Predicting memory: How study techniques influence delayed judgment-of-learning accuracy

A thesis submitted by

Gregory Isaac Hughes

in partial fulfillment of the requirements for the degree in

Master of Science

in

Psychology

TUFTS UNIVERSITY

May 2018

Advisor:

Dr. Ayanna K. Thomas

Committee Members:

Dr. Paul J. Muentener & Dr. Jonathan G. Tullis

Abstract

Multiple factors influence judgment-of-learning accuracy. One factor is timing: JOLs are most accurate when delayed from encoding (Nelson & Dunlosky, 1991). Another factor is type of encoding: some study techniques lead to higher JOL accuracy than others (Jang, Wallsten, & Huber, 2012). However, study techniques have only been shown to influence the accuracy of JOLs made immediately after encoding. I investigated the possibility that study techniques influence *delayed* JOL accuracy. In three experiments, participants made delayed JOLs after encoding material by passively reading (*study practice)*, generating keywords (*elaborative encoding*), or taking practice tests (*retrieval practice*). Retrieval practice led to higher JOL accuracy than study practice when the delay was 48 hr, but not 15 min. The effect of elaborative encoding, relative to study practice, was inconsistent across experiments. The results also suggest that study techniques influence JOL accuracy by affecting how much of the studied material is retrieved at the time of the JOL.

Key words: *metamemory, judgments of learning, encoding, retrieval practice*

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Predicting memory: How study techniques influence delayed judgment-of-learning accuracy

CHAPTER 1: INTRODUCTION

Without the supervision of a teacher, students must rely on metacognitive processes to guide crucial study decisions, such as what to study, how to study, and how long to study (Nelson & Narens, 1994). Metacognition refers to the dynamic process in which individuals evaluate their progress toward goals (*monitoring*) and then make decisions about how to achieve those goals (*control*; Nelson & Dunlosky, 1990). Research has consistently demonstrated that monitoring influences control (e.g., Koriat, Ma'ayan, Sheffer, & Bjork, 2006; see Son & Metcalfe, 2000 for review), and that the accuracy of monitoring influences the efficacy of self-regulated, or controlled study (Thiede, 1999; Thiede, Anderson, & Therriault, 2003). Unfortunately, research has also consistently demonstrated that monitoring is often inaccurate, and such inaccuracies have downstream consequences for control (e.g, Karpicke, 2009; Thomas & McDaniel, 2007). In the present study, I explore how different study techniques influence judgment of learning (JOL) accuracy.

A JOL is a prediction about the likelihood of remembering recently studied material on a future test. In a typical JOL experiment, learners study word pairs and then, on a numeric scale, estimate the likelihood that they will remember the target (second member of the pair) when later prompted with the cue (first member on the pair; Arbuckle & Cuddy, 1969; for a review see Rhodes,

2016). Finally, learners take a cued-recall or recognition test, which allows researchers to assess how well JOLs predict subsequent memory performance.

JOL Accuracy

JOL accuracy has traditionally been measured either by (a) calculating Goodman and Kruskal Gamma (γ) correlations between JOLs and later memory performance on an item-by-item basis (*relative accuracy*), or (b) by comparing mean JOLs to mean memory performance (*absolute accuracy*; for a review see Rhodes, 2016). Relative accuracy is perfect when the JOLs for items that are later remembered are always higher than the JOLs for items that are later forgotten. Absolute accuracy is perfect when there is no difference between mean JOLs and mean performance. In the present study, I focused on the factors that influence relative accuracy (hereafter referred to simply as JOL accuracy) because it is critical to the effective allocation of study time across items.

Delay from Encoding

Delaying JOLs from encoding is perhaps the most effective and reliable method of enhancing JOL accuracy. Using a paired-associate paradigm, Nelson and Dunlosky (1991) were the first to observe the *delayed-JOL effect*, in which JOLs made at a short delay from studying items (at least 30 s) were substantially more accurate than JOLs made immediately after studying items. The delayed-JOL effect is highly robust and has been widely replicated with a variety of materials and designs (Dunlosky & Nelson, 1992, 1994, 1997; Kelemen, 2000; Kelemen & Weaver, 1997; Finn & Metcalfe, 2007; Kimball & Metcalfe, 2003; Koriat & Ma'ayan, 2005; Meeter & Nelson, 2003; Nelson, Narens, & Dunlosky,

2004; Tullis, Finley, & Benjamin, 2013; van Overschelde & Nelson, 2006; see Rhodes & Tauber, 2011 for a review). In a meta-analysis, Rhodes and Tauber (2011) found that the mean weighted accuracy of delayed JOLs, as measured by item-by-item Goodman and Kruskal (γ) correlations between predictions and memory performance, was .77, whereas it was only .42 for JOLs made at no delay.

Several theories have been advanced to explain the delayed-JOL effect. These theories hinge on the idea that delaying JOLs from encoding changes the types of information that learners use to make their predictions. In the first study to demonstrate the effect, Nelson and Narens (1991) proposed the monitoring-dual-memories hypothesis. According to the monitoring-dual-memories hypothesis, learners base their JOLs on information available in both short-term memory and long-term memory. When JOLs are made during or immediately after encoding, the availability of transient information (such as the target) in short-term memory leads to the feeling that the knowledge will likewise be available in the future. However, the presence of knowledge in short-term memory is not indicative of its presence in long-term memory. By delaying JOLs from encoding by at least 30 s, learners must rely solely on information retrieved from long-term memory, which is much more diagnostic of whether knowledge will be available on a later test that also assesses the presence of knowledge in long-term memory. In support of this hypothesis, making the target available in short-term memory when soliciting delayed JOLs abolishes the delayed-JOL effect, presumably because re-presenting the target obviates the retrieval attempt.

For example, Dunlosky and Nelson (1992) found that delay benefits JOL accuracy only when these judgments are prompted by the cue only, and not the cue and target (see also Dunlosky & Nelson, 1992, 1994; Tullis et al., 2013).

Sikström and Jönsson (2005) developed the stochastic-drift model of delayed-JOL accuracy to account for findings that could not be explained by the monitoring-dual-memories hypothesis. Specifically, according the monitoring-dual-memories hypothesis, delaying JOLs beyond the point in which target information exits short-term memory (~30 s) should not further improve JOL accuracy. However, delaying JOLs beyond 30s improves JOL accuracy further. In a meta-analysis, Rhodes and Tauber (2011) found that longer delays (2 to 10 min) resulted in higher JOL accuracy than shorter delays (0 to 2 min). The stochastic-drift hypothesis eschews the traditional distinction between short-term and long-term memory in favor of fast and slow decaying memory traces. According to the stochastic-drift account, memory traces decay at different rates, and thus the longer one delays JOLs from encoding, the more stable, and thus more diagnostic of future memory, the information on which learners base their JOLs. This would explain the benefit of delaying JOLs beyond 30 s.

Finally, the self-fulfilling prophecy account asserts that the delayed-JOL effect owes to changes to memory, and not metacognitive, processes (Kimball & Metcalfe, 2003; Spellman & Bjork, 1992; Soderstrom, Clark, Halamish, & Bjork, 2016). Spellman and Bjork (1992) argue that delayed JOLs are more accurate than immediate JOLs because the act of judgment influences memory for the assessed material. To illustrate the argument, consider a learner who has retrieved

some target knowledge when making a delayed JOL. This learner is likely to express high confidence in their future memory of this item. However, retrieving the item increases the probability of future retrieval, and thus their high JOL has now been rendered more likely to be accurate. That is, the items that receive high JOLs are preferentially strengthened compared to those that receive low JOLs. When making immediate JOLs, the same memorial benefit does not occur because there is no opportunity to retrieve the target from long-term memory. In sum, the self-fulfilling prophecy hypothesis argues that the delay benefits JOL accuracy by potentiating the maintenance of retrieved knowledge. In support of the self-fulfilling prophecy, Kimball and Metcalfe (2003) found that re-exposing participants to the targets after learners made cue-only delayed JOLs reduced the accuracy of those JOLs. Presumably, re-exposing learners to the targets after failed retrieval attempts increased the memory for the initially unretrieved items, but did not benefit the memory for the retrieved items. Nelson et al. (2004) found that the increased accuracy of delayed JOLs owed primarily to distinguishing between retrieved and non-retrieved items, which also provides evidence for the self-fulfilling prophecy account.

One limitation of the self-fulfilling prophecy account is that it specifies retrieval of the target as the sole factor that influences delayed-JOL accuracy. Although research suggests that learners do rely heavily on target retrieval in making their delayed JOLs (e.g., Nelson et al., 2004; Pyc, Rawson, & Aschenbrenner, 2014), evidence suggests that other types of cues contribute to monitoring predictions more generally. The literature on feeling-of-knowing

judgments, which are delayed-JOLs for the future recognition of unretrieved items, has demonstrated that other cues, such as recollection of information associated with the target, influences memory prediction accuracy (Brewer, Marsh, Clarks-foos, & Meeks, 2010; Hertzog, Fulton, Sinclair, & Starlette, 2014; Hicks & Marsh, 2002; Thomas, Bulevich, & Dubois, 2011, 2012). In these experiments, I provide evidence that cues other than target retrieval can influence the accuracy of delayed JOLs.

Study Techniques

Another means of influencing JOL accuracy involves targeting how learners initially study material. A great deal of research has shown that immediate-JOL accuracy increases with the number of times learners review material, such as through repeated study (Lovelace, 1984; Mazzoni, Cornoldi, & Marchitelli, 1990; but see Meeter & Nelson, 2003; Koriat, 1997) or study-test practice (Koriat, 1997; Koriat, Ma'ayan, Sheffer, & Bjork, 2006; Koriat & Shitzer-Reichert, 2002; Leonesio & Nelson, 1990; Jang, Wallsten, & Huber, 2012). In these studies, JOLs are made at each study or study-test phase and are compared to a subsequent test to assess JOL accuracy. Notably, some study techniques result in higher JOL accuracy than others. For example, engaging in practice tests (*retrieval practice*) during learning has been shown to result in higher JOL accuracy compared to simply studying material (*study practice*; Lovelace, 1984; King, Zechmeister, & Shaugnessy, 1980; Shaughnessy & Zechmeister, 1992; but see Meeter & Nelson, 2003). King et al. (1980) found that learners who engaged in three study-test cycles of word pairs exhibited higher

JOL accuracy than those who repeatedly studied five times. This finding remains even when degree of learning is equated between retrieval practice and study practice (Shaughnessy & Zechmeister, 1992).

As with delay, how learners study material influences the types of information that they use to make JOLs. For example, one factor contributing to the benefit of retrieval practice over study practice is the availability of the memory-for-past-test (MPT) heuristic, in which learners predict their future memory based on their recollection of past performance. Past performance is diagnostic of future memory, and thus people who learn with retrieval practice have access to more information that is predictive of their future memory than those who learn via study practice (Finn & Metcalfe, 2007; Hertzog, Hines, & Touron, 2013; Serra & Ariel, 2014; Shaughnessy & Zechmeister, 1992). Finn and Metcalfe (2007) tested the MPT hypothesis by having learners engage in two study-JOL-test trials of word pairs. In each trial, learners studied all word pairs, made immediate JOLs for each, and then took a test. The authors found that Trial-2 JOLs were more related to previous (Trial 1) test performance than to subsequent (Trial 2) test performance. This indicates that participants relied on prior test performance to inform their JOLs while discounting the new learning that occurred in the second trial (see also Hertzog et al., 2013; Serra & Ariel, 2014). Elaborative encoding techniques, in which learners study material by creating *mediator* words or images that link new information with old knowledge, yields its own diagnostic cue. Success or failure in creating a mediator is predictive of later memory; further, recollecting the mediator itself is highly

predictive of future memory success, even when learners fail to retrieve the target (Dunlosky, Hertzog, & Powell-Moman, 2005). Recollection of mediator words has been shown to increase feeling-of-knowing accuracy (Hertzog et al., 2014).

Delay and Study Techniques

Compared to immediate JOLs, different kinds of study techniques have not been shown to influence delayed-JOL accuracy. However, the few studies that exist do not conclusively negate the possibility. Researchers have observed no differences in delayed-JOL accuracy when comparing interactive imagery to study practice (Begg, Duft, Lalonde, Melnick, & Sanvito, 1989), interactive imagery to separate imagery (Dunlosky & Nelson, 1994), and a single study trial to two spaced study trials (Dunlosky & Nelson, 1994), these null results are not surprising because comparing these study techniques do not lead to differences in immediate-JOL accuracy. However, other studies have compared study techniques that lead to differences in immediate-JOL accuracy and still did not find differences in delayed-JOL accuracy (Koriat, et al., 2006; Jang et al., 2012). For example, Koriat, et al. (2006) found that repeated retrieval practice did not result in increasingly higher delayed-JOL accuracy. Likewise, Jang et al. (2012) found no benefit of retrieval practice over study practice on delayed-JOL accuracy. However, the authors of both studies noted that delayed-JOL accuracy was at ceiling after the first study or retrieval practice trial, leaving no room for additional trials to benefit JOL accuracy or for differences to emerge between the two study techniques. In stark contrast to these studies, Peynircioğlu, Brandler, Hohman, and Knutson (2014) found that delayed JOLs for short piano pieces

were more accurate if those pieces were initially learned visually, or visually and aurally, compared to only aurally. Admittedly, there are many important differences between the previous studies that did not find an effect of study technique on JOL accuracy and the one presented by Peynircioğlu et al. (2014). Stimuli, type of study, and general modality differences make it challenging to directly compare this study with those that employ standard verbal-learning methods. However, these results suggest that how learners encode material may influence their delayed JOLs for that material.

CHAPTER 2: OVERVIEW OF THE PRESENT RESEARCH

The purpose of the present experiments was to investigate if study techniques differentially influence delayed-JOL accuracy. Such an investigation has both practical and theoretical significance. From a practical perspective, designing metacognitive interventions for student populations should capitalize on the benefit afforded by delay because it is the most robust, reliable, and efficient means of enhancing JOL accuracy. Consider that Koriat et al. (2006) found that it takes multiple study-test trials for immediate-JOL accuracy to match delayed-JOL accuracy following only a single study-test trial. If certain study techniques can enhance JOL accuracy beyond the potent benefits of delay, then this can yield highly efficient interventions; students could support their monitoring accuracy at the same time as they study material.

From a theoretical perspective, this investigation will provide important information about the factors that influence delayed-JOL accuracy. As previously discussed, a great deal of research demonstrates that learners rely on target retrieval when metacognitive monitoring predictions (e.g., Koriat & Ma'ayan, 2005; Nelson & Dunlosky, 1991; Nelson et al., 2004; Pyc et al., 2014). That is, learners tend to express high JOLs when they retrieve the target, and low JOLs when they do not retrieve the target. If learners rely solely on target retrievability to inform their delayed JOLs, then it is possible that type of initial encoding would not influence delayed-JOL accuracy. This is because no matter how learners study material, learners will have access to the cue of target retrieval (targets are always either retrieved or not retrieved). Observing differences in

delayed-JOL accuracy as a function of initial study technique would suggest that (a) learners rely on cues other than target retrievability, (b) these cues influence JOL accuracy, and (c) the accessibility of these cues depends on how learners originally encoded material. These cues would be *noncriterial*, which refers to any information retrieved about a target (i.e., the *criterial* detail) that is not the target itself (Parks, 2007), such as contextual details from the original encoding event. Of note, although the self-fulfilling prophecy account excludes the influence of noncriterial cues on delayed-JOL accuracy, both the monitoring-dual-memories hypothesis and stochastic-drift model do not.

A growing body of literature suggests that *noncriterial recollection* can influence JOL accuracy (Daniels, Toth, & Hertzog, 2009; McCabe & Soderstrom, 2011) as well as other monitoring judgments, such as retrospective confidence (Kelley & Sahakyan, 2003), and feeling-of-knowing, which are JOLs for unretrieved items (Hertzog, Dunlosky, & Sinclair, 2010; Hertzog et al., 2014; Hicks & Marsh, 2002; Thomas, Bulevich, & Dubois, 2011, 2012). Further, how learners encode material influences the recollection of noncriterial, contextual details (Cook, Marsh, & Hicks, 2006; Hertzog et al., 2010, 2014). For example, Hertzog et al. (2014) found not only that retrieving contextual details enhanced feeling-of-knowing accuracy, but also that increasing the number of times that learners studied material increased the probability of retrieving those contextual details. In Experiment 3, I directly tested the hypothesis that different study techniques influence delayed-JOL accuracy by influencing the accessibility of noncriterial recollection.

CHAPTER 3: EXPERIMENTS 1 AND 2

In Experiments 1 and 2, I examined delayed-JOL accuracy as a function of how participants initially studied material. Using a between-subjects design, participants learned weakly-related word pairs (e.g., mother – child) either by (a) studying these pairs twice (*study practice*), (b) studying these pairs once, followed by a practice test with feedback in the form of the correct answer (*retrieval practice*), or (c) by generating *mediator* words that thematically linked the members of the word pairs (e.g., family for mother – child; *elaborative encoding*). A retention interval followed. In Experiment 1, the retention interval was 15-min. In Experiment 2, the retention interval was 48 hr. After the retention interval, participants then made their JOLs, in which they estimated the likelihood of recognizing a target, when given the cue, from a list of four choices on a later test. After making their JOLs, participants took a four-alternative forced-choice test in which they were presented with each cue and asked to select the corresponding target amongst a list of four possible choices.

I hypothesized that retrieval practice would lead to higher delayed-JOL accuracy than study practice for several reasons. First, because the effect has been reliably demonstrated with immediate JOLs (Koriat et al., 2006). Second, because research suggests that, relative to study practice, retrieval practice increases the accessibility of noncriterial, contextual (Karpicke, Lehman, & Aue, 2014; Lehman, Smith, & Karpicke, 2014) and/or semantic cues (Carpenter, 2009, 2011; Pyc & Rawson, 2010) in long-term memory that can signal the presence of an unretrieved target in memory. Third, because retrieval practice strengthens

recollective processes relative to study practice, thus increasing the comparative likelihood of consciously retrieving noncriterial cues (Jinkun & Lixian, 2009). To reiterate, I expect that retrieval practice would lead to higher delayed-JOL accuracy than study practice because while both techniques yield the cue of target-retrievability, research suggests that retrieval practice will yield more noncriterial cues than study practice.

I made no directional hypothesis concerning the influence of elaborative encoding on delayed-JOL accuracy. There is an ongoing debate about whether retrieval practice and elaborative encoding work through similar or different mechanisms. This debate speaks to the type, and accessibility, of the noncriterial cues that these techniques might engender. According to the semantic elaboration hypothesis, retrieval practice is effective because it encourages more covert semantic elaboration than study practice (Carpenter, 2009, 2011; Pyc & Rawson, 2010). If so, then I would expect that retrieval practice and elaborative encoding would lead to similar, and higher, levels of delayed-JOL accuracy than study practice. However, recent research indicates that retrieval practice and elaborative encoding work through distinct mechanisms (Karpicke, Blunt, & Smith, 2016; Karpicke et al., 2014; Karpicke & Smith, 2012) and if so, retrieval practice and elaborative encoding might influence delayed-JOL accuracy differently.

I also hypothesized that the effects of study techniques on delayed-JOL accuracy would be larger after a longer delay than shorter delay. This is because differences in the influence of study techniques on memory often become more apparent as time passes. For example, although retrieval practice is a far more

potent study technique than study practice in the long-term, it can lead to equivalent or even lower memory performance in the short-term (Roediger & Karpicke, 2006; Tullis et al., 2013). Given that the differential effects of study techniques on long-term memory emerge over longer periods of time, I expected that their effects on metacognitive processes would likewise emerge after longer delays. That is, I expected that the noncriterial cues engendered during retrieval practice would persist longer in memory than those generated during study practice, thus leading to greater differences in delayed-JOL accuracy with time. Thus, in Experiment 1 I used a 15-min delay, and in Experiment 2 I used a 48-hr delay.

Contrary to most JOL studies, I had learners predict recognition performance rather than cued-recall performance. I did so because predictions of recognition are more difficult and typically less accurate than predictions of cued-recall (e.g., Thiede & Dunlosky, 1994; Weaver & Kelemen, 2003), thus minimizing the chances of observing the ceiling levels of JOL accuracy that might have masked the influence of study techniques on delayed-JOL accuracy in previous investigations (cf., Koriat et al., 2006; Jang et al., 2012).

Experiment 1 Method

Participants

Ninety participants volunteered from Tufts University (34 males, 56 females) aged 18 to 25 (*M*age = 18.80, *SD* = 2.42). Participants were either compensated with course credit or $10 per hour. I randomly assigned 30

participants to the study practice (study practice) group, the elaborative encoding (elaborative encoding) group, and the retrieval practice (retrieval practice) group.

Materials

Twenty-nine weakly-related word-pairs. The mean forward associative strength of the cue to target was 3% per the University of South Florida Free Association norms (Nelson, McEvoy, & Schreiber, 1998). These norms were also used to acquire weakly-related words to use as foils for the final four-alternative forced-choice test. The mean of forward associative strength from cue to foil was 6%. See Appendix A for a full list of the cues, targets, and foils.

Design

I used a between-subjects design consisting of three study technique groups: study practice, elaborative encoding, and retrieval practice.

Procedure

Participants were run in groups of three to six on laptop computers programmed with E-Prime software (Version 2.1; Schneider, Eschman, & Zuccolotto, 2002). The procedure consisted of five phases: initial study, strategic learning, retention interval, judgment-of-learning (JOL), and final test. Before the experimental session began, participants engaged in a practice session, during which they studied four word-pairs, then made JOLs, and finally took a four-alternative forced-choice test, in which they were presented with each cue word, and were required to select the corresponding target word on a list of four choices. The purpose of the practice phase was to familiarize participants with the instructions and the tasks. Participants were told that the tasks in the practice

phase and the experiment would be identical. See Appendix B for the instructions of the tasks.

Phase 1. In Phase 1, which was identical across all three groups, participants studied 26 word-pairs. Participants first received written and oral instructions informing them that they would see pairs of words one at a time for a short duration and that they should try to remember these words for a later memory test. Word pairs appeared one at a time for a 1000 ms each. Presentation order was randomized across participants.

Phase 2. Phase 2 immediately followed Phase 1, and differed across the three groups. For each group, participants received written and oral instructions. As in Phase 1, Phase 2 trials were randomized across participants.

Study practice. Participants in the study practice group read identical instructions as those presented in Phase 1. They then saw the same 26 word-pairs, one-at-a-time, for 1000 ms.

Elaborative Encoding. Participants in the elaborative encoding group were told that they would be studying the word pairs again, but this time would be asked to type in a new word for each pair that related to both the cue and target word. Participants in the elaborative encoding group were presented with the 26 word-pairs, one at a time, for 500 ms each. After each word pair disappeared, participants were asked to supply a word (*the mediator*) that thematically related the cue and the target (e.g., *Space* for the pair *Moon – Galaxy*). There was no time limit for responding. On average, it took participants 5436 ms to enter a mediator (*SD* = 1774 ms). Participants left 7.7% of responses blank.

Retrieval Practice. Participants in the retrieval practice group were told that they would be taking a test on the words they studied in Phase 1, in which they would see the cue and provide the target. Participants were encouraged to guess if they were not confident in a retrieved answer. The cue remained on the screen until an answer was provided and there was no time limit to respond. On average, it took participants 4221 ms to enter a response (*SD* = 1269 ms). Mean accuracy during retrieval practice was 40% (*SD* = 17%) on the cued-recall test and participants left only 9.5% of responses blank. After each trial, participants saw the complete word pair for 500 ms as feedback. Thus, the mean time on a given retrieval practice trial was approximately 4700 ms.

Phases 3 and 4. Phase 3 immediately followed Phase 2, and consisted of a 5-min retention interval in which participants performed a non-verbal drawing task. Participants received printed images of snowflakes and were asked to hand draw the snowflakes on blank paper. Immediately after the retention interval, Phase 4 began. Participants were told that they would see the cue from each of the studied pairs and would be instructed to estimate their likelihood of recognizing the corresponding target word on a multiple-choice test that would occur in about 5 min. Each JOL was prompted by the cue and the instruction to predict recognition on the final test. They provided their JOLs on a scale of 0 (*will not remember*) to 10 (*will definitely remember*). Participants were told that a rating of 5 indicates moderate levels of confidence and were encouraged to use the entire range of the scale. Judgments-of-learning were self-paced. Average time to enter a JOL was 2340 ms (*SD* = 742 ms) in the study practice group, 2856 ms (*SD* =

1343 ms) in the elaborative encoding group, and 2617 ms (*SD* = 750 ms) in the retrieval practice group.

Phase 5. Phase 5 consisted of a four-alternative forced-choice test. Cues from the 26 word-pairs were randomly presented one-at-a-time. Participants were asked to select the corresponding target word from a list of three other foils by typing in a numeral, ranging from 1 to 4, which corresponded with potential answers. There was no time limit to respond. Average time to enter a response on each test item was 3980 ms (*SD* = 914 ms) in the study practice group, 4053 ms (*SD* = 1104 ms) in the elaborative encoding group, and 3331 ms (*SD* = 700 ms) in the retrieval practice group.

Experiment 1 Results

Unless otherwise state, I used an alpha level of .05 to determine significance of all subsequent statistical tests and post-hoc tests were adjusted with a Bonferroni correction.

Test Performance

I conducted a one-way between-subjects analysis of variance (ANOVA) to compare the effects of study technique on four-alternative forced-choice test performance. There was a main effect of group, $F(2, 87) = 61.83$, $p < .001$, $\eta_p^2 = .59$. As shown in Table 1, retrieval practice ($M = .89$) resulted in higher test performance than study practice ($M = .58$), $t(59) = 9.58$, $p < .001$, $d = 1.24$, as did elaborative encoding ($M = .90$), $t(59) = 9.75$, $p < .001$, $d = 1.26$. No other differences were significant.

JOL Magnitude

I conducted a one-way between-subjects ANOVA to compare the effects of study technique on JOL magnitude. There was a main effect of group, $F(2, 87) = 7.33$, $p = .001$, $\eta_p^2 = .14$. As shown in Table 1, retrieval practice ($M = 5.88$) led to higher magnitude JOLs than study practice ($M = 4.72$), $t(59) = 3.04$, $p = .009$, $d = 0.39$, as did elaborative encoding ($M = 6.06$), $t(59) = 3.54$, $p = .002$, $d = 0.46$. No other differences were significant.

JOL Accuracy

To measure JOL accuracy, I calculated intra-individual γ correlations between each participant's JOLs and test accuracy on an item-by-item basis. These γ correlations range from -1.0 to 1.0, with higher values indicating better JOL accuracy than lower values. Such correlations were incomputable for seven participants from the retrieval practice group and three from the elaborative encoding group due to lack of variance in either the JOLs or performance on the four-alternative forced-choice test. Thus, these 10 participants could not be included in the analysis.

I conducted a one-way between-subjects ANOVA to compare the effects of study technique on JOL accuracy. There was no main effect of study-technique group, $F(2, 77) = 0.383$, $p = .683$, $\eta_p^2 = .01$. Refer to Table 1 for mean JOL accuracy for each study-technique group.

I conducted three one-sample *t*-tests to assess whether mean γ values for each group were statistically greater than 0, which would indicate above-chance JOL accuracy. Only the mean correlation coefficient associated with the study

practice group (M = .24) was statistically different than 0, $t(29) = 3.731$, $p = .001$, $d = 0.62$.

Bayesian Analysis of JOL Accuracy

Given null result of study technique on JOL accuracy, I conducted a Bayesian one-way ANOVA to gain direct evidence about the probability that this was a true null effect and not an issue of low power (see Rouder, Morey, Speckman, & Province, 2012; Wagenmakers et al., 2017 for discussion of Bayesian ANOVA). The analysis was conducted with the JASP software (Wagenmakers et al., 2017) and used the default recommended uniform prior values. Rather than a p value, the Bayesian ANOVA analysis yields a Bayes factor (BF_{10}), which represents the ratio of the probability of obtaining the data given the alternative hypothesis, relative to the probability of obtaining the data given the null hypothesis. The BF_{10} was .147, which represents strong evidence for the null hypothesis per the cutoffs outlined by Jeffreys (1961). It is unlikely that the null result was an issue of low power.

Experiment 1 Discussion

Contrary to my hypotheses, study techniques did not differentially influence delayed-JOL accuracy. It is possible that the failure to obtain this effect was due to ceiling levels of performance on the final test in both the retrieval practice and elaborative encoding groups (scores were .90 and .89, respectively). The limited variability in final-test performance likely suppressed the values of the γ correlation coefficients; that is, there simply was not much variability in memory performance for the JOLs to predict.

Despite the failure to observe differences in JOL accuracy across groups, there were differences in JOL magnitude across groups. The average JOLs of learners in the retrieval practice and elaborative encoding groups reflected the large memorial advantage over study practice (scores were .31 and .32 higher, respectively), which accords with previous findings (see Tullis et al., 2013 and Begg et al., 1989, respectively). Participants therefore exhibited metacognitive sensitivity to the relative efficacy of study techniques on memory retention.

Experiment 2

Experiment 2 was nearly identical to Experiment 1, except for two changes. First, I increased the retention interval between initial study and JOLs to 48 hr. As previously discussed, I expected that observing differences in JOL accuracy across study-technique groups would be more likely after a longer, rather than shorter, delay. The longer delay would also likely attenuate the ceiling levels of final-test performance observed in Experiment 1. Second, I controlled for differences in the time that participants were exposed to the cues across groups. That is, in Experiment 1, participants in the study practice and elaborative encoding groups saw the cues for a fixed interval, participants in the retrieval practice the cue until an answer was provided. In Experiment 2, participants were exposed to the cues for an equal amount of times in all groups.

Experiment 2 Method

Participants

Ninety-Six Tufts undergraduate students (24 men, 72 women) aged 18 to 25 (*M*age = 19.44, *SD* = 1.46) participated for course credit or $10 per hour. I

randomly assigned 32 participants to the study practice group, 33 participants to the elaborative encoding group, and 31 participants to the retrieval practice group.

Materials

I used the same materials as in Experiment 1.

Design

I used a between subjects-design with three groups: study practice, elaborative encoding, and retrieval practice.

Procedure

As outlined above, the procedure of Experiment 2 matched Experiment 1 except for two changes. First, I increased the Phase 3 retention interval from 15 min to 48 hr. Second, I equated exposure to each cue-word across groups. Participants in all groups were exposed to each cue-word for 2500 ms total per item, as detailed below. Participants began by completing a short practice session in which they practiced each experimental phase with three word-pairs. Presentation of items was randomized across all participants in all phases.

Phase 1. Participants studied all 26 word-pairs for 1000 ms each.

Phase 2. Phase 2 immediately followed Phase 1 and differed across the three groups.

Study Practice. Participants in the study practice group studied all 26 word-pairs for 1500 ms each.

Elaborative Encoding. Participants viewed all 26 word-pairs for 1500 each. After a given word pair disappeared, participants were asked to enter in

their mediator word for that word pair. On average, it took participants 4753 ms (*SD* = 2610 ms) to enter a mediator word. Participants in the elaborative encoding group left .5% of responses blank.

Retrieval Practice. Participants viewed each cue-word for 1000 ms. After each a given cue-word disappeared, participants were asked to enter in the corresponding target word. There was no time limit to respond. After entering a response, participants saw the intact word pair for 500 ms as feedback. On average, it took participants 2993 ms (*SD* = 1046 ms) to produce a response. Mean accuracy during retrieval practice was 33% (*SD* = 17%) and participants left 3.8% of responses blank.

Phase 3 and 4: JOL and Final Test. Participants then made self-paced JOLs after the 48-hour retention interval, in which they were presented the cue one at a time and made their prediction on a 0–10 scale. Average time to enter a JOL was 2736 ms (*SD* = 877 ms) in the study practice group, 2946 ms (*SD* = 936 ms) in the elaborative encoding group, and 3248 ms (*SD* = 1025 ms) in the retrieval practice group. Finally, participants took the self-paced, four-alternative forced-choice test. Average time to enter a response on each test item was 4746 ms (*SD* = 1277 ms) in the study practice group, 4702 ms (*SD* = 1297 ms) in the elaborative encoding group, and 3862 ms (*SD* = 1360 ms) in the retrieval practice group.

Experiment 2 Results

Test Performance

I conducted a one-way between-subjects ANOVA to compare the effects of study technique on four-alternative forced-choice test performance. There was a main effect of study-technique group, $F(2, 87) = 61.827$, $p < .001$, $\eta_p^2 = .53$. As shown in Table 2, retrieval practice ($M = .79$) resulted in higher test performance than study practice ($M = .47$), $t(62) = 9.80$, $p < .001$, $d = 1.23$, as did elaborative encoding ($M = .70$), $t(64) = 7.19$, $p < .001$, $d = 0.89$. Retrieval practice also led to higher test performance than elaborative encoding, $t(63) = 2.74$, $p = .022$, $d = 0.35$.

JOL Magnitude

I conducted a one-way between-subjects ANOVA to compare the effects of study technique on JOL magnitude. There was a main effect of group, $F(2, 87) = 7.33$, $p = .001$, $\eta_p^2 = .17$. As shown in Table 2, retrieval practice ($M = 5.08$) resulted in higher JOL magnitude than study practice ($M = 3.71$), $t(62) = 3.86$, $p = .001$, $d = 0.49$, as did elaborative encoding ($M = 4.98$), $t(64) = 3.65$, $p = .001$, $d = 0.62$. No other differences were significant.

JOL Accuracy

I could not compute γ correlations for three participants from the retrieval practice group and one for the elaborative encoding group due to no variation in JOLs or test performance, and thus these four participants were excluded from the analysis.

I conducted a one-way between-subjects ANOVA to compare the effects of study technique on JOL accuracy. There was a main effect of group, $F(2, 89) = 7.04$, $p = .001$, $\eta_p^2 = .14$. As shown in Table 2, retrieval practice ($M = .37$) resulted

higher JOL accuracy than study practice (M = .16), $t(59) = 2.87$, p = .015, d = 0.37, as did elaborative encoding (M = .41), $t(59) = 3.50$, p = .002, d = 0.45. No other differences were significant.

I conducted three one-sample *t*-tests to assess whether mean γ values for each group were statistically greater than 0, which would indicate above-chance JOL accuracy. The JOL accuracy of learners in the retrieval practice group (M = .37) was higher than 0, $t(27) = 7.65$, $p < .001$, d = 1.45, as it was for learners in the elaborative encoding group, $t(31) = 7.49$, $p < .001$, d = 1.32, and study practice group (M = .16), $t(31) = 7.49$, $p < .001$, d = 0.56.

Bayesian Analysis of JOL Accuracy

Given the null result between retrieval practice and elaborative encoding on delayed-JOL accuracy observed with the frequentist one-way ANOVA, I conducted a Bayesian one-way ANOVA to determine if the null result was likely due to low power. Consistent with the previous analysis, there was moderate evidence for the effect of group (BF_{10} = 23.81). Post-hoc pairwise comparisons were corrected for family-wise error by fixing the prior probability that the null hypothesis holds across comparisons to 0.5 (Westfall, Johnson, & Utts, 1997). The pairwise comparisons showed that there was moderate evidence that retrieval practice (BF_{10} = 10.08) and elaborative encoding (BF_{10} = 14.34) led to higher JOL accuracy than study practice. There was strong evidence for the null when comparing retrieval practice and elaborative encoding (BF_{10} = .30). Therefore, this null result does not appear to be an issue of power.

Experiment 2 Discussion

Study techniques differentially influenced delayed-JOL accuracy. Both retrieval practice and elaborative encoding led to higher JOL accuracy than study practice. As in Experiment 1, the average JOLs of learners reflected the large benefit of retrieval practice and elaborative encoding over study practice on final-test performance (scores were .32 and .29 higher, respectively). However, average JOLs did not reflect the smaller benefit of retrieval practice over elaborative encoding (a difference of .09).

The finding that JOL accuracy differed across study-technique groups provides evidence for the hypothesis that study techniques influence JOL accuracy by affecting the accessibility of noncriterial recollection. This is because, as previously discussed, the cue of target retrieval will always be accessible no matter how learners initially encoded material (targets will either by retrieved or not retrieved). Differences in JOL accuracy, therefore, likely owed to differences in the accessibility of noncriterial recollection across study-technique groups.

One potential confound of the results is differences in time spent learning the word pairs across study-technique groups. That is, participants in the retrieval practice and elaborative encoding groups spent more time studying the word pairs than participants in the study practice group. However, differences in time spent studying the word pairs cannot entirely explain the differences in JOL accuracy across groups—although participants in the elaborative encoding group studied the words pairs longer than those in the retrieval practice group, this did not lead to superior JOL accuracy.

CHAPTER 4: EXPERIMENT 3

The primary purpose of Experiment 3 was to determine how study techniques influence JOL accuracy. I tested the hypothesis that study techniques influence JOL accuracy by affecting the accessibility of two cues: target retrieval and noncriterial recollection. According to the noncriterial recollection hypothesis, individuals can grade the probability of later remembering unretrieved targets based on the whether they have access to noncriterial information associated with that unretrieved target. I hypothesized that retrieval practice and elaborative encoding would improve JOL accuracy relative to study practice by increasing the accessibility of noncriterial recollection at the time of the prediction. That is, I expected that (a) noncriterial recollection would be diagnostic of final-test performance, (b) rates of noncriterial recollection would be highest in the retrieval practice and elaborative encoding groups, and (c) rates of noncriterial recollection would be positively related to JOL accuracy.

I measured target retrieval by having participants make a cued-recall attempt of the target prior to each JOL, and accessibility of noncriterial recollection with the remember-know procedure. In a remember-know task, participants indicate whether a given stimulus evokes the recollection of contextual details pertaining to the initial encoding event (*remember*), merely a feeling of familiarity without access to contextual details (*know*), or nothing at all (*no memory*; Tulving, 1985). Using a remember-know task, Isingrini and colleagues (2016) found that the accuracy of feeling-of-knowing judgments was

positively related to the accessibility of noncriterial recollection at the time of making the prediction.

The secondary aim of Experiment 3 was to determine if the influence of study techniques on JOL accuracy depends on the associative relatedness of word pairs. In addition to weakly-related word pairs (e.g., Throne – Castle), participants also studied unrelated word pairs (e.g., Divorce – Allergy). I hypothesized that the effects of study techniques on (a) JOL accuracy and, (b) accessibility of noncriterial recollection, would be greater with weakly-related compared to unrelated word pairs. This hypothesis was based on the idea that elaborative encoding and retrieval practice would be easier, and thus more effective, with weakly-related word pairs. That is, I expected that it would be easier to generate a mediator word, or retrieve a target, during encoding of weakly-related pairs. As the effectiveness of retrieval practice and elaborative encoding increases relative to study practice, so too should differences in metamemory between these groups increase.

The third aim of Experiment 3 was to examine the influence of correct guessing on JOL accuracy. As previously discussed, JOL accuracy is generally lower when a recognition test, rather than a cued-recall test, is used as the criterion measure of memory. This effect owes to (a) the difficulty of anticipating the influence of automatic memory processes that play a large role in recognition tests (Undorf, Böhm, & Cüpper, 2016), and (b) correctly guessing the answer on the recognition test, which introduces noise into the measurement of JOL accuracy (Schwartz & Metcalfe, 1994; Thiede & Dunlosky, 1994). The reduced

accuracy of JOLs for recognition compared to cued-recall tests is therefore partially an artifact of design. In Experiment 3, participants indicated whether their response on the final test was a guess, which permitted the evaluation of the influence of correct guessing on the final test.

The final aim was to replicate the effects of study techniques on JOL accuracy while improving the methodology of Experiment 2. First, I equated the amount of time that participants in each study-technique group encoded each word pair. In Experiment 2, participants spent more time studying the word pairs in the retrieval practice and elaborative encoding groups compared to those in the study practice group, which was a confound of the design. In Experiment 3, all groups studied each word pair for a total of 12 s. Second, I used a new set of weakly-related word pairs and foils that better controlled for type of speech, concreteness, and length. Third, I changed the scale used to measure JOLs from ordinal (0 to 10) to interval (0 to 100%) to avoid ambiguity in interpretation of the scale (cf. Son & Metcalfe, 2005).

Experiment 3 Method

Participants

One-hundred and twenty participants volunteered from Tufts University (50 men, 70 women) aged 18 to 25 (*M*age = 19.11, *SD* = 1.46). Participants were either compensated with course credit or $10 per hour. An a priori power analysis indicated that to achieve a power of 80% with an alpha rate of .05, a minimum of 72 participants were needed to replicate the effects of study techniques on JOL accuracy for weakly-related items. I randomly assigned participants to the study

practice group, the elaborative encoding group, and the retrieval practice group. As explained in the method section, 21 participants could not be included in the analyses because they did not demonstrate understanding of the remember/know instructions. This drop-out rate was comparable to other remember/know studies (cf. McCabe & Geraci, 2009). As a result, there were 31 participants in the study practice group, 32 in the elaborative encoding group, and 36 in the retrieval practice group.

Materials

Participants studied 44 noun-noun, English word-pairs. Half of these pairs were weakly-related (e.g., *Throne – Castle)* and the remainder were unrelated (e.g., *Allergy – Divorce).* Item-relatedness was operationalized in terms of forward associative strength per the University of South Florida Free Association norms (Nelson, McEvoy, & Schreiber, 1998). The mean forward associative strength from cue to target was 2% (*SD* = 1%) for weakly-related pairs, and 0% for unrelated pairs. For each word pair, I selected three words to use as foils on the four-alternative-forced-choice test, which totaled to 132 words. Average forward associative strength from cue to foil was 4% (SD = 5%) for weakly-related pairs, and 0% for unrelated pairs. All cues, targets, and foils were nouns, ranged in length from four to eight letters, and were selected for high concreteness, which I operationalized as a value of 4 or greater on a scale ranging from 1 (highly abstract) to 7 (highly concrete). Concreteness values were obtained from the South Florida Free Association Norms. Cues were not associated with

the targets or foils of any other pair. See Appendix C for a full list of the cues, targets, and foils.

Procedure

Participants were run in groups of three to six on laptop computers programmed with E-Prime software (Version 2.1; Schneider, Eschman, & Zuccolotto, 2002). The procedure consisted of five phases: initial study, strategic learning, retention interval, judgment-of-learning (JOL), and final test. Prior to each of the two experimental sessions, participants engaged in a practice session with four word-pairs. On the first day, participants practiced studying the word pairs, which varied across study-technique groups. On the second day, participants practice the JOL phase, which included a cued-recall test, making JOLs, making remember/know judgments, and the four-alternative forced-choice test. Participants were told that the tasks in the practice phase and the experiment would be identical.

The primary purpose of the practice phase was to familiarize participants with the instructions and the tasks. The secondary purpose of the practice phase was to include a manipulation check to determine if participants understood the remember/know instructions. Following the recommendations of Yonelinas, Aly, Wang, and Koen (2010), I asked participants to provide written explanations for their remember/know responses after making each of the four remember/know judgments during practice. Only those participants who provided answers consistent with the provided definitions were included in the subsequent analyses.

Phase 1. In Phase 1, which was identical across all three groups, participants studied 40 word-pairs. Participants first received written and oral instructions informing them that they would be see pairs of words one at a time for a short duration and that they should try to remember these words for a later memory test. Word pairs appeared one at a time for a 6000 ms each. Presentation order was randomized across participants.

Phase 2. Phase 2 immediately followed Phase 1, and differed across the three groups. For each group, participants received written and oral instructions. As in Phase 1, Phase 2 trials were randomized across participants.

Study practice. Participants in the study practice group read identical instructions as those presented in Phase 1. They then saw the same 40 word-pairs, one-at-a-time, for 6000 ms.

Elaborative Encoding. Participants in the elaborative encoding group were told that they would be studying the word pairs again, but this time would be asked to type in a new word for each pair that related to both the cue and target word. Participants in the elaborative encoding group were presented with the 40 word-pairs, one at a time, for 6000 ms each. While the word pair was on the screen, participants were asked to supply a word (*the mediator*) that thematically related the cue and the target (e.g., *Space* for the pair *Moon – Galaxy*). There was no time limit for responding. Participants left 6.0% of responses blank.

Retrieval Practice. Participants in the retrieval practice group were told that they would be taking a test on the words they studied in Phase 1, in which they would see the cue and provide the target. Participants were encouraged to

guess if they were not confident in a retrieved answer. Participants in the retrieval practice group were presented with the cues from all 40 word-pairs, one at a time, for 5000 ms each. While the cue remained on the screen, the participant were asked to provide the corresponding target. After the 5000 ms elapsed, participants viewed the intact word pair for 1000 ms. Mean accuracy during retrieval practice was 58% (*SD* = 23%) on the cued-recall test and participants left 21% of responses blank.

Phases 3 and 4. Phase 3 immediately followed Phase 2, and consisted of a 48-hr retention interval. After the retention interval, participants made their JOLs. Judgments of learning were made using the pre-recall and monitoring procedure (Nelson et al., 2004). Participants were presented with the cue from each of the 40 word-pairs, one at a time, and were asked to provide the corresponding target. Participants were instructed to type the word "blank" if they could not retrieve the target. Average time to enter a cued-recall response was 6668 ms (*SD* = 2239 ms) in the study practice group, 7772 ms (*SD* = 3086 ms) in the elaborative encoding group, and 6334 ms (*SD* = 2291 ms) in the retrieval practice group.

After entering their response, participants made their JOLs on a scale of 0 – 100%, in which they estimated the probability of correctly recognizing the corresponding target on a list of four choices, without guessing, on a test that would occur in about 5 min. Average time to enter a JOL was 3081 ms (*SD* = 983 ms) in the study practice group, 2987 ms (*SD* = 986 ms) in the elaborative encoding group, and 2784 ms (*SD* = 644 ms) in the retrieval practice group.

After making their JOL, participants were asked to indicate whether the presented cue evoked a remember, know, or no memory experience. Instructions were adapted from Geraci, McCabe, and Guillory (2009). Remember responses were defined as the conscious recollection of some aspect of the original encoding experience. Participants were told to reply with a remember response only if they could provide details of what was remembered if asked by the experimenter. Know responses were defined as the feeling that a given cue had been encountered before during Phase 1, but did not evoke the recollection of specific information pertaining to the encoding event. A response of no memory was to be used when a given cue evoked neither the recollection of specific details, nor a feeling of familiarity that the cue had been presented during Phase 1. Participants were further instructed that remember, know, and no memory responses did not represent different levels of confidence of future recognition. To avoid interpretational ambiguity of the terms "remember" and "know," I substituted these terms with "Type 1" and "Type 2" memory, respectively, in the instructions (see, McCabe & Geraci, 2009). See Appendix D for the remember/know instructions. After participants read the remember/know instructions, the experimenter asked whether participants felt they understood the instructions, and then to explain what each response meant in their own words and with examples. If the participant did not understand or did not answer correctly, the experimenter reread the written instructions with no further elaboration. Average time to enter a remember/know response was 1925 ms (*SD* = 505 ms) in the study practice

group, 2155 ms (*SD* = 756 ms) in the elaborative encoding group, and 2342 ms (*SD* = 1133 ms) in the retrieval practice group.

All tasks in Phase 4 were self-paced and word pairs were presented randomly. To summarize, for each word pair, participants made a cued-recall attempt, JOL, and remember/know response before proceeding to the next word pair.

Phase 5. Phase 5 consisted of a four-alternative multiple-choice test. Cues from the 40 word-pairs were randomly presented one at a time. Participants were asked to select the corresponding target word from a list of three other distractors by typing in a numeral, ranging from 1 to 4, which corresponded with potential answers. There was no time limit to respond. Average time to enter a response on each final-test item was 4941 ms (*SD* = 1384 ms) in the study practice group, 4749 ms (*SD* = 1624 ms) in the elaborative encoding group, and 4122 ms (*SD* = 1025 ms) in the retrieval practice group. Performance on the final test was measured as the proportion of correct responses. After each response, participants entered their confidence in their answer on a scale of *guess, low*, *medium,* and *high.* There was no time limit to respond. Average time to enter a confidence response on each test item was 1272 ms (*SD* = 402 ms) in the study practice group, 1317 ms (*SD* = 422 ms) in the elaborative encoding group, and 1268 ms (*SD* = 375 ms) in the retrieval practice group.

Experiment 3 Results

All analyses used an alpha-rate of .05, and all post-hoc pairwise comparisons were adjusted with a Bonferroni correction.

JOL Magnitude

Judgment-of-learning magnitude was measured as the mean JOL for each participant. Means were calculated separately for unrelated and weakly-related items. I conducted a 3 (study technique: study practice, elaborative encoding, retrieval practice) x 2 (item type: unrelated, weakly-related) mixed-groups factorial ANOVA on JOL magnitude. There was a main effect of study-technique group, $F(2,96)= 3.13$, $p = .048$, $\eta_p^2 = .06$, and item type, $F(1,96) = 87.61$, $p < .001$, $\eta_p^2 = .48$, as well as an interaction between study-technique group and item type, $F(2,96) = 10.46$, $p < .001$, $\eta_p^2 = .18$.

I followed-up the interaction with simple effects analysis, which showed that study-technique group did not influence JOL magnitude for unrelated items, $F(2,96) = 0.95$, $p = .392$, $\eta_p^2 = .02$, but did influence JOL magnitude for weakly-related items, $F(2,96) = 6.93$, $p = .002$, $\eta_p^2 = .13$. Post-hoc pairwise comparisons showed that retrieval practice ($M = 66.01$) led to higher JOL magnitude for weakly-related items than study practice ($M = 50.45$), $t(66) = 3.33$, $p = .004$, $d = 0.82$, as did elaborative encoding ($M = 65.72$), $t(62) = 3.16$, $p = .006$, $d = 0.80$. Retrieval practice did not lead to higher JOL magnitude for weakly-related items than elaborative encoding ($p > .05$).

Further, post-hoc comparisons showed that within each study-technique group, judgment-of-learning magnitude was higher for weakly-related than unrelated items for retrieval practice group, $t(35) = 8.68$, $p < .001$, $d = 0.68$, and elaborative encoding group, $t(31) = 5.92$, $p < .001$, $d = 2.13$, but not the study practice group, $t(30) = 1.86$, $p = .066$, $d = 2.98$.

Final-Test Performance

Final-test performance was measured as the proportion of correct responses on the four-alternative forced-choice test. Proportions were calculated separately for unrelated and weakly-related items. I conducted a 3 (study technique: study practice, elaborative encoding, retrieval practice) x 2 (item type: unrelated, weakly-related) mixed-groups factorial ANOVA on JOL magnitude. There was a main effect of study-technique group, $F(2,96) = 9.208$, $p < .001$, $\eta_p^2 = .16$. Neither item type, $F(1,96) = 0.05$, $p = .818$, $\eta_p^2 = .00$, nor the interaction between study-technique group and item type, $F(2,96) = 1.79$, $p = .172$, $\eta_p^2 = .04$, were significant. Post-hoc comparisons showed that retrieval practice ($M = .85$) led to higher final-test performance than study practice ($M = .69$), $t(66) = 39.81$, $p < .001$, $d = 9.80$, as did elaborative encoding ($M = .84$), $t(62) = 40.95$, $p < .001$, $d = 10.40$.

JOL Accuracy

As in Experiments 1 and 2, I calculated γ correlations between each participant's JOLs and final-test performance on an item-by-item basis. These γ correlations separately for unrelated items and weakly-related items. Due to invariance either in JOLs or final-test performance, γ correlations for weakly-related and/or unrelated items could not be computed for seven participants in the study practice group, eight participants in the elaborative encoding group, and 12 participants in the retrieval practice group. Refer to Table 3 for the means.

I conducted a 3 (study technique: study practice, elaborative encoding, retrieval practice) x 2 (item type: unrelated, weakly-related) mixed-groups

factorial ANOVA on judgment-of-learning accuracy. There was no main effect of study-technique group, $F(2,69)= 1.27$, $p = .288$, $\eta_p^2 = .04$. There was a main effect of item type, $F(1,69) = 8.17$, $p = .006$, $\eta_p^2 = .11$, as well as an interaction between study-technique group and item type, $F(2,69) = 3.35$, $p = .041$, $\eta_p^2 = .09$.

Simple main effects analysis showed that study-technique group did not influence JOL accuracy for unrelated items, $F(2,69) = 0.10$, $p = .902$, $\eta_p^2 = .00$, but did influence JOL accuracy for weakly-related items, $F(2,69) = 5.83$, $p = .005$, $\eta_p^2 = .15$. Pairwise comparisons showed that retrieval practice ($M = .59$) led to higher JOL accuracy for related items than study practice ($M = .31$), $t(54) = 3.40$, $p = .003$, $d = .93$. Elaborative encoding ($M = .45$) did not lead to different levels of JOL accuracy for related items than retrieval practice and study practice (p's > .05).

Pairwise comparisons between JOL accuracy of unrelated and weakly-related items showed that in the retrieval practice group, JOL accuracy of unrelated ($M = .28$) items was lower than of weakly-related items ($M = .59$), $t(47) = -3.56$, $p < .001$, $d = 1.49$. Judgment-of-learning accuracy did not differ between unrelated and weakly-related items in either the study practice or elaborative encoding group (p's > .05).

I also conducted one-sample t-tests to determine if JOL accuracy for unrelated and weakly-related items was greater than 0 in each group, which would indicate above-chance prediction accuracy. For unrelated items, JOL accuracy was greater than 0 in the study practice group, $t(23) = 4.37$, $p < .001$, $d = 0.89$, elaborative encoding group, $t(24) = 4.27$, $p < .001$, $d = 0.85$, and retrieval

practice group, $t(28) = 3.64$, $p < .001$, $d = 0.68$. For weakly-related items, JOL accuracy was greater than 0 in the study practice group, $t(25) = 5.93$, $p < .001$, $d = 1.16$, elaborative encoding group, $t(27) = 6.80$, $p < .001$, $d = 1.29$, and retrieval practice group, $t(27) = 10.62$, $p < .001$, $d = 1.94$.

Bayesian Analysis of JOL Accuracy

The frequentist mixed ANOVA and post-hoc tests yielded two sets of null results that merit Bayesian follow-up: (a) no effect of study technique on JOL accuracy for unrelated items, and (b) no difference in JOL accuracy for weakly-related items when comparing elaborative encoding to study practice or retrieval practice. Therefore, I conducted a Bayesian 3x2 mixed ANOVA. Consistent with the previous analysis, there was strong evidence for the null for the main effect of group ($BF_{10} = .07$), moderate evidence for the alternative hypothesis for the main effect of item type ($BF_{10} = 7.00$), and moderate evidence for the interaction between study-technique group and item type ($BF_{10} = 3.10$) per the cutoffs of Jeffreys (1961). I followed-up the interaction by conducting separate one-way Bayesian ANOVAs for unrelated and weakly-related items. For unrelated items, there was strong evidence for the null ($BF_{10} = .132$). The null result is unlikely to be an issue of power. For weakly-related items, there was moderate evidence for the alternative hypothesis ($BF_{10} = 5.44$). Post-hoc pairwise comparisons were corrected for family-wise error by fixing the prior probability that the null hypothesis holds across comparisons to 0.5 (Westfall, Johnson, & Utts, 1997). The pairwise comparisons showed that there was moderate evidence that retrieval practice led to higher JOL accuracy for weakly-related items than study practice

(BF_{10} = 15.18), and anecdotal evidence for a difference with elaborative encoding (BF_{10} = 1.31). There was only anecdotal evidence for a difference between elaborative encoding and study practice (BF_{10} = .50). The results of the pairwise comparisons involving elaborative encoding suggests that the null findings could be due to low power.

Correct Guessing and JOL Accuracy

I examined the influence of correct guessing on JOL accuracy by factoring out trials in which participants answered correctly on the final test, but indicated that their response was a guess (cf. Thiede & Dunlosky, 1994). I then recalculated γ correlations between JOLs and final-test performance on an item-by-item basis for both unrelated and weakly-related items. The means slightly increased in all groups. I conducted the same 3x2 ANOVA as above with the recalculated γ correlations, which demonstrated the same pattern of results. As before, there was no main effect of study-technique group, $F(2,63)= .567$, $p = .570$, $\eta_p^2 = .00$, but there was a main effect of item type, $F(1,63) = 4.771$, $p = .038$, $\eta_p^2 = .07$, as well as an interaction between study-technique group and item type, $F(2,63) = 3.801$, $p = .028$, $\eta_p^2 = .11$. Again, study techniques did not influence JOL accuracy for unrelated items, $F(2,63) = .740$, $p = .740$, $\eta_p^2 = .01$, but did for weakly-related items, $F(2,63) = 4.02$, $p = .023$, $\eta_p^2 = .10$. Bonferroni-corrected, pairwise comparisons showed that retrieval practice ($M = .60$) led to higher JOL accuracy for related items than study practice ($M = .34$), $t(41) = 2.8$, $p = .02$, $d = 0.76$. Elaborative encoding ($M = .46$) did not lead to different levels of JOL accuracy for related items than retrieval practice or study practice (p's > .05).

Confidence Accuracy

To measure confidence accuracy, I calculated γ correlations between confidence ratings and final-test performance on an item-by-item basis. The γ correlations were calculated separately for unrelated and weakly-related items. Due to invariance either in confidence ratings or final-test performance, γ correlations could not be computed for seven participants in the study practice group, eight participants in the elaborative encoding group, and twelve participants in the retrieval practice group. Refer to Table 3 for the means.

I conducted a 3 (study technique: study practice, elaborative encoding, retrieval practice) x 2 (item type: unrelated, weakly-related) mixed-groups factorial ANOVA on judgment-of-learning accuracy. There was a main effect of study-technique group, $F(2,70) = 5.26$, $p = .007$, $\eta_p^2 = .13$. There was no main effect of item type, $F(1,70) = 2.47$, $p = .121$, $\eta_p^2 = .03$, nor an interaction between study-technique group and item type, $F(2,70) = 1.45$, $p = .242$, $\eta_p^2 = .04$. Pairwise comparisons showed that retrieval practice ($M = .81$) led to higher confidence accuracy than study practice ($M = .60$), $t(48) = 7.44$, $p = .009$, $d = 2.15$. No other differences were significant (*p's* > .05).

I also conducted one-sample t-tests to determine if confidence accuracy for unrelated and weakly-related items was greater than 0 in each group. For unrelated items, confidence accuracy was greater than 0 in the study practice group, $t(24) = 9.24$, $p < .001$, $d = 1.85$, elaborative encoding group, $t(24) = 16.84$, $p < .001$, $d = 3.37$, and retrieval practice group, $t(28) = 21.64$, $p < .001$, $d = 4.02$. For weakly-related items, confidence accuracy was greater than 0 in the study

practice group, $t(26) = 8.85$, $p < .001$, $d = 1.70$, elaborative encoding group, $t(27) = 9.47$, $p < .001$, $d = 1.79$, and retrieval practice group, $t(27) = 7.25$, $p < .001$, $d = 1.37$.

Cue Accessibility

Accessibility of Target Retrieval. Accessibility of target retrieval was measured as the proportion of targets produced during the cued-recall test prior to making each JOL. Proportions were calculated separately for unrelated and weakly-related items. Refer to Table 3 for the means.

I conducted a 3 (study technique: study practice, elaborative encoding, retrieval practice) x 2 (item type: unrelated, weakly-related) mixed-groups factorial ANOVA on target retrieval. There was a main effect of study-technique group, $F(2,96)= 5.42$, $p = .006$, $\eta_p^2 = .10$, and item type, $F(1,96) = 163.24$, $p < .001$, $\eta_p^2 = .63$, as well as an interaction between study-technique group and item type, $F(2,96) = 14.90$, $p < .001$ $\eta_p^2 = .24$.

Simple main effects analysis showed that study-technique group did not influence target retrieval for unrelated items, $F(2,96) = 1.21$, $p = .303$, $\eta_p^2 = .03$, but did influence target retrieval for weakly-related items, $F(2,96) = 11.51$, $p < .001$, $\eta_p^2 = .19$. Pairwise comparisons adjusted showed that retrieval practice ($M = .56$) led to higher target retrieval of weakly-related items than study practice ($M = .31$), $t(66) = 4.76$, $p < .001$, $d = 1.17$, as did elaborative encoding ($M = .45$), $t(62) = 2.85$, $p = .017$, $d = 0.72$. Retrieval practice did not lead to higher target retrieval for related items than elaborative encoding ($p > .05$).

I also conducted pairwise comparisons between accessibility of target retrieval of weakly-related and unrelated items within each study-technique. Accessibility of target retrieval was higher for weakly-related items than unrelated items in the study practice group, $t(30) = 3.79$, $p < .001$, $d = 4.04$, elaborative encoding group, $t(31) = 6.62$, $p < .001$, $d = 2.38$, and retrieval practice group, $t(35) = -11.96$, $p < .001$, $d = 1.39$.

Accessibility of Noncriterial Recollection. Accessibility of noncriterial recollection was measured as the proportion of remember responses during the JOL phase on trials (on trials that participants did not retrieve the target). Proportions were calculated separately for unrelated and weakly-related items. Refer to Table 3 for the means.

I conducted a 3 (study technique: study practice, elaborative encoding, retrieval practice) x 2 (item type: unrelated, weakly-related) mixed-groups factorial ANOVA on accessibility of noncriterial recollection. There was a main effect of study-technique group, $F(2,96) = 4.97$, $p = .009$, $\eta_p^2 = .09$, a main effect of item type, $F(1,96) = 10.92$, $p = .001$, $\eta_p^2 = .10$, as well as an interaction between study-technique group and item type, $F(2,96) = 6.06$, $p = .003$, $\eta_p^2 = .11$.

Simple main effects analysis showed that study-technique group did not influence accessibility of noncriterial recollection for unrelated items, $F(2,96) = 1.54$, $p = .219$, $\eta_p^2 = .03$, but did for weakly-related items, $F(2,96) = 8.05$, $p = .001$, $\eta_p^2 = .14$. Pairwise comparisons showed that for weakly-related items, elaborative encoding ($M = .51$) led to higher accessibility of noncriterial recollection than study practice ($M = .30$), $t(62) = 3.52$, $p = .002$, $d = 0.89$, and

retrieval practice (M = .29), $t(67) = 3.48$, $p = .002$, $p = 0.85$. The difference between retrieval practice and study practice was not significance ($p > .05$).

I also conducted pairwise comparisons between accessibility of noncriterial recollection of unrelated and weakly-related word pairs within each study-technique. In the elaborative encoding group, noncriterial recollection was higher for weakly-related items (M = .34) than unrelated items (M = .51), $t(31) = 4.58$, $p < .001$, $d = 1.64$. Noncriterial recollection did not significantly differ across item types in the study practice and retrieval practice groups, (p's > .05).

Cue Utilization

Utilization of Target Retrieval. To measure the degree to which participants relied on target retrieval to make their JOLs, I calculated γ correlations between target retrieval during the JOL phase (coded as 0 for unretrieved or 1 for retrieved) and JOLs on an item-by-item basis for each participant. Due to invariance either in targets retrieved (no targets or all targets retrieved) or final-test performance, γ correlations for weakly-related and/or unrelated items could not be computed for 14 participants in the study practice group, four participants in the elaborative encoding group, and four participants in the retrieval practice group. Refer to Table 4 for the means.

I conducted a 3 (study technique: study practice, elaborative encoding, retrieval practice) x 2 (item type: unrelated, weakly-related) mixed-groups factorial ANOVA on cue-utilization of target retrieval. There was a main effect of study-technique group, $F(2,74) = 7.14$, $p = .001$, $\eta_p^2 = .16$, and item type, $F(1,74)$

= 51.29, $p < .001$, $\eta_p^2 = .41$, as well as an interaction between study-technique group and item type, $F(2,74) = 3.14$, $p = .049$ $\eta_p^2 = .08$.

Simple main effects analysis showed that study techniques influenced cue utilization of target retrieval for unrelated items, $F(2,74) = 4.45$, $p = .015$, $\eta_p^2 = .11$, and for weakly-related items, $F(2,74) = 5.454$, $p = .006$, $\eta_p^2 = .13$. I followed up both significant main effects with pairwise comparisons, which showed that for unrelated items, retrieval practice ($M = .99$) led to higher cue utilization of target retrieval than elaborative encoding ($M = .95$), $t(59) = 2.69$, $p = .028$, $d = 0.70$, but not study practice ($M = .99$; $p > .05$). For weakly-related items, the same pattern held; retrieval practice ($M = .92$) led to higher cue utilization of target retrieval than elaborative encoding ($M = .77$), $t(59) = 3.20$, $p = .005$, $d = 0.83$,but not study practice ($M = .81$; $p > .05$).

I also conducted pairwise comparisons between cue utilization of unrelated and weakly-related word pairs within each study-technique group. Utilization of target retrieval was higher for unrelated items than weakly-related items in the study practice group, $t(47) = 4.32$, $p < .001$, $d = 1.09$, elaborative encoding group, $t(47) = 5.53$, $p < .001$, $d = 2.31$, and retrieval practice group, $t(47) = 2.60$, $p = .011$, $d = 1.80$.

Utilization of Noncriterial Recollection. To measure the degree to which participants relied on noncriterial recollection to make their JOLs, I calculated γ correlations between noncriterial recollection (coded as 0 for no noncriterial recollection or 1 for noncriterial recollection) and JOLs on an item-by-item basis for each participant. These correlations were restricted to cases in which the

participants did not retrieve the target. Due to invariance either in noncriterial recollection (remember responses for 0 or all trials) or final-test performance, γ correlations for weakly-related and/or unrelated items could not be computed for 11 participants in the study practice group, seven participants in the elaborative encoding group, and 12 participants in the retrieval practice group. Refer to Table 4 for the means.

I conducted a 3 (study technique: study practice, elaborative encoding, retrieval practice) x 2 (item type: unrelated, weakly-related) mixed-groups factorial ANOVA on utilization of noncriterial recollection. There was no main effect of study-technique group, $F(2,67) = 0.32$, $p = .728$, $\eta_p^2 = .00$, item type, $F(1,67) = 0.21$, $p = .645$, $\eta_p^2 = .00$, or interaction between study-technique group and item type, $F(2,96) = 0.09$, $p = .911$, $\eta_p^2 = .00$.

I also conducted one-sample *t*-tests to determine if the utilization of noncriterial recollection was greater than 0 for each group and item type. For unrelated items, utilization of noncriterial recollection was greater than 0 in the study practice, $t(23) = 17.83$, $p < .001$, $d = 3.64$, elaborative encoding group, $t(28) = 12.39$, $p < .001$, $d = 2.30$, and retrieval practice group, $t(27) = 21.88$, $p < .001$, $d = 4.13$. For weakly-related items, utilization of noncriterial recollection was greater than 0 in the study practice group, $t(24) = 23.37$, $p < .001$, $d = 4.67$, elaborative encoding group, $t(26) = 10.18$, $p < .001$, $d = 1.96$, and retrieval practice group, $t(26) = 21.81$, $p < .001$, $d = 4.20$.

Cue Diagnosticity

Diagnosticity of Target retrieval. To measure the degree to which target retrieval was diagnostic of final-test performance, I calculated γ correlations between target retrieval during the JOL phase (coded as 0 for unretrieved or 1 for retrieved) and final-test performance on an item-by-item basis for each participant. Due to invariance either in target retrieval or final-test performance, γ correlations for weakly-related and/or unrelated items could not be computed for 18 participants in the study practice group, 12 participants in the elaborative encoding group, and 16 participants in the retrieval practice group. Refer to Table 4 for the means.

I conducted a 3 (study technique: study practice, elaborative encoding, retrieval practice) x 2 (item type: unrelated, weakly-related) mixed-groups factorial ANOVA on the diagnosticity of target retrieval. There was no main effect of study-technique group, $F(2,50) = 0.63$, $p = .536$, $\eta_p^2 = .03$, but there was a main effect of item type, $F(1,50) = 732.15$, $p = .005$, $\eta_p^2 = .15$. The interaction between study-technique group and item type was not significant, $F(2,50) = 0.63$, $p = .54$, $\eta_p^2 = .03$. Target retrieval was more diagnostic for unrelated items (M = 1.00) than for weakly-related items (M = .81).

I also conducted one-sample *t*-tests to determine if the diagnosticity of target retrieval was greater than 0 for each group and item type. For unrelated items, the *t* statistic could not be calculated because the γ coefficient of every participant was 1.00 and thus there was no standard error. For weakly-related items, diagnosticity of target retrieval was greater than 0 in the study practice

group, $t(27) = 28.49$, $p < .001$, $d = 5.70$, elaborative encoding group, $t(27) = 6.02$, $p < .001$, $d = 1.14$, and retrieval practice group, $t(27) = 6.67$, $p < .001$, $d = 1.26$.

Diagnosticity of Noncriterial Recollection. To measure the degree to which noncriterial recollection was diagnostic of final-test performance, I calculated γ correlations between noncriterial recollection during the JOL phase (coded as 0 for no noncriterial recollection or 1 for noncriterial recollection) and final-test performance on an item-by-item basis for each participant. These correlations were restricted to cases in which the participants did not retrieve the target. Due to invariance either in noncriterial recollection or final-test performance, γ correlations for weakly-related and/or unrelated items could not be computed for 14 participants in the study practice group, 14 participants in the elaborative encoding group, and 17 participants in the retrieval practice group. Refer to Table 4 for the means.

I conducted a 3 (study technique: study practice, elaborative encoding, retrieval practice) x 2 (item type: unrelated, weakly-related) mixed-groups factorial ANOVA on the diagnosticity of noncriterial recollection. There was no main effect of study-technique group, $F(2,51) = 0.43$, $p = .650$, $\eta_p^2 = .02$, item type, $F(1,51) = 1.99$, $p = .163$, $\eta_p^2 = .04$, or interaction between study-technique group and item type, $F(2,51) = 0.62$, $p = .545$, $\eta_p^2 = .02$.

I also conducted one-sample *t*-tests to determine if the diagnosticity of target retrieval was greater than 0 for each study-technique group and item type. For unrelated items, diagnosticity of noncriterial recollection was greater than 0 in the study practice group, $t(20) = 3.26$, $p = .028$, $d = 0.71$, and retrieval practice

group, $t(21) = 2.37$, $p = .028$, $d = 0.50$, but not the elaborative encoding group ($p > .05$). For weakly-related items, diagnosticity of noncriterial recollection was not greater than 0 in any group (all p's > .05).

Mediation Analysis of JOL Accuracy

Having demonstrated that retrieval practice leads to higher JOL accuracy than study practice for weakly-related items, I conducted a mediation analysis to determine the cause of this effect. Theoretically, this effect could have owed to differences in cue utilization, cue diagnosticity, and/or cue accessibility. However, only cue accessibility significantly differed between retrieval practice and study practice. Thus, I tested the hypothesis that study techniques influenced JOL accuracy indirectly through affecting the accessibility of (a) target retrieval and/or (b) noncriterial recollection, which served as the two mediators in the subsequent analysis. I conducted the mediation analysis using ordinary least squares path analysis. Note that each relationship expressed in the model corresponds to an unstandardized coefficient from the ordinary least squares regression path analysis. Retrieval practice was dummy coded as 1, meaning that each coefficient pertaining to study techniques represent the difference between retrieval practice compared to study practice. Accessibility of target retrieval and noncriterial recollection were centered at the grand mean of the two groups. Refer to Figure 5 for a diagram of the mediation model I tested.

As shown in Table 5 and Figure 5, retrieval practice led to higher rates of target retrieval than study practice ($a_1 = 0.30$, $t(1) = 5.35$, $p < .001$), and higher levels of target retrieval were positively and significantly related to JOL accuracy

for weakly-related items, (b_1 = 0.57, $t(1) = 3.12$, $p = .003$). Retrieval practice did not lead to higher rates of noncriterial recollection than study practice, ($a_2 = .02$, $t(1) = 0.04$, $p = .7267$), and higher rates of noncriterial recollection were not significantly associated with JOL accuracy for weakly-related items ($b_2 = 0.01$, $t(1) = 0.00$, $p = .9951$). I tested the significance of the indirect effects of study technique on JOL accuracy for weakly-related items through target retrieval ($ab_1 = 0.17$) and noncriterial recollection ($ab_2 = 0.00$) by estimating 95% confidence intervals (CI) using bias-corrected bootstrap samples from 5,000 simulations. The 95% CI for the indirect effect of target retrieval was entirely above 0 [0.08, 0.25], but the 95% CI for the indirect effect of noncriterial recollection included 0 [-0.02, 0.03]. The direct effect of retrieval practice on JOL accuracy for related items was not significant ($c' = .096$, $t(1) = 1.06$, $p = .2929$). Therefore, the analysis suggests that the benefit of retrieval practice, over study practice, on JOL accuracy for weakly-related items was entirely mediated by its influence on accessibility of target retrieval during the JOL phase.

Experiment 3 Discussion

In Experiment 3, I partially replicated the effects of study technique on JOL accuracy for weakly-related items that was observed in Experiment 2. That is, retrieval practice, but not elaborative encoding, led to higher JOL accuracy than study practice. Regarding JOL accuracy for unrelated items, there were no differences across study-technique groups. This is consistent with the hypothesis that differences in JOL accuracy across study-technique groups would be greater for weakly-related than unrelated items.

Hypotheses regarding noncriterial recollection were partially verified. Participants did rely on noncriterial recollection to make JOLs in all groups and with both item types; that is, JOLs were higher when participants retrieved noncriterial cues than when they did not in all experimental conditions. However, noncriterial recollection was not diagnostic of final-test performance in most of the experimental conditions. Across all experimental conditions, the diagnosticity of noncriterial recollection was above 0 (i.e., chance levels) only in the study practice and retrieval practice groups, and only for unrelated items. In contrast, target retrieval was highly diagnostic of final-test performance in all experimental conditions.

Contrary to my initial hypothesis, differences in the accessibility of noncriterial recollection did not explain the advantage of retrieval practice, over study practice, on JOL accuracy for weakly-related items. Rather, the mediation analysis demonstrated that this effect was entirely due to the fact that retrieval practice led to higher rates of target retrieval during the JOL phase, which was positively and significantly associated with JOL accuracy for weakly-related items. These results also suggest that study techniques did not differentially influence JOL accuracy for unrelated items because they did not affect the accessibility of target retrieval for those items at the time of the JOL.

It is likely that noncriterial recollection was not positively associated with JOL accuracy because many of the retrieved noncriterial cues were not diagnostic of final-test performance. A cue is only diagnostic if it supports selection of the target amongst the foils, and there are many noncriterial cues that would not

support performance on this task. For example, during the practice phase in which participants explained their remember/know responses, some participants correctly reported that an unretrieved target was related to the cue. However, this cue may not support selection of that target because, for weakly-related items, all foils were related to the cue. Thus, recollecting the associative relatedness of the target would not help rule out foils. As another example, some participants reported remembering a mental image that a given cue evoked at the time of encoding. Given that this information pertained to the cue, and not the unretrieved target, it might not support the recognition of that target on the final test. It is entirely possible that some noncriterial cues did support performance on the four-alternative forced-choice test in this experiment. However, the measurement of noncriterial recollection with a remember/know task precludes an empirical evaluation of this possibility.

Generally, studies that have observed a positive association between noncriterial recollection and final-test performance measured cues that were (a) associated with the target, and (b) useful in supporting selection of the target amongst foils (Hertzog et al., 2010; Hertzog et al., 2014; Thomas et al., 2011, 2012; but see Isingrini et al., 2016). For example, Thomas et al (2011) measured participants' memory for the emotional valence (positive or negative) of unretrieved targets when making feeling-of-knowing judgments. The options on the final six-alternative forced-choice test featured both positively- and negatively-valenced words, and thus remembering valence would help performance by eliminating foils. Similarly, Hertzog et al (2014) measured

participants' memory of the mediator images that they generated during encoding of word pairs when making feeling-of-knowing judgments. Word pairs in that experiment either consisted of two abstract or two concrete nouns, and mediators for these pairs generally reflected the nature of the word pairs. Thus, remembering the mediator image likely cued participants into what type of target word they would need to select. As with Thomas et al (2011), options on the final four-alternative forced-choice test featured both abstract and concrete nouns, thus rendering recollection of the mediator useful in selecting the target by eliminating foils.

Participants were sensitive to the relative efficacy of study-techniques on final-test performance for weakly-related items, but not for unrelated items. That is, for weakly-related items, the JOL magnitude of participants in the retrieval practice and elaborative encoding groups was higher than study practice, which mirrored differences in performance on the final test. However, for unrelated items, JOL magnitude did not differ across groups, despite the fact that retrieval practice and elaborative encoding led to higher final-test performance than study practice for unrelated items. As with JOL accuracy, the lack of differences in JOL magnitude for unrelated items likely owed to the fact that there were no differences in accessibility of target retrieval and noncriterial recollection across groups for unrelated items.

Finally, the results suggest that correct guessing did not significantly influence JOL accuracy for either unrelated or weakly-related items. Factoring out the cases of correct guessing minimally affected the mean JOL accuracy in all

experimental conditions, and did not change the significance tests. As with Experiments 1 and 2, the JOL accuracy for both item types was lower than what is typically observed in experiments that use a cued-recall test as the criterion measure. Given that correct guessing did not appear to influence the data significantly, the lower levels of JOL accuracy likely reflect the difficulty in predicting recognition memory rather than noise due to the design.

CHAPTER 5: GENERAL DISCUSSION

The present study investigated the factors that influence delayed-JOL accuracy. I found that study techniques differentially influenced JOL accuracy when (a) the retention interval between encoding and predictions was long (48 hr; Experiments 2 and 3), but not short (15 min; Experiment 1), and (b) word pairs were weakly-related, but not unrelated (Experiment 3). In both Experiments 2 and 3, retrieval practice led to higher JOL accuracy for weakly-related word pairs than study practice. In contrast, elaborative encoding led to higher JOL accuracy for weakly-related word pairs than study practice only in Experiment 2, but did not differ between the other two groups in Experiment 3.

JOL Accuracy and Target Retrieval

The results of Experiment 3 suggest that study techniques influenced JOL accuracy by affecting the accessibility of target retrieval, but not noncriterial recollection, at the time of the judgment. In Experiment 3, the benefit of retrieval practice over study practice on JOL accuracy for weakly-related items was entirely mediated by differences in the accessibility of target retrieval. That is, participants in the retrieval practice group remembered more targets during the JOL phase than those in the study practice group, and the number of targets retrieved during the JOL phase was positively associated with JOL accuracy. These results suggest that study techniques did not influence JOL accuracy for unrelated items because study techniques did not affect levels of target retrieval of unrelated items.

Differences in the accessibility of target retrieval during the JOL phase can also explain why elaborative encoding led to higher JOL accuracy for weakly-related items than study practice in Experiment 2, but not Experiment 3. Recall that in Experiment 2, participants in all groups initially studied items for a fixed interval (1.5 s). In the study practice group, participants studied the word pairs a second time for a fixed interval (1.5 s; total of 3 s per item). In the elaborative encoding group, participants studied the word pairs in a self-paced fashion, leading to more time spent encoding (~6.5 s; ~8s total per item). The longer duration of study in the elaborative encoding group likely led to higher rates of covert target retrieval at the time of the JOL, and consequently, higher levels of JOL accuracy. In Experiment 3, encoding time was fixed at a total of 12 s per item in all groups (6 s for initial study, 6 s for second study or elaboration). This change likely benefited the quality of encoding in the study practice group more than the elaborative encoding group. Note that the duration of encoding increased by 9 s in the study practice group, but only 4 s in the elaborative encoding group. Further, the increase in encoding time for the elaborative encoding group was entirely allotted to initial study (+4.5 s), which was at the slight expense of the elaborative task (-0.5 s). This would explain why JOL accuracy for weakly-related items in the study practice group nearly doubled from .16 to .31 across Experiments 2 and 3, but remained constant at .41 for participants in the elaborative encoding group. Simply put, the gap of how many targets were accessible during the JOL phase likely decreased between the groups across experiments, and as a result, attenuated the gap in JOL accuracy.

As with study practice, JOL accuracy for weakly-related items in the retrieval practice group increased markedly between Experiments 2 and 3 (from .37 to .58). It is possible that the changes in encoding time between Experiments 2 and 3 accounts for this improvement. Participants in the retrieval practice group first studied word pairs for a fixed interval, and then took a cued-recall test. In Experiment 2, the initial study of word pairs was 1.5 s, which was increased to 6 s in Experiment 3. This appears to have increased the odds of successfully retrieving the target during retrieval practice. In Experiment 2, participants remembered 40% of weakly-related targets during retrieval practice, which increased to 58% in Experiment 3. Given that successful retrieval of a target potentiates future retrieval, it is likely that participants in Experiment 3 remembered more targets during the JOL phase than participants in Experiment 2. This would account for the increase in JOL accuracy for weakly-related items between Experiments 2 and 3.

Memory and Metamemory

The results of this study speak to the prevailing theories about why delay from encoding benefits JOL accuracy. The monitoring-dual-memories hypothesis, the stochastic-drift model, and the self-fulfilling prophecy account all share a common theme but differ in one crucial point. All three theories posit that delay is beneficial for JOL accuracy because it forces participants to base their JOLs on cues retrieved from long-term memory, and not cues that are transiently available in short-term memory. However, these theories differ regarding the *types* of cues from long-term memory that influence delayed-JOL accuracy. Whereas the self-

fulfilling prophecy account specifies target retrieval as the sole cue underlying the accuracy of delayed JOLs, both the monitoring-dual-memories hypothesis and stochastic drift model leave room for other types of cues, such as noncriterial recollection, to influence the accuracy of delayed JOLs.

Consistent with the self-fulling prophecy account, it was the cue of target retrieval, and not noncriterial recollection, that accounted for differences in delayed-JOL accuracy across groups in Experiment 3. However, the results of Experiment 3 cannot be used to rule decisively in favor of the self-fulfilling prophecy account. Participants did rely on noncriterial recollection to make their delayed JOLs; that is, they expressed higher JOLs when they retrieved noncriterial cues about the initial encoding event than when they retrieved nothing at all. This alone is enough to challenge the self-fulfilling prophecy account. Recall that JOL accuracy is a function of (a) the cues used to make the prediction, and (b) the degree to which those cues are diagnostic of final-test performance. The finding that participants use noncriterial cues to make their JOLs allows for the possibility that noncriterial cues *can* influence JOL accuracy; the effect simply depends on how diagnostic noncriterial cues are in a given context. Several experiments investigating feeling-of-knowing judgments have already offered cases in which noncriterial recollection was diagnostic of final-test performance, and also influenced predictive accuracy.

Limitations

Measuring noncriterial recollection with a remember/know procedure did not allow fine-grained analysis of the noncriterial recollection hypothesis. In the

current study, it was not possible to determine why noncriterial recollection was not diagnostic of final-test performance in most experimental conditions, nor mediated JOL accuracy between groups. As previously discussed, it is possible that participants remembered some noncriterial cues that were diagnostic of final-test performance, and others that were not. A "remember" response groups both cues into a single measure of noncriterial recollection. It is also possible that many of the noncriterial cues were not diagnostic because of the construction of the final test. Item relatedness was likely one of the most accessible noncriterial cues. However, recalling the associative relatedness of the target might not be diagnostic of final-test performance because all options on the final test were either unrelated or weakly-related. Measuring specific noncriterial cues would be necessary to explore the noncriterial recollection hypothesis more fully.

The findings may not generalize to all studies investigating JOL accuracy. As previously discussed, the overwhelming majority of JOL studies use cued-recall tests as the memory criterion measure. Consistent with previous research, the use of a recognition test as the criterion measure in the present study resulted in lower levels of JOL accuracy than what is typically observed in delayed-JOL paradigms. It is not uncommon to observe ceiling levels of delayed-JOL accuracy, even with simple study practice, in a typical cued-recall test paradigm. As such, it is likely that the influence of study techniques on delayed-JOL accuracy would not replicate in typical designs. To the author's best knowledge, Jang and colleagues (2012) conducted the only study to compare retrieval practice and study practice in a standard cued-recall, delayed-JOL paradigm. The authors

observed ceiling-levels of JOL accuracy in both groups. However, it is also possible that the effects of study techniques on JOL accuracy observed here would extend to cued-recall paradigms if care is taken to prevent ceiling levels of JOL accuracy. One method would be to interject a long delay between JOLs and the final test. Given that JOL accuracy reflects the degree to which information used to make JOLs matches information available at the final test, JOL accuracy should decrease as the delay between prediction and final test increases. Study techniques that lead to the best long-term memory of targets should also lead to the best JOL accuracy, regardless of the criterion measure, with sufficiently-long delays.

Conclusions

The results of these experiments demonstrate that study techniques differentially influence delayed-JOL accuracy by affecting the accessibility of targets at the time of the prediction. The present study suggests that participants use noncriterial cues when making JOLs. However, unlike previous research, these noncriterial cues were generally not predictive of final-test performance and access to noncriterial cues did not improve JOL accuracy. Further research is necessary to explore the types of cues that are predictive of final-test performance, as well as methods to increase access to these cues.

References

Arbuckle, T. Y., & Cuddy, L. L. (1969). Discrimination of item strength at time of presentation. Journal of Experimental Psychology, 81(1), 126-131.

Begg, I., Duft, S., Lalonde, P., Melnick, R., & Sanvito, J. (1989). Memory predictions are based on ease of processing. *Journal of Memory and Language, 28*(5), 610-632.

Brewer, G. A., Marsh, R. L., Clark-Foos, A., & Meeks, J. T. (2010). Noncriterial recollection influences metacognitive monitoring and control processes. *The Quarterly Journal of Experimental Psychology, 63*(10), 1936-1942.

Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: The benefits of elaborative retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*(6), 1563-1569.

Carpenter, S. K. (2011). Semantic information activated during retrieval contributes to later retention: Support for the mediator effectiveness hypothesis of the testing effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*(6), 1547-1552.

Cook, G. I., Marsh, R. L., & Hicks, J. L. (2006). Source memory in the absence of successful cued recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*(4), 828-835.

Daniels, K. A., Toth, J. P., & Hertzog, C. (2009). Aging and recollection in the accuracy of judgments of learning. *Psychology and Aging*, *24*(2), 494-500.

Dunlosky, J., Hertzog, C., & Powell-Moman, A. (2005). The contribution of mediator-based deficiencies to age differences in associative learning. *Developmental Psychology, 41*(2), 389-400.

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*(4), 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 and Language, 36*(1), 34-49.

Finn, B., & Metcalfe, J. (2007). The role of memory for past test in the underconfidence with practice effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*(1), 238-244.

Finn, B., & Metcalfe, J. (2007). Judgments of learning are influenced by memory for past test. *Journal of Memory and Language, 58*(1), 19-34.

Geraci, L., McCabe, D. P., & Guillory, J. J. (2009). On interpreting the relationship between remember–know judgments and confidence: The role of instructions. *Consciousness and Cognition: An International Journal*, 18(3), 701-709.

Hertzog, C., Dunlosky, J., & Sinclair, S. M. (2010). Episodic feeling-of-knowing resolution derives from the quality of original encoding. *Memory & Cognition, 38*(6), 771-784.

Hertzog, C., Fulton, E. K., Sinclair, S. M., & Dunlosky, J. (2014). Recalled aspects of original encoding strategies influence episodic feelings of knowing. *Memory & Cognition, 42*(1), 126-140.

Hertzog, C., Hines, J. C., & Touron, D. R. (2013). Judgments of learning are influenced by multiple cues in addition to memory for past test accuracy. *Archives of Scientific Psychology, 1*(1), 23-32.

Hicks, J. L., & Marsh, R. L. (2002). On predicting the future states of awareness for recognition of unrecallable items. *Memory & Cognition, 30*(1), 60-66.

Hines, J. C., Hertzog, C., & Touron, D. R. (2015). Younger and older adults weigh multiple cues in a similar manner to generate judgments of learning. *Aging, Neuropsychology, and Cognition, 22*(6), 693-711.

Isingrini, M., Sacher, M., Perrotin, A., Taconnat, L., Souchay, C., Stoehr, H., & Bouazzaoui, B. (2016). Episodic feeling-of-knowing relies on noncriterial recollection and familiarity: Evidence using an online remember-know procedure. *Consciousness and Cognition: An International Journal, 41*, 31-40.

Jang, Y., Wallsten, T. S., & Huber, D. E. (2012). A stochastic detection and retrieval model for the study of metacognition. *Psychological Review, 119*(1), 186-200.

Jeffreys, H. (1961). Theory of probability, (3rd ed.) Oxford, UK. Oxford University Press.

Jinkun, Z., & Lixian, Y. (2009). The process dissociation of the testing effect. *Psychological Science (China), 32*(5), 1180-1182.

Karpicke, J. D. (2009). Metacognitive control and strategy selection: Deciding to practice retrieval during learning. *Journal of Experimental Psychology: General, 138*(4), 469-486.

Karpicke, J. D., Blunt, J. R., & Smith, M. A. (2016). Retrieval-based learning: Positive effects of retrieval practice in elementary school children. *Frontiers in Psychology, 7*, 9.

Karpicke, J. D., & Smith, M. A. (2012). Separate mnemonic effects of retrieval practice and elaborative encoding. *Journal of Memory and Language, 67*(1), 17-29.

Karpicke, J. D., Lehman, M., & Aue, W. R. (2014). Retrieval-based learning: An episodic context account. In B. H. Ross (Ed.), *The psychology of learning and motivation (vol. 61)* (pp. 237-284, Chapter x, 330 Pages).

Kelemen, W. L. (2000). Metamemory cues and monitoring accuracy: Judging what you know and what you will know. *Journal of Educational Psychology, 92*(4), 800-810.

Kelley, C. M., & Sahakyan, L. (2003). Memory, monitoring, and control in the attainment of memory accuracy. *Journal of Memory and Language, 48*(4), 704-721.

Kimball, D. R., & Metcalfe, J. (2003). Delaying judgments of learning affects memory, not metamemory. *Memory & Cognition, 31*(6), 918-929.

King, J. F., Zechmeister, E. B., & Shaughnessy, J. J. (1980). Judgments of knowing: The influence of retrieval practice. *The American Journal of Psychology, 93*(2), 329-343.

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General, 126*(4), 349-370.

Koriat, A., & Ma'ayan, H. (2005). The effects of encoding fluency and retrieval fluency on judgments of learning. *Journal of Memory and Language, 52*(4), 478-492.

Koriat, A., Ma'ayan, H., Sheffer, L., & Bjork, R. A. (2006). Exploring a mnemonic debiasing account of the underconfidence-with-practice effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*(3), 595-608.

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*(2), 147-162.

Koriat, A., & Shitzer-Reichert, R. (2002). Metacognitive judgments and their accuracy. In P. Chambres, M. Izaute & P. Marescaux (Eds.), *Metacognition:*

Process, function and use. (pp. 1-17) Kluwer Academic Publishers, Dordrecht.

Lehman, M., Smith, M. A., & Karpicke, J. D. (2014). Toward an episodic context account of retrieval-based learning: Dissociating retrieval practice and elaboration. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*(6), 1787-1794.

Leonesio, R. J., & Nelson, T. O. (1990). Do different metamemory judgments tap the same underlying aspects of memory? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*(3), 464-470.

Lovelace, E. A. (1984). Metamemory: Monitoring future recallability during study.*Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*(4), 756-766.

Mazzoni, G., Cornoldi, C., & Marchitelli, G. (1990). Do memorability ratings affect study-time allocation? *Memory & Cognition, 18*(2), 196-204.

McCabe, D. P., & Geraci, L. D. (2009). The influence of instructions and terminology on the accuracy of remember-know judgments. *Consciousness and Cognition: An International Journal*, 18(2), 401-413.

McCabe, D. P., & Soderstrom, N. C. (2011). Recollection-based prospective metamemory judgments are more accurate than those based on confidence: Judgments of remembering and knowing (JORKS). *Journal of Experimental Psychology: General, 140*(4), 605-621.

Meeter, M., & Nelson, T. O. (2003). Multiple study trials and judgments of learning. *Acta Psychologica, 113*(2), 123-132.

Metcalfe, J. (2002). What does your brain believe? *Contemporary Psychology, 47*(2), 126-128.

Metcalfe, J. (2009). Metacognitive judgments and control of study. *Current Directions in Psychological Science*, 18(3), 159-163.

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*(4), 267-270.

Nelson, T. O., Gerler, D., & Narens, L. (1984). Accuracy of feeling-of-knowing experiments for predicting perceptual identification and relearning. *Journal of Experimental Psychology: General*, 113, 282-300.

Nelson, T. O., & Narens, L. (1994). Why investigate metacognition? In J. Metcalfe & A. P. Shimamura (Eds.), Metacognition: Knowing about knowing (pp. 1-25). Cambridge, MA, US: The 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*(1), 53-69.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). The University of South Florida Word Association, Rhyme, and Word Fragment Norms.

Peynircioğlu, Z. F., Brandler, B. J., Hohman, T. J., & Knutson, N. (2014). Metacognitive judgments in music performance. *Psychology of Music*, 42(5), 748-762.

Parks, C. M. (2007). The role of noncriterial recollection in estimating recollection and familiarity. *Journal of Memory and Language, 57*(1), 81-100.

Pyc, M. A., & Rawson, K. A. (2010). Why testing improves memory: Mediator effectiveness hypothesis. *Science, 330*(6002), 1-335.

Pyc, M. A., Rawson, K. A., & Aschenbrenner, A. J. (2014). Metacognitive monitoring during criterion learning: When and why are judgments accurate? *Memory & Cognition, 42*(6), 886-897.

Roediger, H.L. III, & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, 17(3), 249-255.

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, 56(5), 356-374.

Rhodes, M. G. (2016). Judgments of learning: Methods, data, and theory. In J. Dunlosky, & S. K. Tauber (Eds.), *The oxford handbook of metamemory.* (pp. 65-80) Oxford University Press, New York, NY.

Rhodes, M. G., & Tauber, S. K. (2011). The influence of delaying judgments of learning on metacognitive accuracy: A meta-analytic review. *Psychological Bulletin, 137*(1), 131-148.

Schneider, W., Eschman, A., & Zuccolotto, A. (2002). E-Prime: User's guide. Pittsburgh, PA: Psychology Software Tools.

Shaughnessy, J. J., & Zechmeister, E. B. (1992). Memory-monitoring accuracy as influenced by the distribution of retrieval practice. *Bulletin of the Psychonomic Society, 30*(2), 125-128.

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. 93-113, Chapter xiii, 334 Pages) The MIT Press, Cambridge, MA.

Serra, M. J., & Ariel, R. (2014). People use the memory for past-test heuristic as an explicit cue for judgments of learning. *Memory & Cognition*, *42*(8), 1260-1272.

Sikström, S., & Jönsson, F. (2005). A model for stochastic drift in memory strength to account for judgments of learning. *Psychological Review, 112*(4), 932-950.

Son, L. K., & Metcalfe, J. (2000). Metacognitive and control strategies in study-time allocation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*(1), 204-221.

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*(5), 315-316.

Soderstrom, N. C., Clark, C. T., Halamish, V., & Bjork, E. L. (2015). Judgments of learning as memory modifiers. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*(2), 553-558.

Son, L. K., & Metcalfe, J. (2005). Judgments of learning: Evidence for a two-stage process. *Memory & Cognition*, 33(6), 1116-1129.

Thiede, K. W. (1999). The importance of monitoring and self-regulation during multitrial learning. *Psychonomic Bulletin & Review, 6*(4), 662-667.

Thiede, K. W., Anderson, M. C. M., & Therriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. *Journal of Educational Psychology, 95*(1), 66-73.

Thiede, K. W., & Dunlosky, J. (1994). Delaying students' metacognitive monitoring improves their accuracy in predicting their recognition performance. *Journal of Educational Psychology, 86*(2), 290-302.

Thomas, A. K., Bulevich, J. B., & Dubois, S. J. (2011). Context affects feeling-of-knowing accuracy in younger and older adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*(1), 96-108.

Thomas, A. K., Bulevich, J. B., & Dubois, S. J. (2012). An analysis of the determinants of the feeling of knowing. *Consciousness and Cognition: An International Journal, 21*(4), 1681-1694.

Thomas, A. K., & McDaniel, M. A. (2007). Metacomprehension for educationally relevant materials: Dramatic effects of encoding--retrieval interactions. *Psychonomic Bulletin & Review, 14*(2), 212-218.

Tulving, E. (1985). Memory and consciousness. *Canadian Psychology/Psychologie Canadienne*, 26(1), 1-12.

Tullis, J. G., Finley, J. R., & Benjamin, A. S. (2013). Metacognition of the testing effect: Guiding learners to predict the benefits of retrieval. *Memory & Cognition*, 41(3), 429-442.

Undorf, M., Böhm, S., & Cüpper, L. (2016). Do judgments of learning predict automatic influences of memory? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 42*(6), 882-896.

Van Overschelde, J. P., & Nelson, T. O. (2006). Delayed judgments of learning cause both a decrease in absolute accuracy (calibration) and an increase in relative accuracy (resolution). *Memory & Cognition, 34*(7), 1527-1538.

Wagenmakers, E., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., . . . Morey, R. D. (2017). Bayesian inference for psychology. part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review*.

Weaver, Charles A., I.,II, & Kelemen, W. L. (2003). Processing similarity does not improve metamemory: Evidence against transfer-appropriate monitoring. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29(6), 1058-1065.

Westfall, P., Johnson, W., & Utts, J. (1997). A Bayesian Perspective on the Bonferroni Adjustment. *Biometrika*, 84(2), 419-427.

Yonelinas, A. P., Aly, M., Wang, W., & Koen, J. D. (2010). Recollection and familiarity: Examining controversial assumptions and new directions. *Hippocampus*, 20(11), 1178-1194.

Table 1

Experiment 1. Mean JOLs, Final-Test Performance, and JOL Accuracy (mean γ correlation coefficient).

Study Technique Group	JOL	Final Test	JOL Accuracy
Study Practice	4.72 (*1.47*)	.58 (*.17*)	.24 (*.35*)
Elaborative Encoding	6.06 (*1.47*)	.90 (.06)	.21 *(.56)*
Retrieval Practice	5.88 (*1.48*)	.89 (*.11*)	.12 *(.53)*

Note: Standard deviations given in parentheses.

Table 2

Experiment 2. Mean JOLs, Final-Test Performance, and JOL Accuracy (mean γ correlation coefficient).

Study Technique Group	JOL (All)	Final Test	JOL Accuracy
Study Practice	3.71 (*1.48*)	.47 (*.12*)	.16 (*.28*)
Elaborative Encoding	4.98 (*1.20*)	.70 (*.12*)	.41 (*.31*)
Retrieval Practice	5.08 (*1.51*)	.79 (*.15*)	.37 (*.26*)

Note: Standard deviations given in parentheses.

Table 3

Experiment 3. Mean accessibility of target (proportion of targets retrieved during JOL phase), noncriterial recollection (proportion of remember responses for unretrieved items during the JOL phase), JOL magnitude, JOL accuracy (mean γ correlation between JOLs and final-test performance), and confidence accuracy (mean γ correlation between confidence responses and final-test performance), as a function of item type (unrelated and weakly-related word pairs).

	Associative Relatedness											
	Unrelated						Weakly-Related					
Study Technique	Target Retriev.	Noncrit. Recoll.	Mean JOL	Final Test	JOL Acc.	Conf. Acc.	Target Retriev.	Noncrit Recoll.	Mean JOL	Final Test	JOL Acc.	Conf. Acc.
Study Practice	.17 (.24)	.30 (.27)	46.30 (22.80)	.69 (.23)	.31 (.35)	.67 (.36)	.28 (.26)	.29 (.22)	50.45 (23.47)	.70 (.15)	.31 (.30)	.56 (.33)
Elaborative Encoding	.26 (.22)	.34 (.22)	53.57 (19.10)	.86 (.17)	.35 (.41)	.81 (.24)	.45 (.19)	.51 (.30)	65.72 (16.10)	.83 (.13)	.41 (.32)	.72 (.40)
Retrieval Practice	.24 (.22)	.24 (.18)	49.31 (19.24)	.83 (.16)	.28 (.41)	.81 (.20)	.56 (.25)	.30 (.24)	66.01 (17.50)	.86 (.15)	.58 (.27)	.70 (.51)

Note: Standard deviations given in parentheses.

Table 4

Experiment 3. Mean cue utilization of target retrieval and noncriterial recollection (mean γ correlation of each cue and final-test performance on an item-by-item basis), as a function of associative relatedness of word pairs, for each study-technique group.

	Associative Relatedness							
	Unrelated				Weakly-Related			
	Utilization		Diagnosticity		Utilization		Diagnosticity	
Study Technique	Target Retriev.	Noncrit. Recoll.	Target Retriev.	Noncrit Recoll.	Target Retriev.	Noncrit Recoll.	Target Retriev.	Noncrit. Recoll.
Study Practice	.99 (.02)	.87 (.24)	1.00 (0.00)	.42 (.59)	.81 (.23)	.87 (.19)	.94 (.17)	.17 (.69)
Elaborative Encoding	.95 (.09)	.83 (.36)	1.00 (0.00)	.08 (.85)	.77 (.19)	.82 (.42)	.68 (.60)	.12 (.79)
Retrieval Practice	.99 (.04)	.85 (.21)	1.00 (0.00)	.36 (.70)	.92 (.13)	.83 (.20)	.73 (.58)	.06 (.88)

Note: Standard deviations given in parentheses.

Table 5

Coefficients, standard error, and p values from the Experiment-3 mediation analysis of the effect of retrieval practice, compared to study practice, on JOL accuracy for weakly-related items.

		Consequent										
		Mediator 1: Target retrieval				Mediator 2: Noncriterial Recollection				JOL Accuracy (Weakly-Related)		
Variable		Coeff	SE	*p*		Coeff	SE	*p*		Coeff	SE	*p*
Intercept	i_{01}	-.24	.04	<.001	i_{02}	-.02	.04	<.001	i_{03}	.45	.07	<.001
Retrieval Practice	a_1	.30	.05	<.001	a_2	.02	.06	.727	**c'**	.10	.09	.292
Target retrieval	-	-	-	-	-	-	-	-	b_1	.57	.18	.003
Noncriterial Recollection	-	-	-	-	-	-	-	-	b_2	.01	.17	.965
		$R^2 = .36$ $F(1,52) = (28.65)$, $p < .001$				$R^2 = .002$ $F(1,52) = (.12)$, $p = .727$				$R^2 = .32$ $F(1,52) = (7.79)$, $p < .001$		

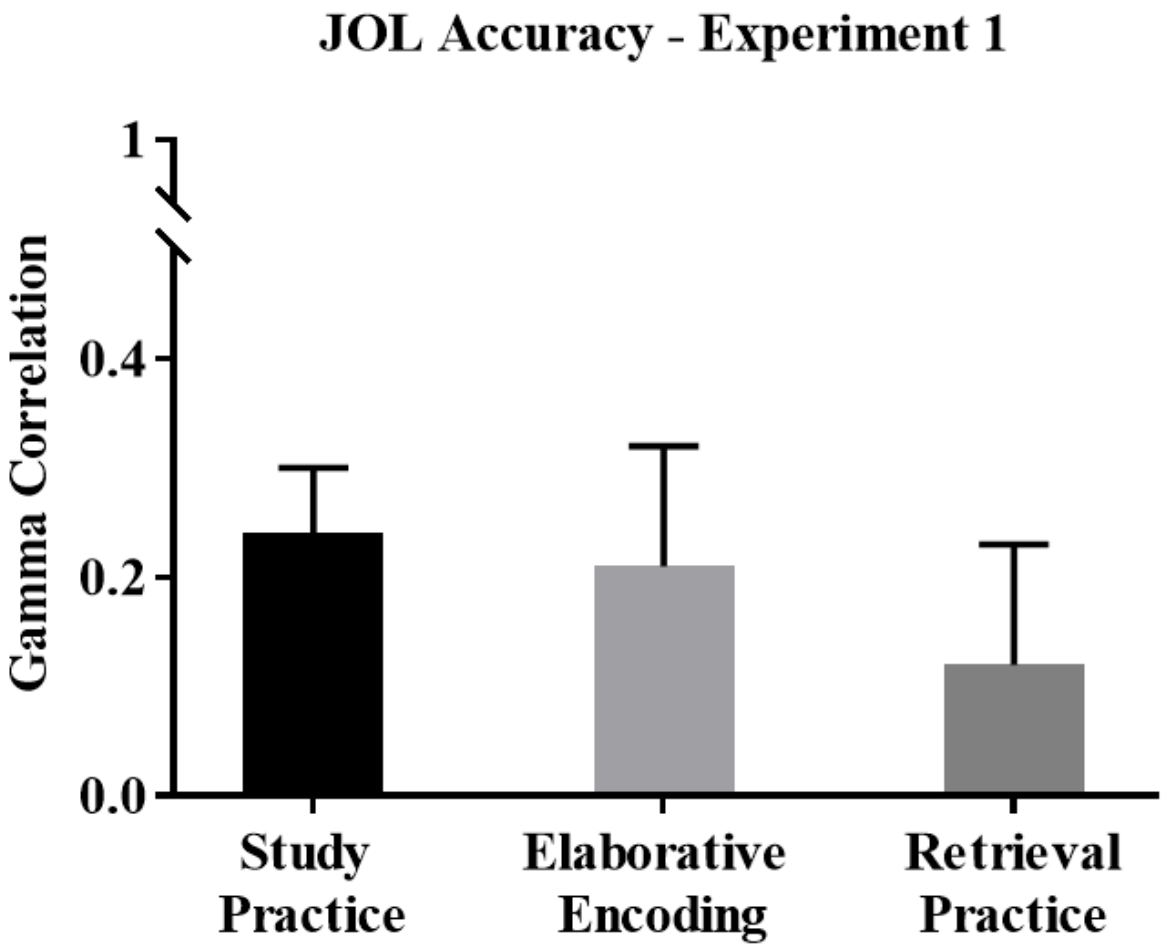

Figure 1. Delayed judgment-of-learning accuracy of participants in Experiment 1 as a function of study-technique group. Error bars represent standard error of the mean. *p < .05, **p < .01, ***p < .001.

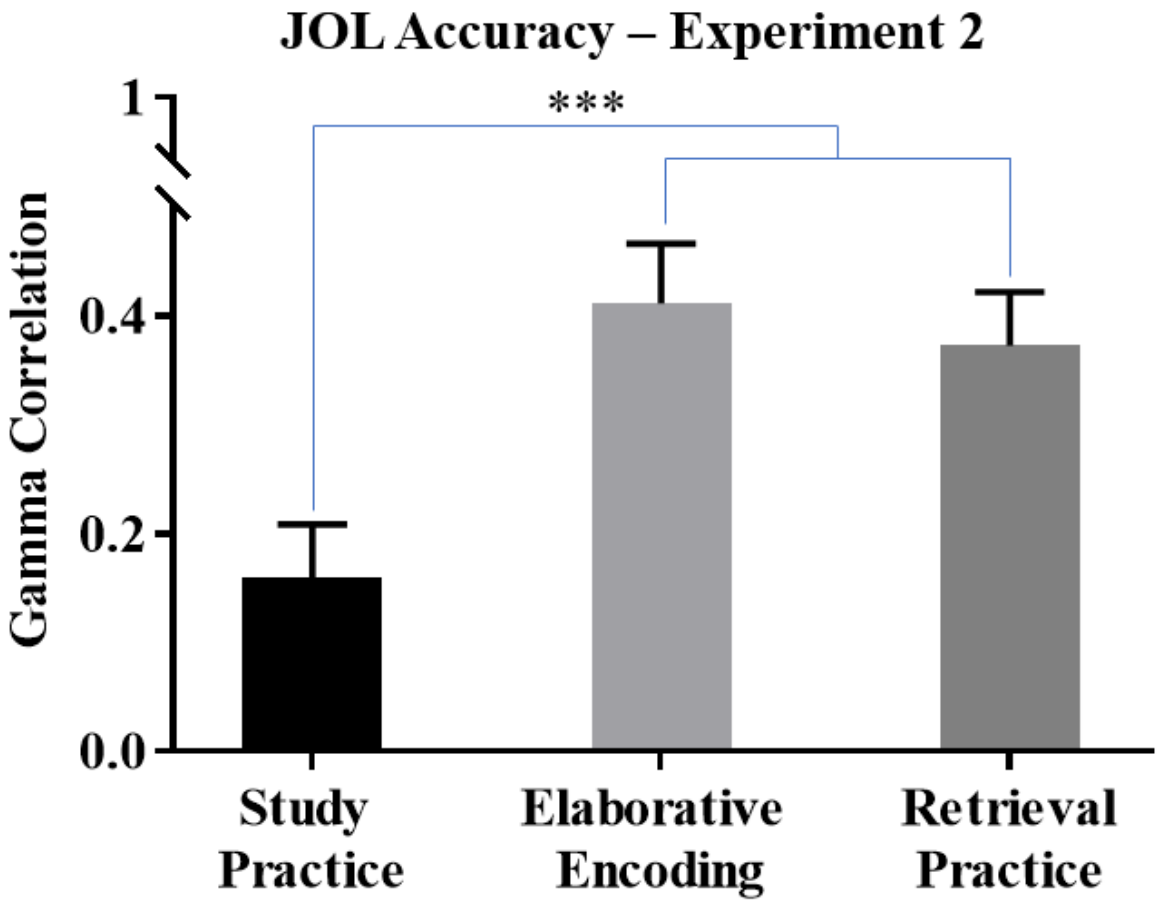

Figure 2. Delayed judgment-of-learning accuracy of participants in Experiment 2 as a function of study-technique group. Error bars represent standard error of the mean. *p < .05, **p < .01, ***p < .001.

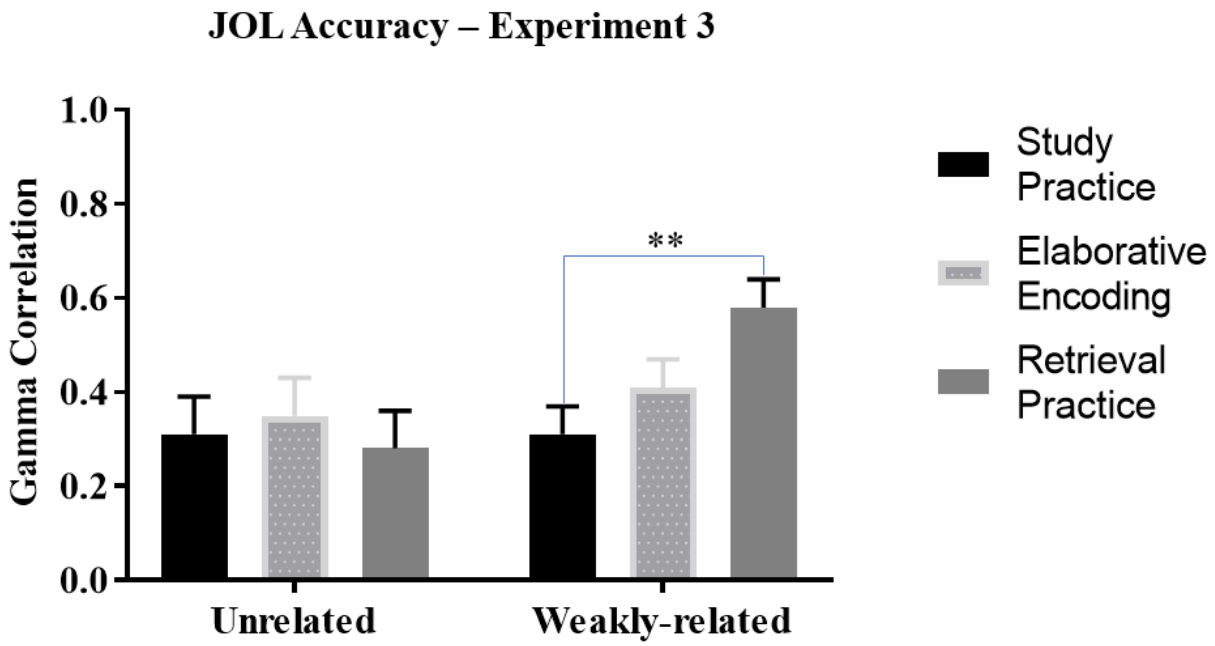

Figure 3. Delayed judgment-of-learning accuracy of participants in Experiment 3, as a function of study-technique group and associative relatedness of the word pairs. Error bars represent standard error of the mean. *$p < .05$, **$p < .01$, ***$p < .001$.

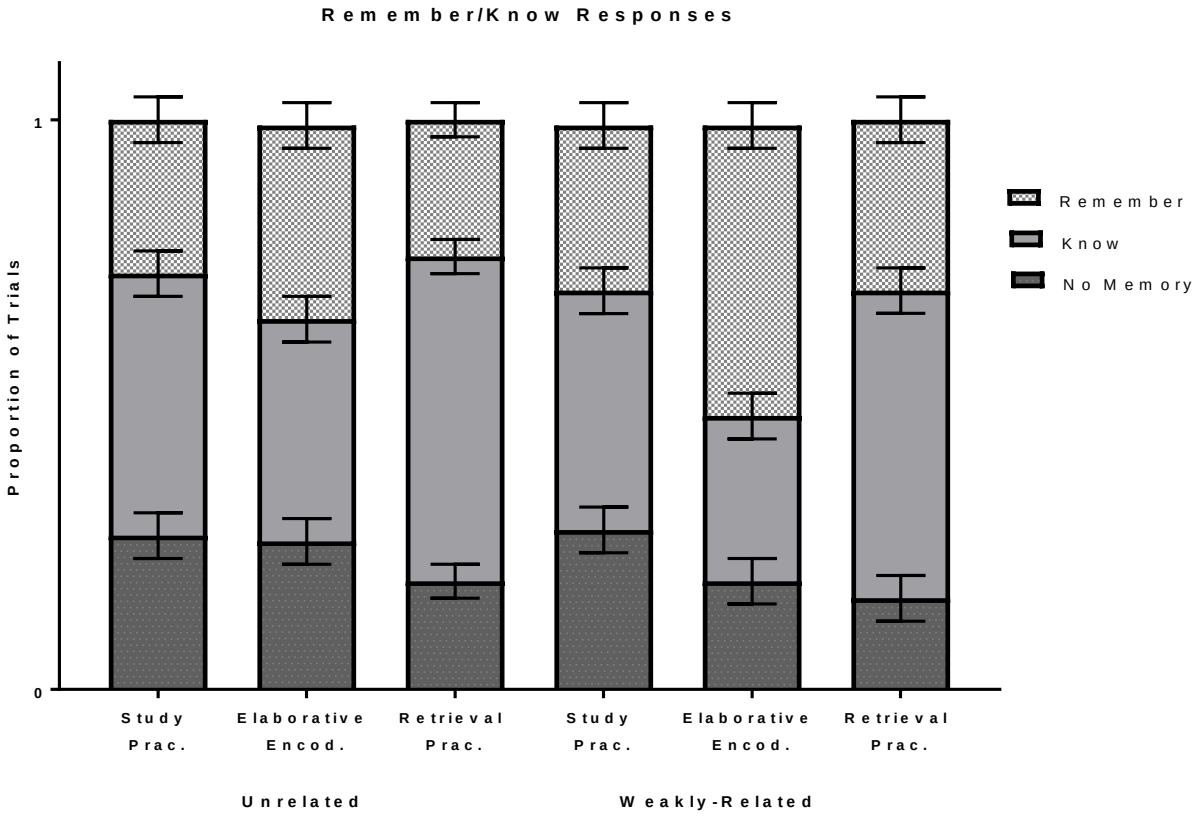

Figure 4. Proportion of remember, know, and no memory responses during the JOL phase in Experiment 3, split by type of word pair. Proportions calculated for trials in which the participant did not retrieve the target. Error bars represent standard error of the mean.

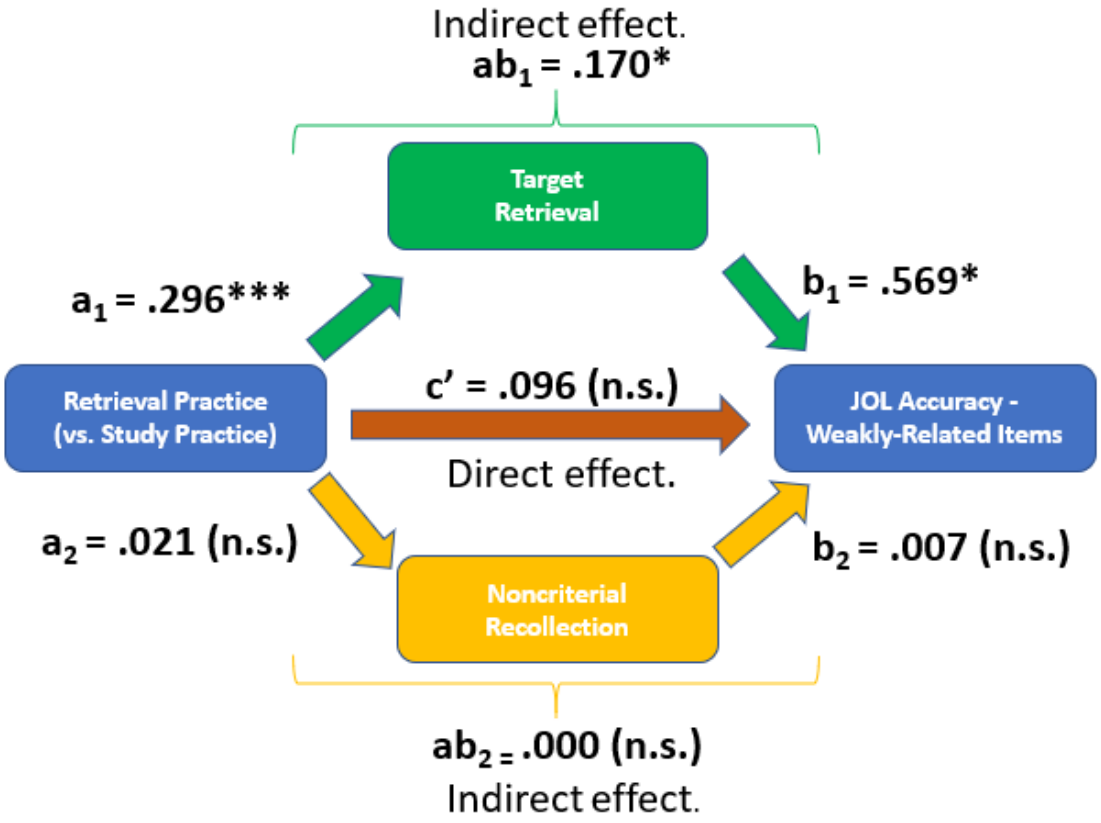

Figure 5. Experiment-3 mediation model of the benefit of retrieval practice, over study practice, on JOL accuracy for weakly-related items. The two mediators were (a) accessibility of target retrieval and (b) accessibility of noncriterial recollection. Values represent unstandardized coefficients from parallel mediation path analysis. Coefficients a_1 and a_2 represent the difference between retrieval practice and study practice on the mediators. Coefficient c' represents the effect of retrieval practice on JOL accuracy independent of its influence on the mediators (indirect effects ab_1 and ab_2). The beneficial effect of retrieval practice on JOL accuracy for related items was entirely mediated by the indirect effect of target retrieval (indirect effect ab_1). *p < .05, **p < .01, ***p < .001.

Appendix A

Cues, targets, and foils used in Experiments 1 and 2.

Cue	Target	Foil 1	Foil 2	Foil 3
MOTHER	CHILD	LOVE	DAUGHTER	MOM
PRESCRIPTION	DOCTOR	MEDICINE	PHARMACY	PILL
SOIL	EARTH	GROUND	PLANT	FARM
DUSK	EVENING	SUNSET	DARK	NIGHT
DONOR	HEART	VOLUNTEER	GIVE	ORGAN
WEAPON	KNIFE	KILL	DANGER	WAR
SONNET	MUSIC	SHAKESPEARE	SONG	ENGLISH
EMPLOYMENT	OFFICE	AGENCY	MONEY	WORK
TRASH	PAPER	CAN	BAG	SMELL
SEA	RIVER	FISH	WATER	WAVES
VOCABULARY	SCHOOL	ENGLISH	SPELLING	DICTIONARY
JACKET	SHIRT	COLD	SWEATER	TIE
PEDESTRIAN	STREET	PERSON	CROSSING	SIDEWALK
BREEZE	SUMMER	BLOW	COOL	AIR
COFFEE	TABLE	CAFFEINE	BLACK	CREAM
FRAME	WINDOW	GLASSES	ACCUSE	BICYCLE
STORE	MARKET	BUY	GROCERIES	PLACE
GOBLIN	CREATURE	HALLOWEEN	GHOUL	FRONT
DIAMOND	PEARL	GOLD	EMERALD	PAST
DISLIKE	MEAN	DISHONEST	ENEMY	CHAMPION
EMPTY	NEST	GONE	HEAD	SUFFER
END	FINISH	START	STOP	SHINE
REMEMBER	MIND	MEMORY	RECALL	MIND
VICTORY	WAR	DEFEAT	LOSE	WAR
PAIN	HEADACHE	AGONY	OUCH	HEADACHE
LIGHT	SHINE	SUN	DAY	BRIGHT

Appendix B

Instructions for Study Practice

You are about to see a series of two words presented together (like “Moon – Galaxy”). Each pair of words will be presented one at a time. Your job is to try and remember these word pairs for a later test of your memory.

Instructions for Elaborative Encoding

You are about to see same pairs of words again. For each pair of words, I want you to think of a new word which links the two together. You have as much time as you need for entering in your answer.

For example, if you saw "Moon - Galaxy," you might type in the word "Space."

Instructions for Retrieval Practice

You will now be presented with some of the words from the pairs you saw earlier. You will see one word at a time, and your goal is to try and remember the word that goes with it.

For example, if you saw "Moon - Galaxy," we will show you the word “Moon” and want you to type in the word “Galaxy.” You have as much time as you need for entering in your answer. After you enter a word, you will see the answer.

If you do not remember, please guess

Instructions for Judgments of Learning

We will show you some of the words you studied earlier, and want you to tell us how likely you will be to recognize the word that goes with it on a test about 5 minutes from now. On the test, we will give you the first word and ask you to choose the word the goes with it on a list of four choices.

For example, if you studied “Moon – Galaxy,” we will show you “Moon,” and want you to tell us how likely you will be to recognize “Moon” on the later test.

Use the following scale to make this prediction

0 1 2 3 4 5 6 7 8 9 10

0 = will not remember 10 = will definitely remember

If you think you will not be able to remember the word later, give it a low rating. If you think maybe you can, give it higher rating, like 4 or 5. If you think you definitely can remember then give it a high rating.

JOL Screen

Cue - _____

How likely do you think you'll be able to **recognize** the second word on a list of choices?

0 1 2 3 4 5 6 7 8 9 10

0 = will not remember 10 = will definitely remember

Appendix C

Cues, targets, and foils used in Experiment 3.

Item Type	Cue	Target	Foil 1	Foil 2	Foil 3
Unrelated	ALLERGY	DIVORCE	BIRDS	BRAIN	DAMAGE
Unrelated	BACON	MAYOR	LINT	ELEVATOR	FRECKLE
Unrelated	BALLOON	ACRE	SPOUSE	LIBRARY	TAPE
Unrelated	BELLY	STOVE	OATS	CELLAR	EXAM
Unrelated	BEVERAGE	PHYSICS	ARTERY	TEAM	LIQUID
Unrelated	BOOZE	ICING	EMPLOYER	READER	MEETING
Unrelated	BREEZE	TOASTER	PUPPY	FEVER	POND
Unrelated	BUTLER	SEASHORE	PLAQUE	CUSTARD	DUNE
Unrelated	COTTON	SAILING	SANDALS	TEXT	PECAN
Unrelated	CRAYON	HALO	CHORE	PEDAL	CARROTS
Unrelated	DYNASTY	GRIP	SCENE	RAISIN	SLEEVE
Unrelated	EXIT	BROOM	MULE	PILL	ANTIDOTE
Unrelated	GHOST	PORCH	SHOE	SKELETON	BEAVER
Unrelated	LAMB	PARADE	LOBSTER	COMPASS	PAINT
Unrelated	LUNG	SHADOW	SHIELD	CABOOSE	MUMMY
Unrelated	NICOTINE	PACKAGE	TURTLE	SOCCER	SPATULA
Unrelated	PLUM	HELMET	EAGLE	ORCHID	SIGNAL
Unrelated	POSSUM	GLACIER	FLOOD	CARBON	FLAVOR
Unrelated	THIEF	SNOW	ACROBAT	PARROT	LAWN
Unrelated	TROUSERS	CHEF	TRAITOR	PADDLE	PRIZE
Weakly-Related	BREAD	JELLY	ROLL	SANDWICH	WHEAT
Weakly-Related	CAVERN	MOUNTAIN	CABIN	HOLE	TUNNEL
Weakly-Related	CHAPEL	TEMPLE	STEEPLE	PRIEST	CROSS
Weakly-Related	COCOON	WORM	MOTH	NEST	SHELL
Weakly-Related	COFFIN	TOMB	BURIAL	GRAVE	VAMPIRE
Weakly-Related	DAGGER	BLADE	SWORD	BLOOD	MURDER
Weakly-Related	DENTIST	CAVITY	OFFICE	DOCTOR	DRILL
Weakly-Related	DIAMOND	EMERALD	GOLD	PEARL	RUBY
Weakly-Related	DOCK	SHIP	PIER	LAKE	PORT
Weakly-Related	DRESSER	CLOSET	CLOTHES	DESK	TABLE
Weakly-Related	GLOBE	CIRCLE	SPHERE	EARTH	ATLAS
Weakly-Related	HARP	SONG	VIOLIN	PIANO	FLUTE
Weakly-Related	INFERNO	FLAME	HEAT	VOLCANO	BLAZE
Weakly-Related	MARSH	WEED	JUNGLE	LAND	GRASS
Weakly-Related	MUSTACHE	RAZOR	FACE	MOUTH	HAIR
Weakly-Related	OREGANO	HERB	PIZZA	GARLIC	SAUCE
Weakly-Related	REPTILE	FROG	MAMMAL	SCALES	LIZARD
Weakly-Related	SUNRISE	MOON	MORNING	DAWN	BEACH
Weakly-Related	THRONE	CASTLE	SEAT	QUEEN	CROWN
Weakly-Related	TOOL	MACHINE	WRENCH	KITCHEN	SHOVEL

Appendix D

Remember/Know instructions from Experiment 3. Rather than use remember/know in the instructions, I used “Type 1” and “Type 2” memory, respectively.

Instructions:

We will ask you about your memory for studying the word pair. Even if you could not remember the second word of the pair, we want to know whether you have a memory for studying that word. For example, if you had studied the word pair “Planet – Galaxy,” but could not remember “Galaxy,” you still might have a memory of studying the word pair. We will ask you to tell us what your memory for studying the word pair is like. We will ask you to indicate whether you have a Type 1 memory, Type 2 memory, or no memory.

Type 1 Memory: This response indicates that you have a specific, conscious memory of studying the word pair. This could include such things as recollecting what you were thinking about when the word was presented, what the word looked like, any image or thought that the pair evoked, or what it sounded like. You should reply with Type 1 memory only if you can, if asked, tell the experimenter what you recollected about the study event.

Type 2 Memory: This response indicates that you recognize the word pair, and feel that you did study it, but cannot recollect anything specific about when you studied that word pair.

No memory: This response indicates that you do not remember anything specific about studying the word pair, nor recognize it or feel that you studied it at all.

Please note, you can have a Type 1 memory even if you did not remember the second word. As a real-life example, you might forget the name of an acquaintance, but have a specific memory for the moment you met them. An example of a Type 2 memory would be when you recognize someone, but can't remember anything specific about them.