT'< 'SSJT' ≈ 'SJSJT'< 'STSJT' ≈ 'SJTSJT' where ≈ indicates non-significant difference and < indicates a significant directional difference (the results for each level of JOL timing are summarized in Table A1). These results provide evidence of the successful manipulation of type of practice and a replication of Cohen and Johansson (1967) and King et al. (1980) that memory performance improves with item repetitions only when an overt response is involved. The magnitude of immediate JOLs did not differ among the 5 practice conditions, $F(4,$

220) < 1, which is inconsistent with the results of previous studies, in which practice was a within-subject variable (e.g., Dunlosky & Nelson, 1994; Jang & Nelson, 2005; Koriat, 1997).

JOL accuracy in all conditions. Figure 1 shows the mean gamma as a function of JOL timing and type of practice. A two-way ANOVA showed that both main effects were significant, $F(4, 202) = 3.55$, $MSE = .06$, ES = .07; $F(1, 202) = 226.92$, $MSE = .05$, ES = .53, for type of practice and JOL timing, respectively, and that the interaction was significant, $F(4, 202) = 7.02$, $MSE = .05$, ES = .12. All gammas for delayed JOLs were close to ceiling, and as a consequence, the 5 practice conditions did not differ, $F(4, 202) < 1$. Applying Tukey's HSD test to immediate JOLs, the significance pattern can be represented as 'SJT' ≈ 'SSJT' ≈ 'SJSJT' < 'STSJT' ≈ 'SJTSJT'. These results confirm the hypothesis that practice with test increases JOL accuracy and suggest that retrieval practice improves the ability to discriminate items that can be recalled from items that cannot.

Further analyses to compare the delayed-JOL effect and the effect of practice with test showed that gamma was greater in SJT with delayed JOLs than in STSJT or in SJTSJT with immediate JOLs, $t(80) = 5.98$; $t(78) = 5.76$, respectively. The delayed-JOL effect was found even in STSJT, which had the smallest gamma difference between immediate and delayed JOLs, $t(40) = 5.42$. These results disconfirm the hypothesis from the cue-utilization framework that the effects of delay and practice are mediated by a similar process.

JOL accuracy in SJTSJT. The practice condition, SJTSJT itself, as used by Koriat and his colleagues (e.g., Koriat, 1997; Koriat et al. 2002; Koriat & Shitzer-

Reichert, 2002), treats practice as a within-subject variable (i.e., control [first SJT] vs. practice [second SJT]). Hence, further analyses of gammas on this condition are critical to refer to the practice effect and are reported below; all results of recall and JOL magnitude are reported in Appendix B.

Four gamma correlations were calculated from SJTSJT for each of immediate JOLs and delayed JOLs (i.e., a total of 8 gamma correlations): (1) gamma between the first JOL rating and the first recall (i.e., SJTSJT where underlines identify the JOL rating and recall used to calculate gamma), (2) gamma between the second JOL rating and the second recall (i.e., SJTSJT), (3) gamma between the second JOL rating and the first recall (i.e., SJTSJT), and (4) gamma between the first JOL rating and the second recall (i.e., SJTSJT). Figure 2 shows the mean gamma as a function of the variables being correlated and JOL timing. A two-way ANOVA showed that the main effects of the variables being correlated and JOL timing were significant, $F(3, 99) = 40.29$, $MSE = .07$, ES = .55; $F(1, 33) = 75.90$, $MSE = .06$, ES = .75, respectively, and that the interaction was significant, $F(3, 99) = 15.98$, $MSE = .08$, ES = .34. By Tukey's HSD test for immediate JOLs, the significance pattern can be represented as 'SJTSJT' ≈ 'SJTSJT' < 'SJTSJT' < 'SJTSJT' (the corresponding outcome for delayed JOLs is presented in Figure 2). These results replicated most of the findings that Koriat (1997) showed using immediate JOLs, and suggest that (immediate) JOLs are based on information pertaining to the outcome of the previous recall (i.e., the highest gamma is for SJTSJT) and are at least partially postdiction based on memory for remembered items (King et al., 1980; Koriat, 1997; Lovelace, 1984). Indeed, gamma for SJTSJT with immediate JOLs was so high that it was not possible to find

a delayed-JOL effect in this condition, $t(33) = 1.38$, $p = .18$.

Gamma was greater for SJTSJT with delayed JOLs than for SJTSJT with immediate JOLs, $t(33) = 3.51$, and the delayed-JOL effect occurred in SJTSJT, $t(33) = 3.36$. In fact, gamma was even greater for SJTSJT with delayed JOLs than for SJTSJT with immediate JOLs, $t(33) = 2.27$.

All results from analyses on SJTSJT provide further evidence that practice with test increases immediate JOL accuracy (i.e., 'SJTSJT' < 'SJTSJT'), and that practice with test did not improve accuracy as much as delay did, as explored above where this condition was manipulated between subjects.

Linking JOL distributions to JOL accuracy. In another attempt to understand the practice effect on JOL accuracy, I report how people used the rating scale when making JOLs.[4] Figures 3 and 4 show the distributions of JOL ratings for items correctly and incorrectly recalled, respectively, at the final test. For the control and practice without test, participants used extreme values far more frequently than middle values in the delayed-JOL condition whereas the reverse was true in the immediate-JOL condition; most delayed-JOL items receiving extremely high and low values were recalled correctly and incorrectly at the final test, respectively.

Participants also used somewhat more extreme values when items were repeated with test than without test (except for STSJT for incorrectly recalled items), and most items in the practice-with-test conditions receiving high and low values were recalled correctly and incorrectly at the final test, respectively. In particular, the

[4] To confirm whether distributions of JOL ratings differ from one another, the Kolmogorov-Smirnov test was applied. However, the inferential statistics from the test are not reported here because they are not closely relevant to the hypotheses under investigation. The complete results of the distribution data analyses across all experiments are available-upon-request from the author.

difference between the practice-with-test and the practice-without-test distributions (of immediate JOLs) occurs more clearly for correctly recalled items at the final test than for incorrectly recalled items. However, the distributions of delayed JOLs are clearly different from those of practice with test in the immediate-JOL condition; the delayed JOLs were more polarized than the immediate JOLs with test trials. An implication of the different distributions is that the greater accuracy of delayed JOLs results from an obvious bidirectional shift (i.e., extremely high and low JOL ratings), and that the moderately great accuracy of practice with test relative to study-alone practice arises from a unidirectional one (i.e., more high values of JOL ratings).

Summary and Discussion

Experiment 1 supports the hypothesis that intervening tests (manipulated as both a between- and a within-subject variable) are critical to improve one's ability to discriminate items that can be recalled from items that cannot, when JOLs are made immediately following item presentation. However, the two important factors affecting JOL accuracy, practice with test and delay, involve different processes, as indicated by greater accuracy and the very different rating distributions, more polarized after delay than following test practice.

Presumably, the increased accuracy from practice with test relies largely on the knowledge of the outcomes of previous tests, which is somewhat but not fully diagnostic of subsequent recall. Practice without test provides no opportunities for such diagnosticity. Of importance, the practice-with-test effect is achieved through overt but not covert responses, which suggests that self-feedback gained through intervening tests updates knowledge about the to-be-judged item, and that people are

aware of the knowledge that makes JOLs accurate. This idea is consistent with the finding that JOL accuracy increased in both feedback and no feedback conditions with repeated items having test trials (Koriat, 1997).

In regard to the evaluation of JOL accuracy, Experiment 1 did not make any assumption that a given item was or was not retrievable when it received a JOL. However, both monitoring-dual memories and self-fulfilling-prophecy hypotheses assume hypothetical aspects of retrieval occur at the time of JOLs. Experiment 2 more specifically investigates the contribution of intervening tests to the increase of JOL accuracy, not only offering the possibility of making the hypothetical aspects of retrieval observable, but also yielding more information about the memory mechanism giving rise to gamma accuracy.

Chapter 3: Experiment 2

Experiment 2 employed a revised methodology for metacognitive research, which is called "Pre-judgment Recall And Monitoring (PRAM)", provided by Nelson et al. (2004). In the PRAM methodology, a stage of recall, termed pre-judgment recall, is inserted right before JOL rating. The PRAM methodology allows items to be categorized in terms of whether or not they are recalled in the pre-judgment recall phase. Gamma, then, is computed separately for three partitions of item dyads: (1) dyads in which both of the items were recalled during pre-judgment recall (RR dyads), (2) dyads in which one was recalled and the other was not during pre-judgment recall (RN dyads), and (3) dyads in which neither of the items was recalled during pre-judgment recall (NN dyads). One can compute separate gamma statistics, γ_{RR}, γ_{RN}, and γ_{NN}, for the three item dyad partitions. The overall gamma ($\gamma_{..}$) is the frequency-weighted average of the three components. That is,

$$\gamma_{..} = \frac{(f_{RR} \times \gamma_{RR}) + (f_{RN} \times \gamma_{RN}) + (f_{NN} \times \gamma_{NN})}{f_{RR} + f_{RN} + f_{NN}}, \qquad (1)$$

where f_{ij} is the frequency of occurrence of dyads in partition ij (e.g., f_{RR} is the frequency of RR dyads). Simplifying, Equation 1 can be expressed as

$$\gamma_{..} = (p_{RR} \times \gamma_{RR}) + (p_{RN} \times \gamma_{RN}) + (p_{NN} \times \gamma_{NN}), \qquad (2)$$

where each p is the proportion of all dyads in the partition (or the weight for each component gamma: e.g., $p_{RR} = f_{RR}/(f_{RR} + f_{RN} + f_{NN})$, where $p_{RR} + p_{RN} + p_{NN} = 1$).

Calculating the component gammas and their weights, Nelson et al. (2004)

reported that people who predict their future recall discriminate between items more accurately when discriminating between a recalled and a non-recalled item than when discriminating between two recalled items. They also found that for delayed JOLs, most of the relevant discriminations are between RN dyads and relatively few are between RR dyads; whereas for immediate JOLs, most of the relevant discriminations are between RR dyads and relatively few are between RN dyads. These results ascribe most of the greater accuracy of delayed JOLs to different ratios of easier versus more difficult discriminations between items.

Experiment 1 discovered that practice with test, but neither practice with JOL rating nor study-alone practice, improved JOL accuracy, and that the advantage of practice with test was different from the advantage of delay. Experiment 2 used the PRAM methodology to investigate the effects of practice more analytically. Specifically, this experiment asks two questions; how practice with versus without test affects accuracy for immediate JOLs within RR, NN, and RN dyads; and whether the advantage of practice with test is most pronounced in RN dyads, mirroring the typical delayed-JOL effect (without practice) – that would be true if the influences of delay and practice derive from the same source.

Method

Participants. Two hundred and seventy-nine undergraduate students at the University of Maryland were recruited and received credit for psychology courses in return for their participation. Ninety-three participants were assigned to each of the 3 groups by block randomization.

Materials and procedure. The materials and procedure were identical to those

of Experiment 1 except that a stage of pre-judgment recall was inserted right before the JOL rating through all conditions. The pre-judgment recall consisted of self-paced paired associate recall as in the final recall test.

Design. The design was a between-subject design comprised of three conditions: SJT with delayed JOLs, SSJT with immediate JOLs, and STSJT with immediate JOLs (i.e., the control with delayed JOLs vs. two practice conditions with immediate JOLs).

Results

The descriptive statistics and the results from separate one-way ANOVAs on percent correct recall and on JOL magnitude are reported in Appendix C. By Tukey's HSD test, the significance pattern of recall can be represented as 'SJT with delayed JOLs' < 'SSJT with immediate JOLs' < 'STSJT with immediate JOLs'. The prerequisite recall results allow for analyses of gamma, as described next.

JOL accuracy. The mean overall gammas were .95 (*SE* = .02), .55 (*SE* = .03), and .68 (*SE* = .02) for SJT with delayed JOLs, SSJT with immediate JOLs, and STSJT with immediate JOLs, respectively. A one-way ANOVA showed that the main effect of type of practice was significant, $F(2, 266) = 64.43$, $MSE = .06$, ES = .33. By Tukey's HSD test, the significance pattern can be represented as 'SSJT with immediate JOLs' < 'STSJT with immediate JOLs' < 'SJT with delayed JOLs'. These results are consistent with those of Experiment 1; in the immediate-JOL conditions, practice with test yielded better JOL accuracy than did study-alone practice but did not improve accuracy as much as delay did.

Figures 5 and 6 show the mean component gammas of RR and RN dyads (γ_{RR}

and γ_{RN}), and the mean weights of RR and RN dyads (p_{RR} and p_{RN}), respectively, from the PRAM methodology. I do not include the γ_{NN} and p_{NN} results because that NN dyads occurred so infrequently (i.e., only 7, 2, and 2 of each of the 93 participants had estimates of γ_{NN} for SJT with delayed JOLs, SSJT with immediate JOLs, and STSJT with immediate JOLs, respectively), and any difference played a negligible role in overall JOL accuracy (but each mean of γ_{NN} and p_{NN} is reported below where the overall gamma is calculated with the three component gammas and their weights). I conducted one-way ANOVAs for each component gamma and for each weight. First, γ_{RR} was significantly different among the three conditions, $F(2, 229) = 4.57$, $MSE = .18$, ES = .04. By Tukey's HSD test, the significance pattern of γ_{RR} can be represented as 'SJT with delayed JOLs' ≈ 'SSJT with immediate JOLs' < 'STSJT with immediate JOLs'. While Nelson et al. (2004) found greater γ_{RR} for delayed JOLs than for immediate JOLs, the pattern of results was opposite if practice with test but not without test was added. Second, γ_{RN} was significantly different among the three conditions, $F(2, 149) = 11.61$, $MSE = .09$, ES = .14. By Tukey's HSD test, the significance pattern of γ_{RN} can be represented as 'SSJT with immediate JOLs' < 'STSJT with immediate JOLs' ≈ 'SJT with delayed JOLs'. The finding that γ_{RN} was greater for delayed JOLs than for immediate JOLs (Nelson et al., 2004) was replicated when items were repeated without test but not with test.

Both p_{RR} and p_{RN} were significantly different among the three conditions, $F(2, 266) = 4185.66$, $MSE = .01$, ES = .97; $F(2, 266) = 4018.95$, $MSE = .01$, ES = .97, respectively. By Tukey's HSD test, not surprisingly, the significance pattern can be

represented as ‘SJT with delayed JOLs’ < ‘STSJT with immediate JOLs’ ≈ ‘SSJT with immediate JOLs’, for p_{RR}; and as ‘SSJT with immediate JOLs’ ≈ ‘STSJT with immediate JOLs’ < ‘SJT with delayed JOLs’ for p_{RN}. These results are consistent with those of Nelson et al. (2004) (in which practice was not manipulated) that p_{RR} was greater for immediate JOLs than for delayed JOLs whereas p_{RN} was greater for delayed JOLs than for immediate JOLs. The differences of the weights between STSJT with immediate JOLs and SJT with delayed JOLs are difficult to reconcile with the speculation that the effects of practice and delay are mediated by a similar process.

Linking JOL distributions to JOL accuracy. Figures 7 and 8 show the distributions of JOL ratings for items that were and were not recalled, respectively, at both the pre-judgment and final recall tests. As in Experiment 1, most delayed-JOL items receiving extremely high and low values were recalled correctly and incorrectly at the final test, respectively, and most items in STSJT receiving high values, relative to those in SSJT, were recalled correctly at the final test. However, the distributions of SJT with delayed JOLs are different from those of STSJT with immediate JOLs for both correct and incorrect responses.

Most importantly, the data of pre-judgment recall were almost exactly identical to those of final recall in SJT with delayed JOLs for both correct and incorrect responses. By contrast, for items in the two immediate-JOL conditions, the distributions between pre-judgment and final recall tests were different from each other for both correct and incorrect responses. An interpretation of the rating distribution data is that the information retrieved about the to-be-judged item at the

time of the delayed JOL is strongly predictive of eventual memory performance, and then delayed JOLs are highly accurate.

Summary and Discussion

Experiment 2 replicated the findings of the effect of practice with test in Experiment 1. By Equation 2, the overall gamma for each condition is calculated with negligible differences of decimal points as follows:

$$.95 \approx (.03 \times .48) + (.96 \times .97) + (.001 \times (-.40)) = .94,$$

$$.55 \approx (.97 \times .54) + (.02 \times .66) + (.0001 \times 1.00) = .54,$$

$$.68 = (.95 \times .68) + (.04 \times .86) + (.0002 \times 1.00) = .68,$$

where the far left-hand side of each dual equation is the overall gamma for the conditions, SJT with delayed JOLs, SSJT with immediate JOLs, and STSJT with immediate JOLs, respectively, and the far right-hand side is the calculated gamma with the three component gammas and their weights. The practice effect with versus without test mainly results from two component gammas (i.e., higher γ_{RR} and γ_{RN} for practice with test); intervening tests generally boost up ability to discriminate items that can be recalled from items that cannot at the final test. Most obviously, the overall accuracy of delayed JOLs is greater than that of practice-with-test immediate JOLs whereas a particular component gamma, γ_{RR}, shows the opposite pattern. It is no less dubious to connect that the advantage of practice with test on overall JOL accuracy does not mirror the typical delayed-JOL effect because the practice-with-test effect depends largely on RR dyads (i.e., extremely high p_{RR} and greater γ_{RR}) whereas the delayed-JOL effect does largely on RN dyads (i.e., extremely high p_{RN}).

As another benefit from the PRAM methodology, the comparisons of JOL

rating distributions at the pre-judgment and final recall tests show markedly the difference between the delayed-JOL effect (i.e., identical distributions between the two tests) and the effect of practice with test (i.e., more frequent lower values for items correctly and incorrectly recalled at the pre-judgment recall test). When they learn items repeatedly with intervening tests, in all possibility, individuals rely so much on the outcomes of the previous test that they may somewhat but not fully predict the memory performance of future test.

Experiment 3 further investigates whether the main function of test is to raise JOL accuracy and rules out an alternative explanation of the beneficial effect of practice with test, as described next.

Chapter 4: Experiment 3

Experiments 1 and 2 revealed the advantage of practice with test (STSJT) over study-alone practice (SSJT) on immediate-JOL accuracy. However, because one more phase (T) was inserted in STSJT than in SSJT, it is plausible that the increased processing for the additional phase caused the increase of JOL accuracy (in fact, recall was greater in STSJT than in SSJT). Although the explanation is inconsistent with the fact that there was no advantage of study-alone practice (SSJT) over no repetition (SJT) in Experiment 1, the explanation deserves attention. Manipulating the type of practice, SSSJT and STSJT, in which both conditions included 5 processing steps, Experiment 3 tested whether the beneficial effect of STSJT was due to the function of the test or just to the increased processing.

Experiment 3, just as did Experiments 1 and 2, also compared the two effects of delay and practice.

Method

Participants. One hundred and six undergraduate students at the University of Maryland were recruited and received credit for psychology courses in return for their participation. Fifty-three participants were assigned to each of the 2 groups by block randomization.

Materials and procedure. The materials were identical to those of Experiment 1. STSJT was the same condition as in Experiment 1, and SSSJT had three consecutive study phases and JOL rating (during the third one) followed by test phase. Each procedure for study, JOL rating, and test was identical to that of

Experiment 1.

Design. The design of Experiment 3 was a 2 × 2 mixed factorial with type of practice (SSSJT vs. STSJT) manipulated between subjects and JOL timing (immediate vs. delayed) manipulated within subjects.

Results

All results for both recall and JOL magnitude are reported in Appendix D. As seen in Table D2, Experiment 3 failed to yield a retrieval practice effect on memory performance, presumably due to relatively short test-spacing and/or due to the type of task, cued-recall; tests are typically most effective if items are first tested some time soon, but not immediately after the presentation and if free-recall is required instead of cued-recall or recognition (see Dempster, 1996 for a review; Glover, 1989).

JOL accuracy. Figure 9 shows the mean gamma as a function of JOL timing and type of practice. A two-way ANOVA showed that gamma was greater for delayed JOLs than for immediate JOLs, $F(1, 92) = 67.71$, $MSE = .09$, ES = .42; that there was no main effect of type of practice, $F(1, 92) = 3.93$, $MSE = .10$, $p = .0504$; and that the interaction was significant, $F(1, 92) = 4.91$, $MSE = .09$, ES = .05. Follow up simple-effect tests showed that gamma of immediate JOLs was reliably greater for STSJT than for SSSJT, $t(92) = 2.37$, and that gamma of STSJT was reliably greater for delayed JOLs than for immediate JOLs, $t(45) = 4.17$. These results confirm the hypothesis that the increase of JOL accuracy is due to the intervening test itself, and disconfirm the hypothesis from the cue-utilization framework, as in Experiments 1 and 2, that the effects of delay and practice are mediated by a similar process.

Linking JOL distributions to JOL accuracy. Figures 10 and 11 show the

distributions of JOL ratings for items that were correctly and incorrectly recalled, respectively, at the final test. For both SSSJT and STSJT, participants used extreme JOL ratings more frequently in the delayed-JOL conditions than in the immediate-JOL conditions.

Although for both immediate and delayed JOLs in STSJT, participants used more high values of JOL ratings for items they correctly recalled at the final test and more low ones for items they failed to correctly recall, the immediate JOLs in STSJT were not as extreme as were the delayed JOLs. An implication of these rating distribution data is that at the time of JOLs, strongly successful differentiation between subsequent correct and incorrect recall yields better accuracy, and that it is more successful for delayed JOLs than for immediate JOLs with test trials.

Summary and Discussion

The results of Experiment 3 provide further support for the hypotheses that intervening tests, but not just increased processes, are crucial to enhance JOL accuracy, and that the contribution of intervening tests to the increase of JOL accuracy differs from that of delay. Experiment 3 shows that greater recall through test experience is not needed for accurate JOLs (a conclusion also supported by the findings that there was no recall difference between immediate and delayed JOLs across all three experiments). An additional study substituted for test does not serve as a critical factor to improve JOL accuracy; the gamma results of SSSJT in Experiment 3 are similar to those of SSJT in Experiment 1.

Chapter 5: General Discussion

The results of this research support the hypothesis that retrieving items only through test or overt retrieval improves the accuracy of individuals' predictions regarding their memory performance. Neither additional study trials nor additional JOL ratings affected the fundamental attribute of judging which items were more versus less well-learned. However, the advantage of practice with test was different from the advantage caused by delay. Not only did delay improve accuracy more than practice with test did, but the improvement was due to effects in different dyads of items (i.e., RN and RR dyads for delay and practice with test, respectively). Moreover, the distribution of JOL ratings was more polarized for delayed JOLs than for practice with test, which reflects better accuracy. Thus, individuals' item-by-item JOLs can be accurately predictive of the effects of intervening tests, but such accuracy also depends on the timing of JOLs.

Different Influences on Delay and Practice with vs. without Test

The JOL theories mentioned earlier, monitoring-dual memories hypothesis, self-fulfilling-prophecy hypothesis, and cue-utilization framework, cannot fully account for the beneficial effect of practice with test because they are not concerned with the distinction between the effects of study and test experience. In the memory literature, however, such a distinction has been discussed (e.g., Bjork & Bjork, 1992; see also Dempster, 1996 for a review). Bjork and Bjork (1992) distinguished between two hypothetical factors, storage strength (or study practice) and retrieval strength (or retrieval/test practice), which may increase the likelihood of future recall. The act of

re-studying an item contributes to storage strength, or the extent to which the item is well learned whereas test experience helps build retrieval strength, or the ease with which the item can be accessed. Although Bjork and Bjork (1992) did not investigate this distinction directly, it can be applied to the present experiments, with an assumption from the JOL theories that the valid information retrieved from memory about the to-be-judged item yields the increase of JOL accuracy. In similar vein, Koriat and Ma'ayan (2005) distinguished encoding fluency, which refers to the ease with which items are learned well during study, and retrieval fluency, which refers to the ease with which they come to mind (see also, Benjamin, Bjork, & Schwartz, 1998). Inferring encoding fluency from self-paced study time and retrieval fluency from the success and latency of pre-judgment recall, they found that (1) JOLs decreased with increasing self-paced study time whereas JOLs increased with experimentally manipulated study time, and (2) JOLs increased with retrieval fluency. Although Koriat and Ma'ayan's (2005) distinction appears not to clarify the different influences of study-alone practice and practice with test on JOL accuracy because study-alone practice is assumed to be relevant for both encoding and retrieval fluency, such a distinction certainly merits consideration.

How do practice with test and delay improve JOL accuracy? Do they affect accuracy in the same way? The answers to these questions are important to understand the underlying process of JOLs. While Koriat and Shitzer-Reichert (2002) found evidence confirming the hypothesis of the same source (i.e., the increased fluency of the experience-based mnemonic cues underlying JOLs) between delay and practice (with test), the present experiments did not. Different participants (i.e.,

children vs. college students) might give rise to these different results. As may be ascertained from the findings of the PRAM methodology and distributions of JOL ratings, however, the effects of delay and practice with test differ from each other. Indeed, all findings of this research suggest that the influence of practice with test is intermediate between the effects of immediate JOLs (without test trials) and delayed JOLs. The results of the present study are also in accord with a formulation that ascribes the difference of accuracy between delay and practice with test to different ratios of easier versus more difficult discriminations between items (Nelson et al., 2004). To speak in the context of the cue-utilization framework, the effects of delay and practice with test may differ from each other in terms of the degree that the basis for JOLs changes from a theory-based inference towards a greater reliance on experience-based mnemonic cues (i.e., a much greater reliance on experience-based mnemonic cues after delay than with test practice). From the findings of the PRAM methodology, the type of retrieval attempts itself may not be a critical factor unless the time duration is controlled because overt retrieval attempts were made in all conditions. Another important factor from the results of SJSJT, because covert retrieval did not increase JOL accuracy immediately after study trial, is whether the study trial and the covert retrieval attempt are spaced far enough apart. At the very least, the question of how different the effects of delay and practice with test are (i.e., quantitatively or qualitatively) remains open.

The Nature of the Benefit of Practice with Test

Although the results of the current study indicate a benefit of practice with test, a question remains as to the nature of the benefit. The retrieval hypothesis (e.g.,

Bjork, 1988) suggests that it is the processing engendered by acts of retrieval that accounts for the effects of intervening tests, not merely the amount of processing. This retrieval hypothesis is appealing because tests generally afford fewer retrieval cues than additional study trials (e.g., only the stimulus of a pair in cued-recall rather than both the stimulus and target of a pair in paired-associate learning). It is also relevant to the reason that there is an extremely robust delayed-JOL effect when the cue for JOLs is the stimulus alone whereas there is little delayed-JOL effect when the cue for JOLs is the stimulus-target pair (Dunlosky & Nelson, 1992).

The specifics of the retrieval hypothesis exist in a variety of forms. At a global level, first, retrieval attempts during intervening tests may help build a context similar to that of the final test at the time of JOLs. However, the notion of global context similarity does not seem to have much merit because similarity between the cue for JOLs and the cue for test is not the primary determinant of the accuracy (Dunlosky & Nelson, 1997). Second, intervening tests provide opportunities for general practice that boost the likelihood of correct retrieval at the final recall (e.g., Runquist, 1983) and then increase overall accuracy. This account is difficult to reconcile with the facts that the beneficial effect of prior test on memory performance applies mainly to items that were successfully retrieved on that test (Carrier & Pashler, 1992; Glover, 1989), and that the proportion of items in RR dyads was large when items were repeated with test.

At the level of individual items, retrieval attempts may either strengthen existing retrieval routes to the representation of the item in memory (Birnbaum & Eichner, 1971; Bjork, 1975) or result in the creation of new routes (Bjork, 1975) or

even a context change for retrieved items. It is assumed that these strengthened or new routes/contexts will raise both the accessibility to valid information of the item to be judged at the time of JOLs and the probability of correct recall at the final recall. Although such hypotheses seem reasonable, there is little evidence that multiple retrieval routes/contexts are sufficient to increase memory performance and JOL accuracy, and the results of this research cannot entirely support any of the hypotheses. These possible explanations are open to question.

Concluding Comments

This study has clear implications for applied learning situations. When individuals have the option of monitoring their memories either immediately after study or after a brief delay and there is only one learning opportunity (e.g., learning new concepts in the classroom without any preview until the time of review), they should wait for a short delay before making their JOLs. Such delayed JOLs will yield a more informed choice of study activities that will be more effective for learning those items. When reviewing items (e.g., re-studying the concepts for exams), learners should take self-tests because they yield a better long-term retention than study-alone practice trials, as previous studies in memory research have found. For that reason, including test trials has been often referred to as an optimal method of learning (Carrier & Pashler, 1992). The results of the current study support this idea suggesting that intervening tests are also important in monitoring one's knowledge, which affects metacognitive control. At the time of review with test trials, monitoring memories after a delay is also critical because the combination of intervening tests and delayed JOLs allows individuals to develop their best strategies for improving

acquisition based on valid information regarding retrievability and for enhancing the most long-term retention of the materials.

The potential of practice with test for improving learning and retention is vast. As an example, the present study disclosed different effects on JOL accuracy of practice with versus without test. The results suggest that current JOL theories need to be modified to reflect that retrieval practice through test provides the relative validity individuals rely on to predict their memory performance. It is important to achieve valid information about memory for effective learning. The results also suggest that there needs to be differentiation between the validity monitored from studied items after some delay without test experience and the validity acquired from the previous outcomes based on test experience. One way to understand the underlying process of JOLs may be to clarify directly and indirectly updated access to one's own knowledge.

Figure 1. Mean gamma as a function of JOL timing and type of practice in Experiment 1.
Each vertical hash mark depicts the standard error of the mean. Means of the type of practice that do not share alphabetic subscripts differ in Tukey's honestly significant difference (HSD) comparisons for immediate JOLs. S = Study; J = JOL rating; T = Test; JOLs = Judgments of learning.

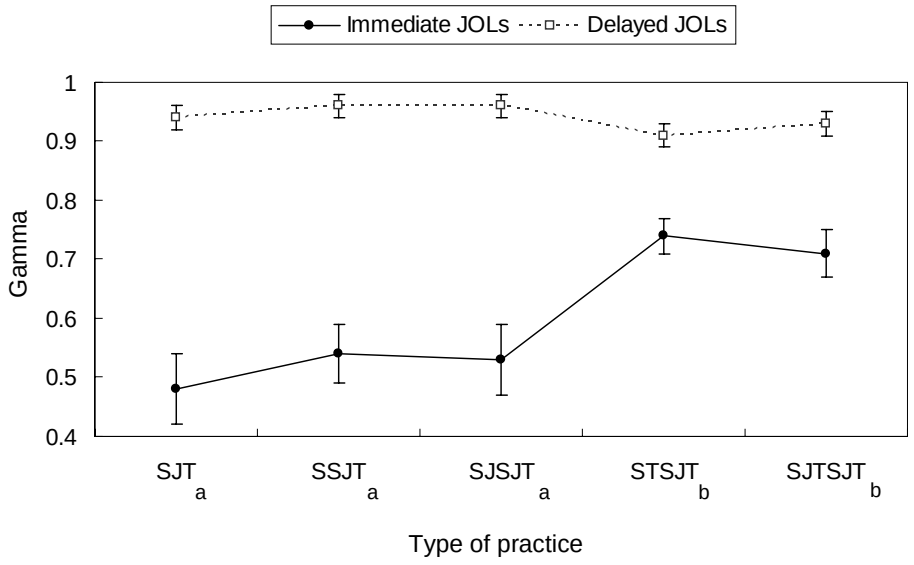


Figure 2. Mean gamma as a function of the variables being correlated and JOL timing in SJTSJT of Experiment 1.
Underlines of SJTSJT identify the measures of JOL magnitude and recall used to calculate gamma. Each vertical hash mark depicts the standard error of the mean. Means that do not share alphabetic and numeric subscripts differ in Tukey's HSD comparisons for immediate and delayed JOLs, respectively. S = Study; J = JOL rating; T = Test; JOLs = Judgments of learning.

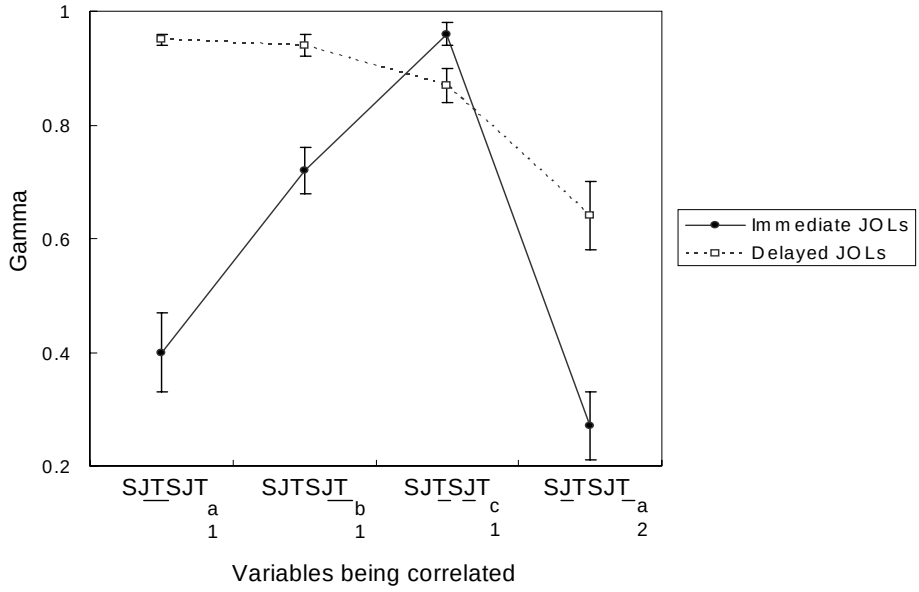


Figure 3. Distributions of JOL ratings for items correctly recalled at the final test in Experiment 1.
S = Study; J = JOL rating; T = Test; JOL = Judgment of learning.

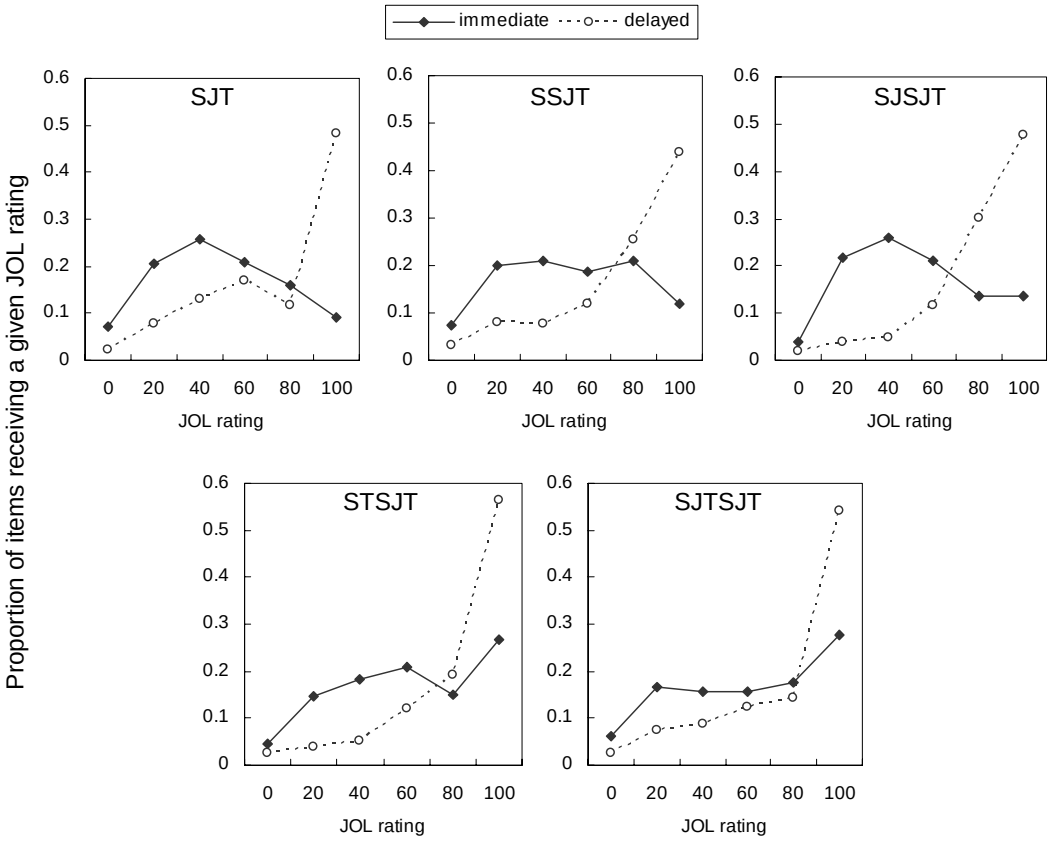


Figure 4. Distributions of JOL ratings for items incorrectly recalled at the final test in Experiment 1.
S = Study; J = JOL rating; T = Test; JOL = Judgment of learning.

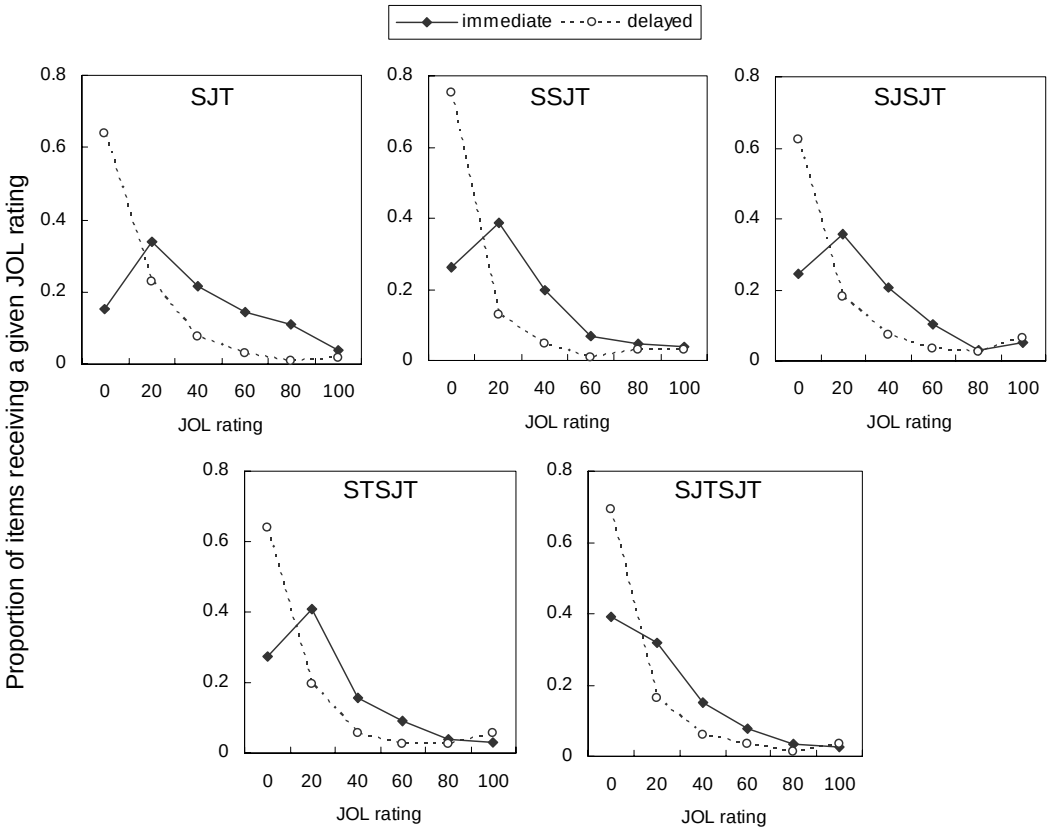


Figure 5. Mean γ_{RR} and γ_{RN} as a function of type of practice in Experiment 2. Each vertical hash mark depicts the standard error of the mean. Means of the type of practice that do not share alphabetic and numeric subscripts differ in Tukey's HSD comparisons for γ_{RR} and γ_{RN}, respectively. S = Study; J = JOL rating; T = Test; JOLs = Judgments of learning.

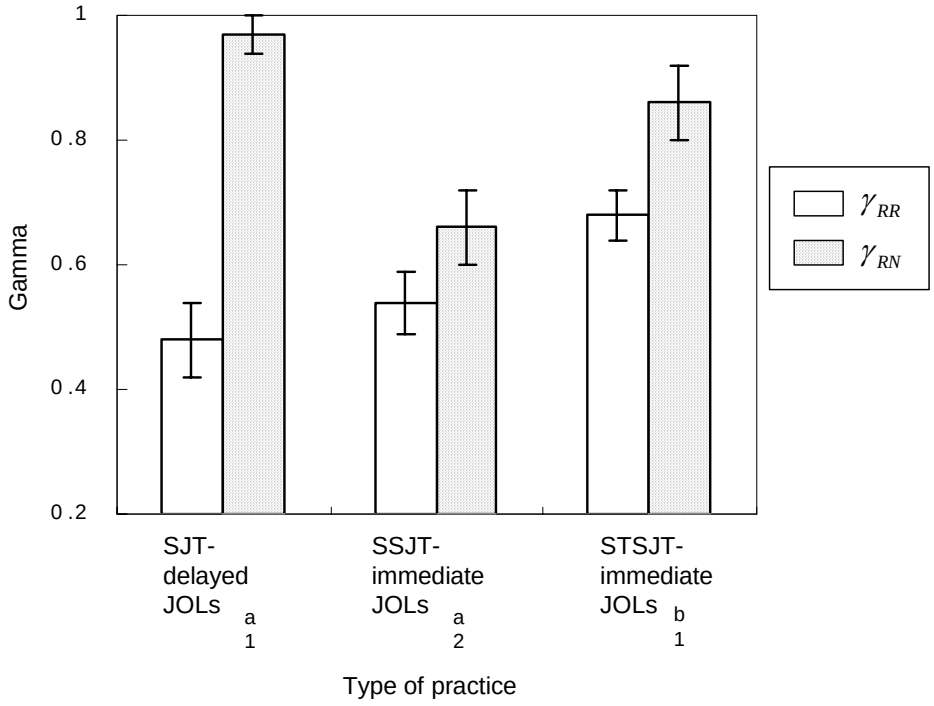


Figure 6. Mean p_{RR} and p_{RN} as a function of type of practice in Experiment 2. Each vertical hash mark depicts the standard error of the mean. S = Study; J = JOL rating; T = Test; JOLs = Judgments of learning.

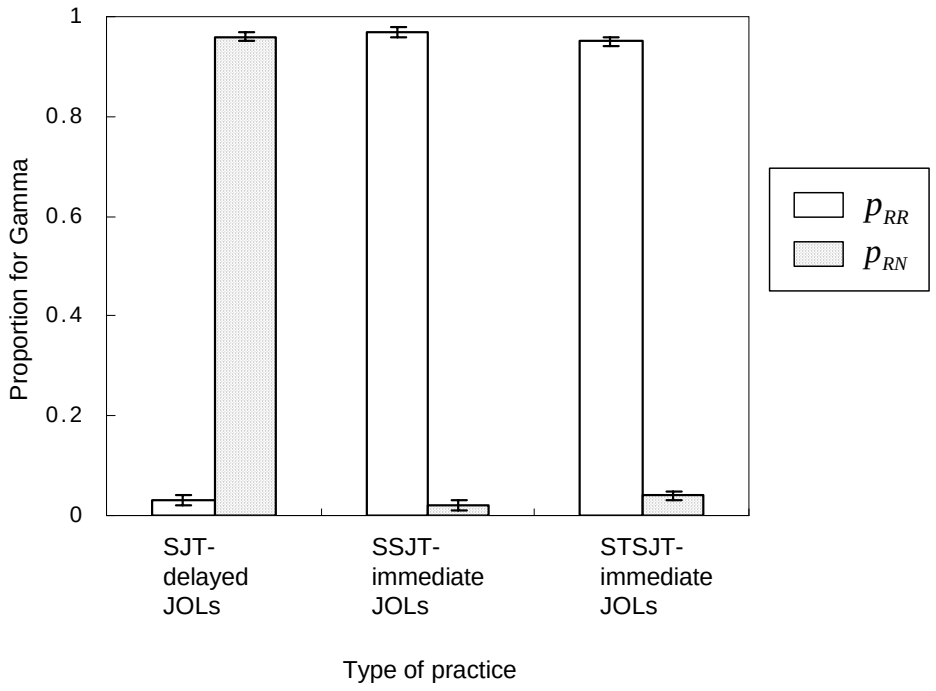


Figure 7. Distributions of JOL ratings for items correctly recalled at both pre-judgment and final recall tests in Experiment 2.
S = Study; J = JOL rating; T = Test; JOLs = Judgments of learning.

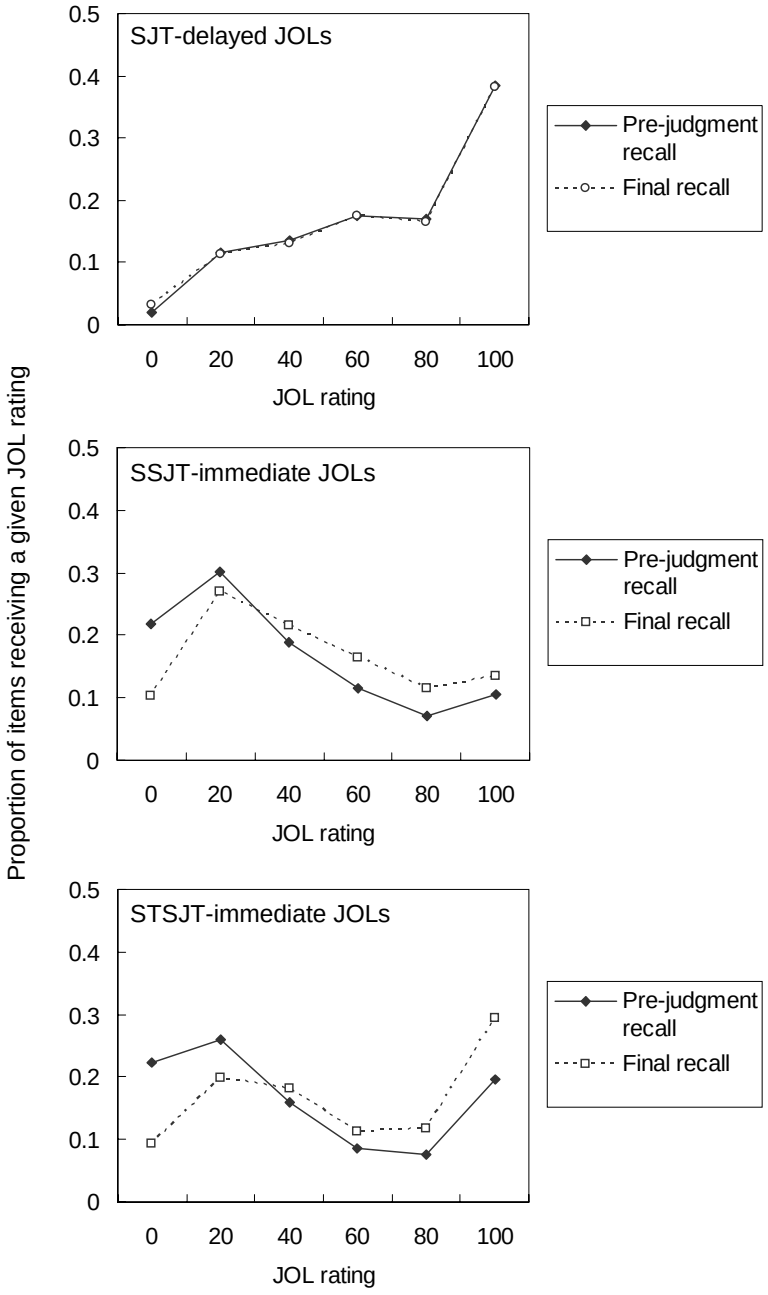


Figure 8. Distributions of JOL ratings for items incorrectly recalled at both pre-judgment and final recall tests in Experiment 2.
S = Study; J = JOL rating; T = Test; JOLs = Judgments of learning.

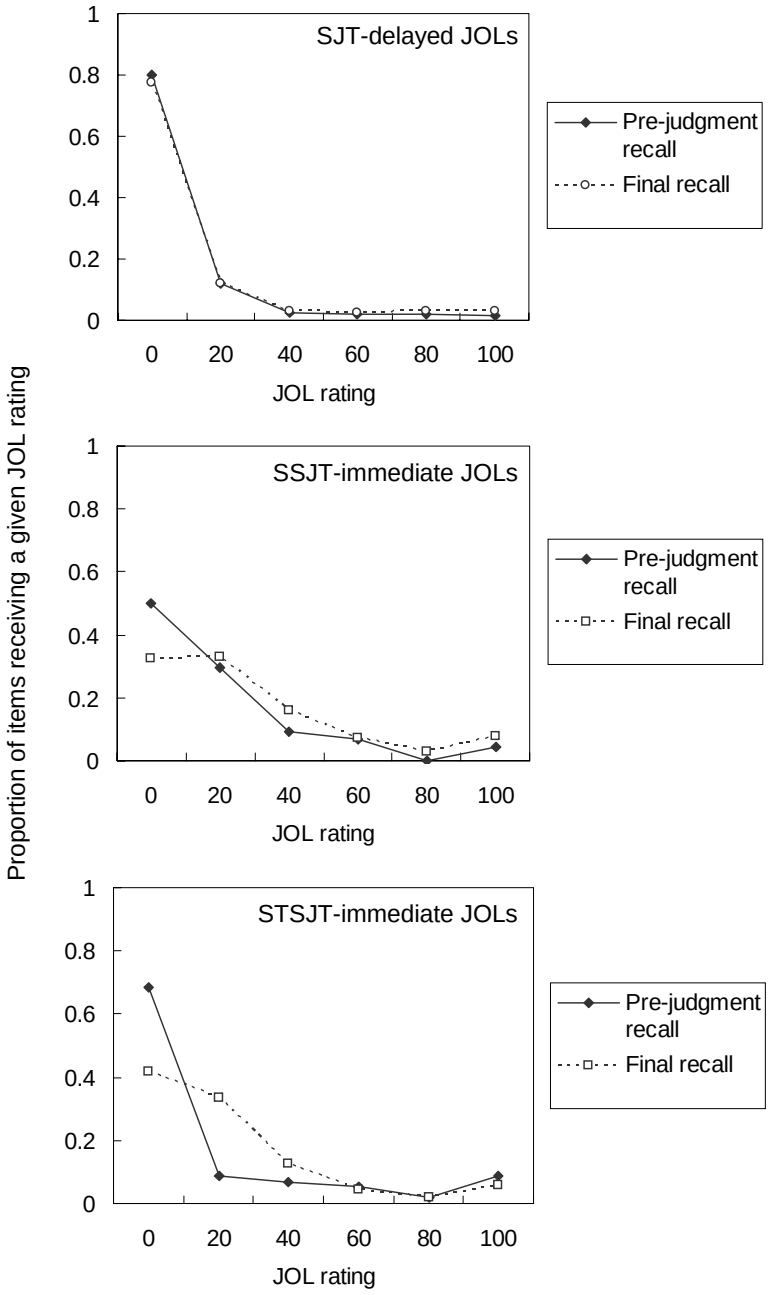


Figure 9. Mean gamma as a function of JOL timing and type of practice in Experiment 3.
Each vertical hash mark depicts the standard error of the mean. S = Study; J = JOL rating; T = Test; JOLs = Judgments of learning.

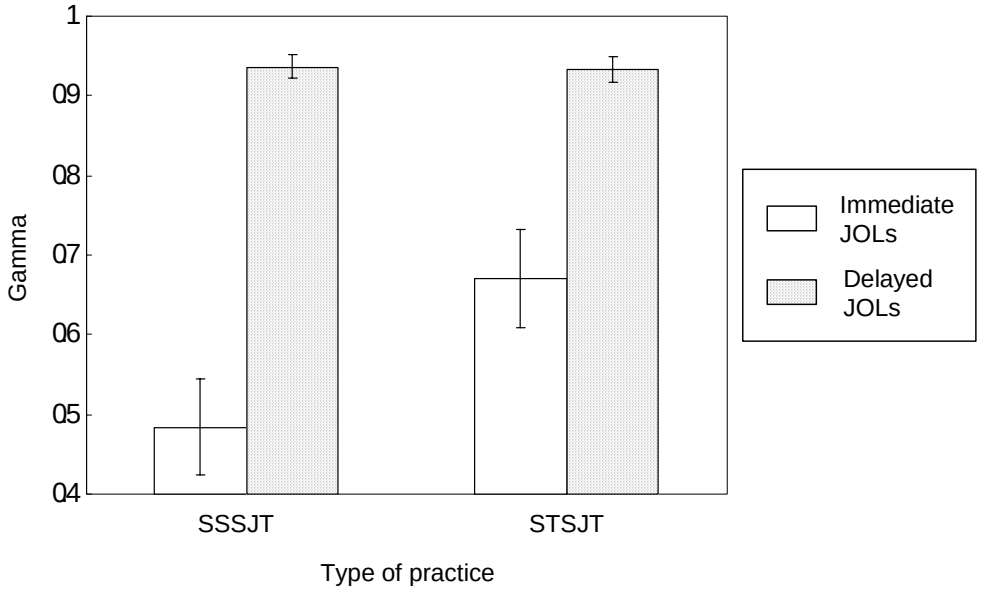


Figure 10. Distributions of JOL ratings for items correctly recalled at the final test in Experiment 3.
S = Study; J = JOL rating; T = Test; JOL = Judgment of learning.

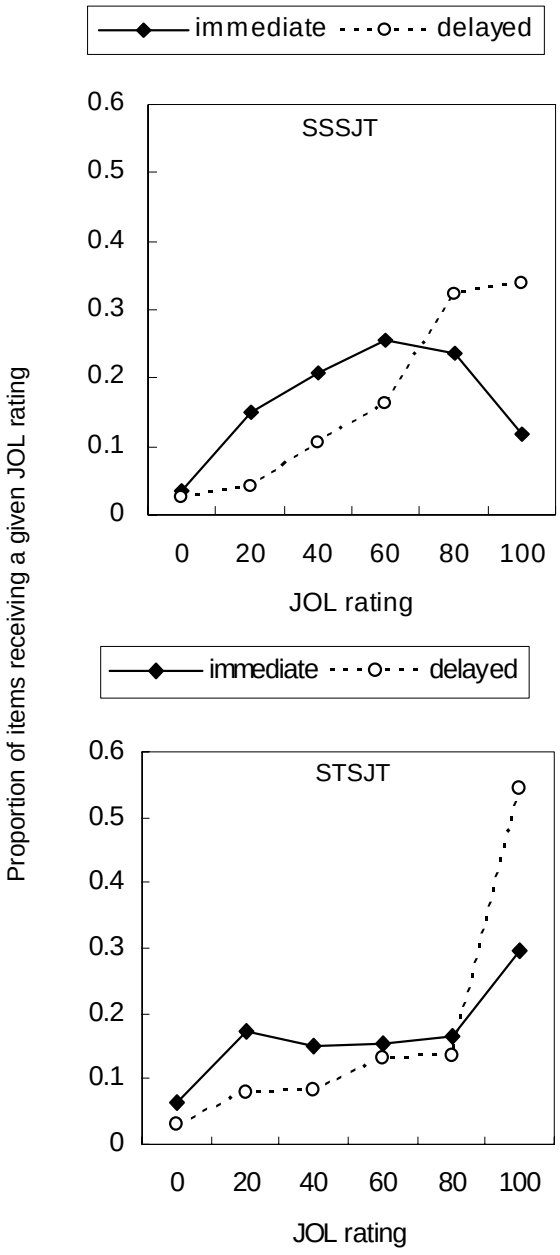
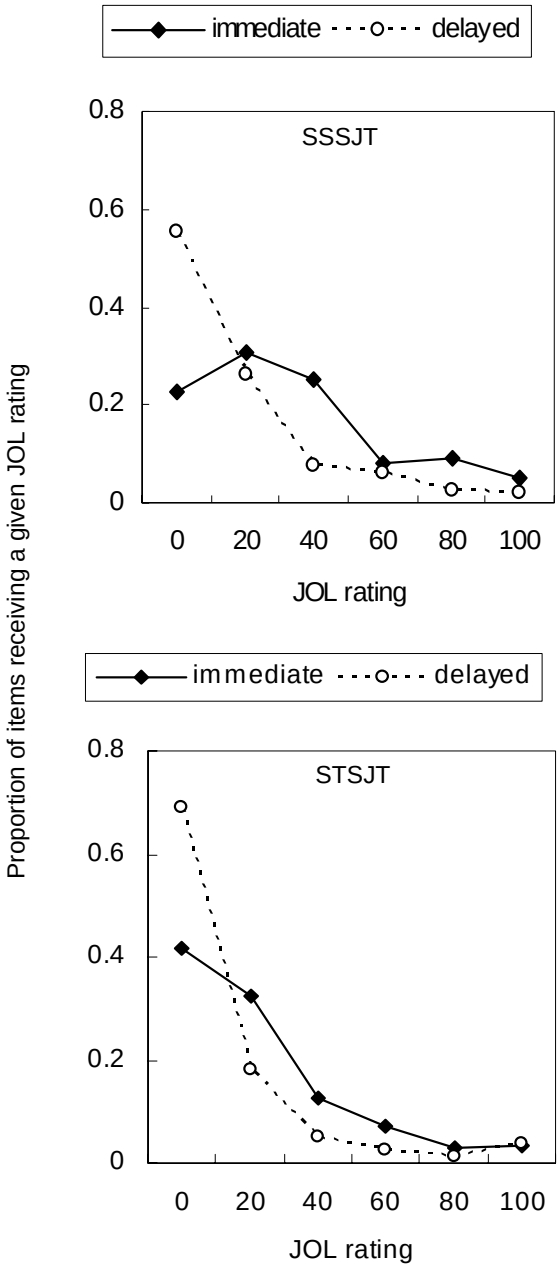


Figure 11. Distributions of JOL ratings for items incorrectly recalled at the final test in Experiment 3.
S = Study; J = JOL rating; T = Test; JOL = Judgment of learning.



Appendix A: Results of Recall and JOL Magnitude in All Conditions of Experiment 1

Mean percent correct recall and mean JOL magnitude in Experiment 1 as a function of JOL timing and type of practice are reported in Table A1, and the results from analyses of variance are reported in Table A2.

Table A1
Mean Percent Correct Recall and Mean JOL Magnitude in Experiment 1 as a Function of JOL Timing and Type of Practice

		Type of practice				
DV	JOL timing	SJT	SSJT	SJSJT	STSJT	SJTSJT
Recall	Immediate	32 (3.27)[a]	50 (4.22)[b]	52 (3.54)[b]	60 (3.93)[b]	62 (3.92)[b]
	Delayed	37 (3.71)[1]	51 (4.18)[1]	49 (3.70)[1,3]	64 (3.63)[2]	62 (3.84)[2,3]
JOLs	Immediate	40 (3.03)	40 (3.13)	41 (2.79)	47 (3.01)	46 (3.90)
	Delayed	35 (3.48)[1]	44 (3.92)[1,2]	48 (3.25)[1,3]	58 (3.64)[3]	53 (4.31)[2,3]

Note. Standard error of the mean is in parentheses; Means of the type of practice that do not share alphabetic and numerical superscripts differ in Tukey's honestly significant difference (HSD) comparisons in the immediate- and delayed-JOL conditions, respectively, for each of recall and JOL magnitude; S = Study; J = JOL rating; T = Test; JOLs = Judgments of learning; DV = Dependent variables.

Table A2

Results from Analyses of Variance of Percent Correct Recall and JOL Magnitude in Experiment 1

	Percent correct recall					JOL magnitude				
IV	*F*	*df*	*MSE*	*p*	ES	*F*	*df*	*MSE*	*p*	ES
Jt	2.14	1,220	92.11	.14		20.49	1,220	126.33	< .001	.08
P	9.27	4,220	1211.01	< .001	.14	3.28	4,220	961.16	< .05	.06
Jt×P	2.71	4,220	92.11	< .05	.05	6.59	4,220	126.33	< .001	.11

Note. Effect size (ES) is reported only when the *F* value was significant; JOL = Judgment of learning; IV = Independent variables; Jt = JOL timing; P = Type of practice.

Appendix B: Results of Recall and JOL Magnitude in SJTSJT of Experiment 1

Mean percent correct recall and mean JOL magnitude for condition SJTSJT in Experiment 1 as a function of JOL timing and practice are reported in Table B1, and the results from analyses of variance are reported in Table B2.

Table B1

Mean Percent Correct Recall and Mean JOL Magnitude for condition SJTSJT in Experiment 1 as a Function of JOL Timing and Practice

		Practice	
DV	JOL timing	Control (first SJT)	Practice (second SJT)
Recall	Immediate	27.24 (3.21)	61.53 (3.92)
	Delayed	30.27 (3.37)	62.47 (3.84)
JOLs	Immediate	40.70 (3.05)	46.20 (3.90)
	Delayed	28.24 (3.51)	53.41 (4.31)

Note. Standard error of the mean is in parentheses; S = Study; J = JOL rating; T = Test; JOLs = Judgments of learning; DV = Dependent variables.

Table B2

Results from Analyses of Variance of Percent Correct Recall and JOL Magnitude for Condition SJTSJT in Experiment 1

	Percent correct recall				JOL magnitude			
IV	*F*(1, 44)	*MSE*	*p*	ES	*F*(1, 44)	*MSE*	*p*	ES
Jt	1.11	158.49	.30		2.24	138.83	.14	
P	275.16	180.75	< .001	.86	38.96	271.54	< .001	.47
Jt×P	< 1				72.90	59.69	< .001	.62

Note. Effect size (ES) is reported only when the *F* value was significant; JOL = Judgment of learning; IV = Independent variables; Jt = JOL timing; P = Practice.

Appendix C: Results of Recall and JOL Magnitude in Experiment 2

Mean percent correct recall and mean JOL magnitude in Experiment 2 as a function of type of practice are reported in Table C1, and the results from analyses of variance are reported in Table C2.

Table C1

Mean Percent Correct Recall and Mean JOL Magnitude in Experiment 2 as a Function of Type of Practice

	Type of practice		
DV	SJT – delayed JOLs	SSJT – immediate JOLs	STSJT – immediate JOLs
Recall	37.57 (2.28) [a]	46.73 (2.40) [b]	57.95 (2.60) [c]
JOLs	32.34 (1.96) [1]	36.56 (2.36) [1, 2]	42.10 (2.35) [2]

Note. Standard error of the mean is in parentheses; Means of the type of practice that do not share alphabetic and numerical superscripts differ in Tukey's HSD comparisons for recall and magnitude of JOLs, respectively; S = Study; J = JOL rating; T = Test; JOLs = Judgments of learning; DV = Dependent variables.

Table C2

Results from Analyses of Variance of Percent Correct Recall and JOL Magnitude in Experiment 2

	Percent correct recall				JOL magnitude			
	F(2, 276)	*MSE*	*p*	ES	*F*(2, 276)	*MSE*	*p*	ES
P	17.59	551.27	< .001	.11	4.81	463.32	< .01	.03

Note. Effect size (ES) is reported only when the *F* value was significant; JOL = Judgment of learning; P = Type of practice.

Appendix D: Results of Recall and JOL Magnitude in Experiment 3

Mean percent correct recall and mean JOL magnitude in Experiment 3 as a function of JOL timing and type of practice are reported in Table D1, and the results from analyses of variance are reported in Table D2.

Table D1

Mean Percent Correct Recall and Mean JOL Magnitude in Experiment 3 as a Function of JOL timing and Type of Practice

		Type of practice	
DV	JOL timing	SSSJT	STSJT
Recall	Immediate	68.00 (3.70)	59.40 (3.70)
	Delayed	69.81 (3.66)	59.66 (3.66)
JOLs	Immediate	49.32 (3.23)	45.32 (3.23)
	Delayed	56.87 (3.72)	51.24 (3.72)

Note. Standard error of the mean is in parentheses; S = Study; J = JOL rating; T = Test; JOLs = Judgments of learning; DV = Dependent variables.

Table D2

Results from Analyses of Variance of Percent Correct Recall and JOL Magnitude in Experiment 3

	Percent correct recall			JOL magnitude			
IV	*F*(1, 104)	*MSE*	*p*	*F*(1, 104)	*MSE*	*p*	ES
Jt	< 1			25.06	95.98	< .001	.19
P	3.40	1370.17	.07	< 1			
Jt×P	< 1			< 1			

Note. Effect size (ES) is reported only when the *F* value was significant; JOL = Judgment of learning; IV = Independent variables; Jt = JOL timing; P = Type of practice.

References

Allen, G. A., Mahler, W. A., & Estes, W. K. (1969). Effects of recall tests on long-term retention of paired associates. *Journal of Verbal Learning and Verbal Behavior, 8,* 463-470.

American Psychological Association (1992). Ethical principles of psychologists and code of conduct. *American Psychologist, 47,* 1597-1611.

Arbuckle, T. Y., & Cuddy, L. L. (1969). Discrimination of item strength at time of presentation. *Journal of Experimental Psychology, 1,* 126-131.

Benjamin, A. S., Bjork, R. A., & Schwartz, B. L. (1998). The mismeasure of memory: When retrieval fluency is misleading as a metamnemonic index. *Journal of Experimental Psychology: General, 127,* 55-68.

Birnbaum, I. M., & Eichner, J. T. (1971). Study versus test trials and long-term retention in free-recall learning. *Journal of Verbal Learning and Verbal Behavior, 12,* 516-521.

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.

Bjork, R. A. (1975). Retrieval as a memory modifier: An interpretation of negative recency and related phenomena. In R. L. Solso (Ed.), *Information processing and cognition: The Loyola symposium* (pp. 123-144). Hillsdale, NJ: Erlbaum.

Bjork, R. A. (1988). Retrieval practice and the maintenance of knowledge. In M. Gruneberg, P. Morris, & R. Sykes (Eds.), *Practical aspects of memory:*

Current research and issues, Vol. 2 (pp. 396-401). Chichester: John Wiley.

Bjork, R. A., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. In A. Healy, S. M. Kosslyn, & R. M. Shiffrin (Eds.), *From learning processes to cognitive processes: Essays in honor of William K. Estes, Vol. 2* (pp. 35-67). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Bregman, A. S., & Wiener, J. R. (1970). Effects of test trials in paired-associate and free-recall learning. *Journal of Verbal Learning and Verbal Behavior, 9,* 689-698.

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition, 20,* 633-642.

Cohen, R. L., & Johansson, B. S. (1967). The activity trace in immediate memory: A re-evaluation. *Journal of Verbal Learning and Verbal Behavior, 6,* 139-143.

Dempster, F. N. (1996). Distributing and managing the conditions of encoding and practice. In E. L. Bjork & R. A. Bjork (Eds.), *Memory* (pp.317-344). San Diego: Academic Press.

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

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 & Language, 33,* 545-565.

Dunlosky, J., & Nelson, T. O. (1997). Similarity between the cue for judgments of learning (JOL) and the cue for test is not the primary determinant of JOL

accuracy. *Journal of Memory & Language, 36,* 34-49.

Gardiner, J. M., & Klee, H. (1976). Memory for remembered events: An assessment of output monitoring in free recall. *Journal of Verbal Learning and Verbal Behavior, 15,* 227-233.

Glover, J. A. (1989). The "testing" phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology, 81,* 392-399.

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

Jang, Y., & Nelson, T. O. (2005). How many dimensions underlie judgments of learning and recall? Evidence from state-trace methodology. *Journal of Experimental Psychology: General, 134,* 308-326.

Kimball, D. R., & Metcalfe, J. (2003). Delaying judgments of learning affects memory, not metamemory. *Memory & Cognition, 31,* 918-929.

King, J. F., Zechmeister, E. B., & Shaughnessy, J. J. (1980). Judgments of knowing: The influence of retrieval practice. *American Journal of Psychology, 95,* 329-343.

Klee, H., & Gardiner, J. M. (1976). Memory for remembered events: Contrasting recall and recognition. *Journal of Verbal Learning and Verbal Behavior, 15,* 471-478.

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

Koriat, A., Bjork, R. A., Sheffer, L., & Bar, S. K. (2004). Predicting one's own

forgetting: The role of experience-based and theory-based processes. *Journal of Experimental Psychology: General, 133,* 643-656.

Koriat, A., & Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review, 103,* 490-517.

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

Koriat, A., & Shitzer-Reichert, R. (2002). Metacognitive judgments and their accuracy: Insights from the processes underlying judgments of learning in children. In P. Chambres, M. Izaute, & P. Marescaux (Eds.), *Metacognition: Process, function, & use* (pp. 1-17). Kluwer Academic Publishers.

Koriat, A., Sheffer, L., & Ma'ayan, H. (2002). Comparing objective and subjective learning curves: Judgments of learning exhibit increased underconfidence with practice. *Journal of Experimental Psychology: General, 131*, 147-162.

Lovelace, E. A. (1984). Metamemory: Monitoring future recallability during study. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 10*, 756-766.

Meeter, M., & Nelson, T. O. (2003). Multiple study trials and judgments of learning. *Acta Psychologica, 113,* 123-132.

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 (JOLs) are extremely accurate at predicting subsequent recall: The "delayed-JOL

effect." *Psychological Science, 2,* 267-270.

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.

Paivio, A., Yuille, J. C., & Madigan, S. A. (1968). Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of Experimental Psychology Monograph, 76* (1, Pt. 2).

Robinson, J. A., & Kulp, R. A. (1970). Knowledge of prior recall. *Journal of Verbal Learning and Verbal Behavior, 9,* 84-86.

Runquist, W. N. (1983). Some effects of remembering on forgetting. *Memory & Cognition, 11*, 641-650.

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.

Tulving, E., & Arbuckle, T. Y. (1963). Sources of intratrial interference in immediate recall of paired associates. *Journal of Verbal Learning and Verbal Behavior, 1,* 321-334.

Underwood, B. J. (1964). Degree of learning and the measurement of forgetting. *Journal of Verbal Learning and Verbal Behavior, 2,* 112-129.

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