DISSERTATION

THE INFLUENCE OF TESTING ON MEMORY, MONITORING, AND CONTROL

Submitted by

Megan K. Littrell

Department of Psychology

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Summer 2011

Doctoral Committee:

Advisor: Edward DeLosh

Anne Cleary
Michael De Miranda
Kurt Kraiger

ABSTRACT

THE INFLUENCE OF TESTING ON MEMORY, MONITORING, AND CONTROL

The current set of experiments investigated the role of testing in enhancing subsequent memory performance, a phenomenon known as the *testing effect*. The current study also assessed whether testing improves assessments of learning and influences subsequent study behaviors that serve to further enhance learning. In Experiments 1 and 2 participants studied lists of words in an initial phase and then either restudied or took a memory test on the words in an intervening phase. They were also asked to predict the likelihood that they would recall each item on a later memory test and indicate whether or not they would like another chance to restudy the item before the final memory test. The difference between the two experiments was that in Experiment 1 participants were allowed to restudy the items they chose, whereas they were not allowed to restudy those items in Experiment 2. Results for both experiments showed that initial testing compared to restudying enhanced final memory accuracy, and produced stronger correlations between predictions of recall and actual recall and between predictions and restudy choices. Experiment 3 examined the effects of testing on predictions of memory and the allocation of study time given to each item. Additionally, some participants' study time choices were honored, while other participants' choices were not. This manipulation was included to examine the differential effects of having control over what materials are restudied, depending on whether they have been simply restudied or subjected to prior test. Overall, the data suggest that testing enhances memory performance as well as relative metacognitive judgments.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

The Influence of Testing on Memory, Monitoring and Control

Researchers have long been interested in the benefits of testing on memory performance (see McDaniel, Roediger, & McDermott, 2007; Roediger & Karpicke, 2006a). Of particular interest is a phenomenon called the *testing effect*, a finding in which materials that are subjected to an initial test are more likely to be remembered on a later test compared to items that have not been tested or have been subjected to additional study (McDaniel et al., 2007; Roediger & Karpicke, 2006a). Many researchers have attempted to identify the boundary conditions of the effect by using various types of materials and experimental procedures. They have found testing effects using materials typically used in lab settings, such as lists of single words (Carpenter & DeLosh, 2006; Kuo & Hirshman, 1996; McDaniel & Masson, 1985; Thompson, Wenger, Bartling, 1978; Tulving, 1967) and paired associates (Cull, 2000; Kuo & Hirshman, 1996). Similar effects have also been observed in laboratory settings using materials that are more applicable to those found in the classroom such as text passages (Agarwal, Karpicke, Kang, Roediger, & McDermott, 2007; Butler, Karpicke, & Roediger, 2007; Glover, 1989; Roediger & Karpicke, 2006b), research articles (Kang, McDermott, & Roediger, 2007), general knowledge questions (Butler, Karpicke, & Roediger, 2008; McDaniel & Fisher, 1991), recorded lectures (Butler & Roediger, 2007), English vocabulary (Cull, 2000; Karpicke & Roediger, 2007), and foreign language words paired with their English

equivalent (Carpenter, Pashler, & Vul, 2006; Carrier & Pashler, 1992; Karpicke & Roediger, 2008). Some studies have even demonstrated testing effects in simulated or actual classroom settings (e.g., Bangert-Drowns, Kulik, & Kulik, 1991; Butler & Roediger, 2007; Carpenter, Pashler, & Cepeda, 2008; Gates, 1917; Spitzer, 1939). This volume of research suggests that the testing effect is robust and typically occurs with many different types of materials and experimental designs, as well as in both experimental and applied settings.

The standard testing effect procedure involves an initial study phase, in which items (e.g., lists of words, text passage, etc.) are encoded. Then during an intervening phase, some items are subjected to a memory test (e.g., cued recall, or free recall, or recognition) while others are either presented for an additional "restudy" opportunity or are not presented again (i.e., a "no test" condition). Testing effects are revealed as a memory advantage for tested items relative to either a no-test or restudy control condition (e.g., Carpenter & DeLosh, 2006; Carrier & Pashler, 1992; Glover, 1989). Although much research has investigated the conditions under which testing benefits memory, very few have examined the effects of testing on an individuals' predictions of their later memory performance, an aspect of metacognition. This is the primary purpose of the present study.

Metacognition and Testing

Flavell (1976) defined metacogniton as "one's knowledge concerning one's own cognitive processes or anything related to them." The cognitive processes involved in learning and engaging in metacognition about that learning has been described as involving two levels of awareness, the *object level* and the *meta-level* (Nelson & Narens,

1990). Initial information processing necessary for learning occurs at the object level. For example, if a student is studying their notes to prepare for a test, the cognitive processes involved in learning the content of the notes would occur at the object level. At the meta-level, people are able to observe and make judgments about the processing that occurred at the object level. In other words, they are able to think about their own cognitive processes through a process called *monitoring*. The monitoring that people engage in can then be used to directly affect the processing going on at the object level. For instance, at the meta-level the student studying for a test may observe that his/her understanding of the materials is not very strong. As a result of this monitoring of the students' own cognition, the student may decide to actively attempt to change the processing going on at the object level. This process is referred to as *control* and may be measured by assessing what strategies are used to change the processing of information at the object level. For example, participants may vary the amount of time that they allocate to studying items based on their assessment of how well they know those items and what their learning goals are (e.g., Metcalfe & Kornell, 2005; Theide & Dunlosky, 1999). In regards to the testing effect, it is possible that taking an initial test may have positive effects on monitoring and control processes that directly affect later memory performance. Therefore, one way to further our understanding of the benefits of testing on memory would be to study the relationship testing has with such metacognitive processes.

Although there are several ways of assessing metacognitive monitoring (e.g., ease-of-learning, Leonesio & Nelson, 1990; feelings of knowing, Hart, 1965; Nelson & Narens, 1990; tip-of-the-tongue, Brown & McNeil, 1966; judgments of remembering and

knowing, McCabe & Soderstrom, in press; remember/know judgments, Gardiner & Richarson-Klavehn, 2000), one common method is the *Judgment of Learning* (JOL). A JOL is a prediction of later memory performance, typically in the form of a percentage or probability judgment of the likelihood that information will be remembered (e.g., 0-100%) or a judgment of the proportion of items that will be remembered (e.g., 23/40). Researchers more finely estimate the correspondence between metacognitive judgments and performance on criterion tasks with observations of *calibration* and *resolution* (Koriat, 2007). Calibration represents the match between participants' overall performance on a task and their predictions of overall performance. For example, if a participant is asked to predict what percentage of items they will recall from a list of words on a later memory test, calibration would be measured by comparing this prediction with the actual percentage of items recalled. Resolution, on the other hand, represents a measure of an individual's ability to discriminate between information that is known relative to information that is not known. For example, participants may be asked to predict the likelihood of recalling each individual item on a list. These individual item predictions would then be compared to the items actually recalled on a later test.

Very few studies in the testing effect literature have examined the effects of testing on monitoring. In one testing effect study, Roediger and Karpicke (2006b) compared recall of passages of text in a condition involving four consecutive study trials (SSSS condition) with a condition in which there was one study trial followed by three test trials (STTT condition). After these trials, participants were asked to predict their performance on a memory test to take place one week later. Roediger and Karpicke discovered that even though participants recalled more information if they were tested

repeatedly than if they studied repeatedly, they reported greater confidence in later memory if they were in the repeated study condition. This mismatch between predictions and actual memory performance suggested that participants were not aware of the memory benefits of testing over those conferred by studying.

Similarly, Agarwal et al. (2007) asked participants to assess the percentage of information (on a scale of 0-100%) from text passages that they were likely to remember after either taking an open or closed-book test or engaging in additional study. If participants repeatedly studied the passages, their predictions were higher than if they were tested on the passages. However, consistent with Roediger and Karpicke's (2006b) findings, prior testing actually produced greater recall than simply restudying the passages.

Although this work suggests that participants are generally unaware of the benefits of testing, the results are limited in that they only pertain to participants' predictions when they are making overall, aggregate judgments of the likelihood of recalling information (i.e., calibration). This work does not indicate whether participants are better or worse at making predictions about individual items learned during the initial and intervening phases, depending on the type of intervening task. The latter is an issue of resolution, that is, how well participants' predictions discriminate between information that is learned and information that is less well-learned or not learned at all. Ideally, participants' predictions should differentiate between items such that they show high levels of confidence in well-learned items and relatively less confidence in items that either have not yet been learned or have not been learned adequately.

Although participants seem to be inaccurate with regard to their aggregate judgments about whether restudying or testing leads to better memory performance, it is possible that individual item judgments (i.e., resolution) produce a different result. Mazzoni and Nelson (1995) found that there are general differences in judgments made on an item-by-item basis compared to aggregate judgments. Specifically, they found that mean item-by-item judgments yield higher predictions than mean aggregate judgments. In addition, there is reason to suspect different results for item-by-item judgments than for aggregate judgments when comparing tested versus restudied items. Testing may help participants distinguish between what has or has not been learned. For instance, making metacognitive judgments about individual items may lead to greater accuracy in making relative predictions about tested items compared to restudied items because the act of retrieval gives participants greater insight into how well they have learned each item relative to restudying. Moreover, if testing improves resolution, might participants' control processes then differ for restudied versus tested items? The current study seeks to examine these questions by examining participants' predictions of performance on individual restudied and tested items as well as the control processes they engage in after making such predictions. Specifically, the study examines participants' decisions to engage in additional study for restudied versus tested items and how these control processes affect subsequent memory performance.

The Delayed JOL Effect and The Testing Effect

The literature on delayed JOLs may be of particular relevance in the discussion of metacognition and testing effects. In metacognitive studies, JOLs are either made immediately after the item is presented or after a delay (e.g., Nelson & Dunlosky,

1991). Nelson and Dunlosky (1991) showed that the accuracy with which JOLs predict actual memory performance is greater for delayed JOLs than for immediate JOLs. This is referred to as the *delayed JOL effect*. As an explanation for this phenomenon, Nelson and Dunlosky suggested that when JOLs are made immediately after studying the item, participants access the item in primary or short-term memory. After a delay, these items are no longer in short-term memory. Therefore, they must access the item in long-term or secondary memory. By doing so, they engage in an act similar to what is required at the final memory test, and their judgments therefore tend to more accurately reflect final test performance. In a similar vein, other researchers have suggested a *transfer-appropriate processing* account of the delayed JOL effect (e.g., Begg, Duft, Lalonde, Melnick, & Sanvito, 1989; Dunlosky & Nelson, 1992). This explanation suggests that delayed JOLs are more accurate than immediate JOLs because the processes that participants engage in for the delayed JOLs are more similar to the delayed memory test. When extended to the testing effect, these explanations suggest that testing may improve metacognitive accuracy by providing the rememberer with better insight into what they are and are not likely to recall on later tests, by better approximating what it is that the rememberer must do on those later tests.

Given this, it is worth considering the similarities and differences in the procedures used in the testing and delayed JOL literatures. As noted previously, the standard testing effect procedure involves an initial study phase in which a list of items is studied. This is followed by an intervening phase in which some items are tested while others are either not tested or are restudied. Finally, a criterion memory test is given over all items from each study list. In the current study, this procedure has been

modified by the addition of a JOL and a restudy decision after each item is presented during the intervening phase. The typical procedure for delayed JOL effect experiments bears resemblance to this procedure, but there are also differences. Some studies include an initial study phase in which the entire list is studied prior to making JOLs, whereas others do not (e.g., Dunlosky & Nelson, 1992; Theide & Dunlosky, 1994). During either the single initial phase or an intervening phase, some items are immediately followed by a JOL (i.e., the immediate JOL condition) whereas other items are followed by several intervening items prior to making a JOL for the earlier item (i.e., the delayed JOL condition). After this phase, a criterion memory test is given over all of the immediate and delayed JOL items.

A particularly important difference is that testing effect studies explicitly require participants to attempt to retrieve items, but it is only speculation that participants in delayed JOL studies may attempt to retrieve items as the basis for making their JOLs. Noting that standard delayed JOL tasks do not explicitly require retrieval, Nelson, Narens, and Dunlosky (2004) introduced a new methodology for studying immediate and delayed JOLs. This methodology is called the Pre-judgment Recall And Monitoring (PRAM) method. The major addition to the standard procedure was an explicit retrieval attempt for each item, prior to making a JOL for that item. When Nelson et al. used this procedure for immediate versus delayed JOLs, their results paralleled prior findings within the delayed JOL literature. They found greater accuracy of JOLs in predicting actual recall on a later test for delayed compared to immediate JOLs. Later recall was also greater for delayed compared to immediate JOL items. Nelson et al. were able to directly examine recall of items that occurred

immediately after presentation versus after a delay. They found that pre-judgment recall was more successful in the immediate JOL compared to the delayed JOL condition, a finding consistent with Nelson and Dunlosky's theory about retrieval from short-term versus long-term memory as a factor in the differential effects of the timing on JOL accuracy. Additionally, Van Overschelde and Nelson (2006) also used the PRAM methodology and discovered that delaying JOLs led to decreasing absolute accuracy (i.e., calibration) and increasing relative accuracy (i.e. resolution) as the delay between item presentation and making the JOL increased. Given that the conditions surrounding delayed JOLs may be similar to the test condition in the testing effect paradigm, this evidence may provide a reason to anticipate differences in the accuracy of JOLs for tested compared to restudied items. Although these data provide strong predictions about what might be revealed when implementing JOLs within a testing effect procedure, it is unclear how well the overall JOL literature translates to investigations of the testing effect given that most studies to not involve an explicit retrieval attempt (see Rhodes & Tauber, 2011, for a meta-analysis of the delayed JOL effect). Additionally, with the PRAM method introduced by Nelson et al., *all* items are subjected to a test (i.e., pre-judgment recall) prior to making a JOL for that item. This does not allow for the direct comparison between tested and restudied (or non-tested) items, as is typical in studies of the testing effect.

Note, however, that Dunlosky and Nelson (1992) did directly compare delayed and immediate JOLs for cue-target pairs when the cue was presented alone (i.e., similar to a test condition) and when the cue and target word were presented together (i.e., similar to a restudy condition). They found that delaying JOLs improved

resolution compared to immediate JOLs in the cue only condition, but not in the cue-target condition. This finding may lend support to the suggestion that testing is more conducive to improving monitoring than restudying.

Self-Regulated Study

One major component that is lacking in the delayed JOL literature (as well as the testing effect literature) is an examination of the effect of retrieval on subsequent control processes that participants may engage in after assessing their learning. In the metacognitive literature, it has been suggested that participants utilize control processes to change learning. In particular, researchers suggest that monitoring learning directly influences control processes such as deciding whether or not to continue studying information or deciding how much time to allocate to studying (e.g., Dunlosky & Hertzog, 1988; Mazzoni & Cornoldi, 1993; Mazzoni, Cornoldi, & Marchitelli, 1990; Nelson, 1996; Nelson, Dunlosky, Graf, & Narens, 1994; Thiede & Dunlosky, 1999). Dunlosky and Hertzog (1988) suggested the *Discrepancy Reduction Theory* to explain how participants decide which items to choose for additional study (see Metcalfe & Kornell, 2005; Son & Metcalfe, 2000, for an alternative theory). They suggest that participants tend to choose those items that have the largest gap between how well they have been learned and how well they need to be learned in order to reach some learning goal. Thiede and Dunlosky (1999) also suggest that people study information until they have met or exceeded their learning goals. In order to reach learning goals, they may allocate more time to items that are deemed to be less well-learned.

The current study will examine the relationships between monitoring and control for items that have been either tested or given a restudy opportunity. Along these lines,

Kornell and Son (2009) recently investigated self-regulated learning in an examination of the testing effect by asking participants to choose a study strategy. Using a flashcard style method, participants studied pairs of words. During the intervening phase, they were asked whether they would like to restudy the items in the list or take a test on them. Although participants' aggregate predictions of performance were higher for restudied items, they tended to choose testing as a study strategy more often than restudying. Furthermore, when asked *why* they chose testing as a study method, the majority of participants indicated that they chose testing to assess how well they knew the items, rather than as a means to enhance learning. The present experiments expand on prior work by examining how restudying and testing directly influences JOLs and self-regulation of further study on an item-by-item basis. If testing does indeed allow participants to better assess their learning, it would be expected that testing oneself compared to restudying information would improve participants' ability to differentiate between what they know and do not know (i.e., improving resolution). In addition, the suggestion that monitoring influences control processes leads to the prediction that participants would be more likely to choose the items in need of further study and to allocate more study time to those items when those items have been tested previously than when they have been restudied.

Overview of the Current Study

The current study seeks to bring together the two literatures on metacognition and the testing effect in order to better understand the benefits of testing on various aspects of memory and metacognitive processes as compared to restudying. In Experiment 1, a standard testing effect procedure was used in which each list of cue-target word pairs was

initially studied. Then half of the list items were restudied and half were subjected to a cued recall test during the intervening phase. During the intervening phase, participants were also asked to provide a JOL for each item, predicting the probability of remembering that item on a later memory test. After making their JOL, participants decided whether or not they would like to restudy the item before the final test. After all of the lists were presented, participants were given the opportunity to restudy the items that they chose during the intervening phase. In the last phase, participants were given a cued recall test over all of the previously encountered items. One potential limitation of Experiment 1 is that allowing some items to be restudied before the final memory test may enhance recall of those items. While it is important to examine the effects of restudy choices on later recall, this boost in exposure to some items may confound the relationship between restudy choices and whether or not those items are destined to be recalled. The procedures used in Experiment 2 were the same as in Experiment 1 except that participants were not given the opportunity to restudy the items that they chose during the intervening phase.

In Experiment 3, a different control process was examined to provide evidence that the findings of Experiments 1 and 2 generalize to other study behaviors. In particular, participants were asked to choose how much time they would like to allocate to each item if given the opportunity to restudy that item before the final memory test. Additionally, a between-subjects condition was added in which participants' restudy choices were either honored or dishonored in order to examine whether having control over study behaviors confers a greater memory benefit to restudied or tested items.

CHAPTER 2: EXPERIMENT 1

Method

Participants

Fifty Colorado State University students in freshman and sophomore level psychology classes participated for course credit. One participant was removed from the analyses because they did not recall any correct items on the final test. Participants were tested individually in a session that lasted approximately 1 hr.

Materials and Design

The experiment used a 2 (Intervening Task: test vs. restudy) x 2 (Measure: JOL vs. recall) within-subjects design. The materials included 192 words selected using the MRC Psycholinguistic database (Wilson, 1988). The words contained four to eight letters, had an average Kucera and Francis (1967) frequency of 65.72 (*SD* = 89.13) and an average concreteness value of 495.54 (*SD* = 107.86). Ten lists of eight unrelated word pairs were generated using the selected words. Each pair of words was checked for relatedness with the Nelson, McEvoy, and Schrieber (1998) word association norms. Any pairs that were related (i.e., that had a non-zero value in the Nelson et al. norms), or for which the cue word was not listed in the database, were replaced. Four additional word pairs were selected to be used as a practice list at the beginning of the experiment. The assignment of words to conditions was counterbalanced such that items were tested or restudied during the intervening phase equally often. Participants were provided with paper answer sheets to record their responses during the experiment.

Procedure

Participants first studied lists of word pairs during the initial study phase, with instructions that they should prepare for a later test in which they would be given the first word of the pair (i.e., the cue word) followed by a blank, which would serve as a cue to remember the second word (i.e., the target). Word pairs were presented one-at-a-time for 4 s each, with a 500 ms interstimulus interval (ISI). After the initial presentation of each list, participants engaged in a 15 s distracter task in which they added together a string of single digits. Next, the list of word pairs was presented again during the intervening phase. In this phase, half of the word pairs from the list appeared on the screen intact (e.g., PRINCE - OCEAN), whereas the other half of the word pairs were presented with the cue word and a blank (e.g., PRINCE - _____). If the pair was presented intact, participants were asked to study the pair again and write down target word on their answer sheet. This was called the *restudy* condition. If they were given the cue word alone, they were asked to recall the word that was paired with it during the first presentation of the list and write it on the answer sheet. This was the *test* condition. Each item was presented for 5 s, followed by a 500 ms ISI. The items were presented in a fixed, random order for the intervening phase that differed from the order used in the study phase.

After writing down each target word, a second screen appeared in which participants were asked to make a judgment of learning (JOL) indicating the likelihood on a scale from 0%-100% that they would be able to recall the target word when given the cue word on a later memory test. Participants were asked to give a rating of 0 if they were absolutely unlikely to recall the item later, a rating of 100 if they were absolutely

likely to recall the item later, and to give varying predictions between 0 and 100 for intermediate levels of confidence. They were given 4 s to make each JOL, followed by a 500 ms ISI. Next, participants indicated whether they would like the opportunity to restudy that item at a later time by choosing "yes" or "no." Participants were given 4 s to respond followed by a 500 ms ISI before the next restudy or test item appeared. After restudying or being tested on the final word pair during the intervening phase, participants received the next study list and repeated this procedure until all 10 lists had been given.

Following the final list, participants were given the opportunity to restudy the items that they chose during the intervening phase. Each of these items was presented again for 3 s followed by a 500 ms ISI. After all items chosen were restudied, participants engaged in a distracter task in which they attempted to list as many U.S. states as they could for 5 min. After the distracter task, participants were given a cued recall test for all word pairs. For each pair, participants were given the cue word and were asked to recall the target (e.g., PRINCE - _____). Each cue word was presented for 8 s with a 500 ms ISI. Cue words were also presented in a fixed, random order at test that differed from the order used during the encoding and intervening phases.

Results and Discussion

Calibration Measures

JOLs and Recall. Predicted and actual recall performance is presented in Figure 1.

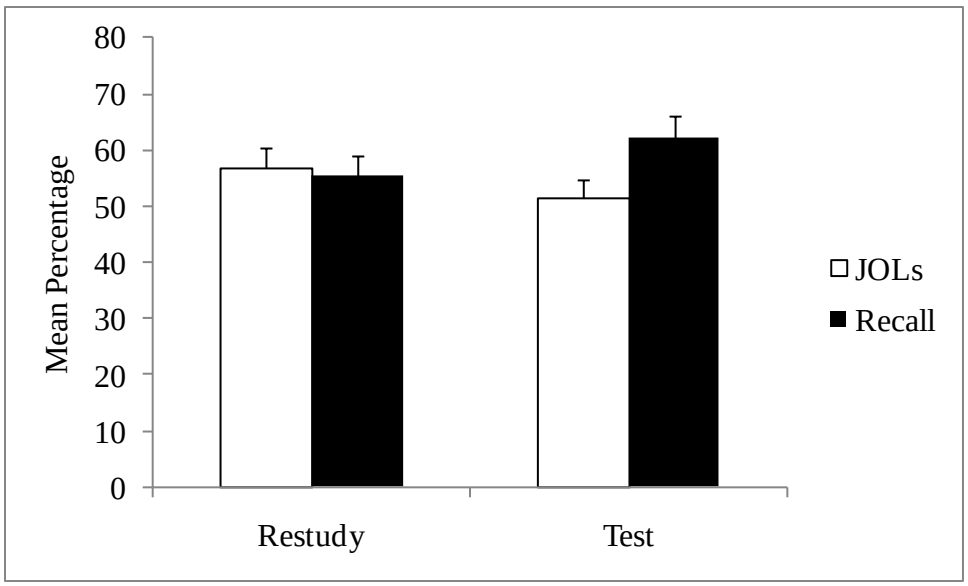

Figure 1. Experiment 1 mean JOL and recall percentages for restudied and tested items.

As a measure of calibration, the averages of item-by-item JOLs made for restudied items and for tested items were calculated. These JOL averages along with average recall performance for each condition were submitted to a 2 (Intervening Task: test vs. restudy) x 2 (Measure: JOL vs. recall) repeated-measures ANOVA. There was no main effect of intervening task, $F(1,48) = 1.51$, $p > .05$, and no differences overall between JOLs and recall performance, $F < 1$. However, there was a reliable Intervening Task x Measure interaction, $F(1,48) = 20.08$, $p < .05$, $MSE = 88.41$, $\eta_p^2 = .30$.

Post-hoc tests were conducted to examine the interaction. To protect against Type I error, the probability was set at $p < .01$ for the four tests conducted in the post-hoc

analyses. There was a significant testing effect, with a higher percentage of items recalled for previously tested items (M = 62.35, SD = 25.70) compared to restudied items (M = 55.51, SD = 25.03), $t(48) = -5.13$, $p < .01$, *Cohen's d* = .27. There was a non-significant trend found in which JOLs were higher for restudied (M = 56.77, SD = 24.74) compared to tested items (M = 51.56, SD = 23.10), $t(48) = 2.30$, $p = .026$, *Cohen's d* = .22. Thus, although participants recalled more tested than studied items overall, participants' JOLs did not reflect the benefits of testing on memory compared with restudying. The post-hoc tests also showed that participants' JOLs for restudied items were not reliably different from recall performance for those items, suggesting that predictions were calibrated with recall, $t(48) = 0.26$, $p > .01$. However, participants' recall performance exceeded JOLs made for tested items, suggesting that participants were underconfident in their abilities to recall those items on a later test, $t(48) = -3.30$, $p < .01$, *Cohen's d* = .44.

JOLs and Recall for Items Chosen vs. Not Chosen for Restudy. Additional analyses were conducted to compare data for items that were selected to be restudied again before the final memory test to those that were not selected. Predicted and actual recall performance for items chosen versus not chosen for additional restudy is presented in Figure 2. To examine JOLs, a 2 (Intervening Task: test vs. restudy) x 2 (Restudy Choice: restudied vs. non-restudied) repeated-measures ANOVA was conducted. There was a main effect of intervening task, in which higher JOLs were made for items that were restudied (M = 54.34, SD = 26.42) compared to tested (M = 47.52, SD = 30.79) during the intervening phase, $F(1,36) = 6.57$, $p < .05$, MSE = 261.84, $\eta_p^2 = .15$. There was also a main effect of restudy choice in which higher JOLs were given for items not chosen for restudy (M = 66.24, SD = 23.37) compared to items chosen (M = 35.62, SD =

26.15), $F(1,36) = 65.74$, $p < .05$, $MSE = 527.55$, $\eta_p^2 = .65$. A reliable Intervening Task x Restudy Choice interaction was also found, suggesting that the disparity between JOLs made for items chosen for restudy and not chosen differed depending on whether the items were restudied or tested during the intervening phase, $F(1,36) = 29.41$, $p < .05$, $MSE = 176.30$, $\eta_p^2 = .45$.

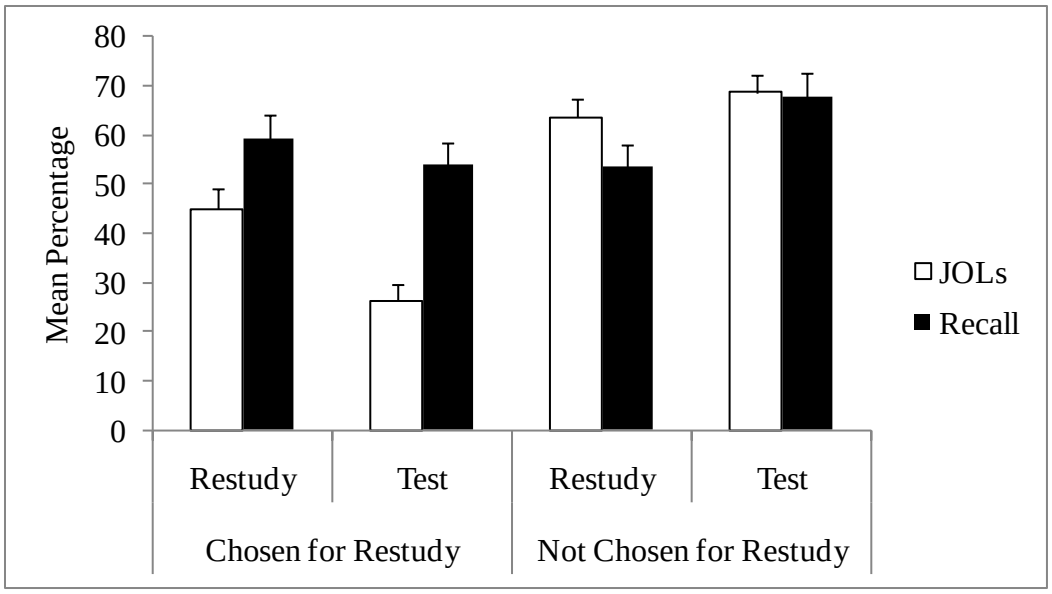

Figure 2. Experiment 1 mean JOL and recall percentages for restudied and tested items that were chosen versus not chosen for restudy.

Post-hoc tests were conducted to examine the interaction further. For items chosen for restudy, JOLs were higher for items that were initially restudied ($M = 44.09$, $SD = 26.20$) compared to those tested during the intervening phase ($M = 26.50$, $SD = 20.83$), $t(39) = 4.77$, $p < .01$, *Cohen's d* = .74. There was no difference, however, between JOLs made for previously restudied and tested items that were not chosen for additional study, $t(45) = -1.37$, $p > .01$. When examining only items restudied during the intervening phase, it became evident that participants made higher JOLs for items that were not chosen for later restudy ($M = 63.73$, $SD = 22.05$) compared to those chosen for restudy ($M = 44.95$, $SD = 26.61$), $t(36) = -5.40$, $p < .01$, *Cohen's d* = .77. When tested

items were examined alone, the same pattern resulted in which JOLs were higher for items not chosen (M = 66.55, SD = 23.71) compared to those chosen for restudy (M = 26.68, SD = 23.83), $t(44) = -8.77$, $p < .01$, *Cohen's d* = 1.68. This suggests that participants' control processes corresponded well with their predictions of later recall in that they desired to restudy items that they thought were less likely to be remembered later.

To examine recall, a 2 (Intervening Task: test vs. restudy) x 2 (Restudy Choice: chosen for restudy vs. not chosen) repeated-measures ANOVA was conducted. There was a marginally significant main effect of intervening task such that overall recall was higher for tested (M = 60.98, SD = 31.50) than restudied (M = 56.57, SD = 29.59) items, $F(1,39) = 3.85$, $p = .06$, $MSE = 202.00$, $\eta_p^2 = .09$. There was no main effect of restudy choice, $F<1$. However, the interaction between intervening task and restudy choice was reliable, $F(1,39) = 15.43$, $p < .05$, $MSE = 242.73$, $\eta_p^2 = .28$.

Post-hoc tests were conducted to examine the interaction with a probability set at .01. These tests revealed a testing effect for items not chosen for restudy, with higher recall for tested (M = 65.73, SD = 29.82) compared to restudied items (M = 52.74, SD = 28.04), $t(48) = -5.819$, $p < .01$, *Cohen's d* = .45. However, there was no difference between recall of restudied and tested items that were chosen for restudy, $t(39) = 1.35$, $p > .01$. Recall was also higher for tested items not chosen for restudy (M = 66.37, SD = 29.79) compared with tested items chosen for restudy (M = 52.13, SD = 31.98), $t(47) = -2.77$, $p > .01$, *Cohen's d* = .44. Overall, it appears that testing only improved recall above and beyond restudying for the presumably less difficult items that were not chosen for another restudy opportunity.

Resolution Measures: Gamma Correlations

In all of the experiments reported in the present study, Goodman-Kruskal Gamma correlations (Goodman & Kruskal, 1954) were used as a measure of resolution. Gamma (*G*) is a nonparametric measure of association that varies from -1.0 to +1.0. It is determined by a ratio of concordant and discordant pairs of items. Concordant pairs of items exhibit consistent ordering for the predictor and criterion variables (e.g., item 1: JOL of 80, recall of 1; item 2: JOL of 50, recall of 0), while discordant pairs exhibit inconsistent ordering (e.g., item 1: JOL of 80, recall of 0; item 2: JOL of 50, recall of 1). The formula for Gamma is:

$$G = \frac{\text{\# concordances - \# discordances}}{\text{\# concordances + \# discordances}}$$

(Goodman & Kruskal, 1954; Nelson, 1984).

JOLs and Final Recall. Gamma correlations were calculated between item-by-item JOLs and recall performance on the final test (See Table 1). Separate one-sample t-tests showed that mean gammas were different from zero for both restudied items, $t(46) = 5.40$, $p < .05$, and tested items, $t(46) = 9.79$, $p < .05$. A paired-samples t-test showed that gamma correlations were significantly higher for previously tested items ($M = .52$, $SD = .37$) compared with restudied items ($M = .30$, $SD = .40$), $t(44) = -3.52$, $p < .05$, *Cohen's d* = .57. These results suggest that participants were better at discriminating between items that they would or would not remember on a later test when they were tested on those items compared with restudying the items.

Table 1. Mean Gamma Correlations for Restudied and Tested Items

Comparison	Experiment	Restudy	Test
JOLs and Recall	1	.30 (.40)	.52 (.37)
	2	.35 (.30)	.74 (.20)
JOLs and Restudy Choice	1	-.53 (.54)	-.77 (.35)
	2	-.67 (.37)	-.93 (.09)

Note. Standard deviations in parentheses.

JOLs and Restudy Choices. Gamma correlations were also calculated to examine the relationship between JOLs and the decision to restudy (See Table 1). These correlations were negative, suggesting that participants were more likely to choose to restudy items that had received low JOLs. Separate one-sample t-tests showed that mean gammas were different from zero for both restudied items, $t(37) = -6.014$, $p < .05$, and tested items, $t(44) = -11.56$, $p < .05$. A paired-samples t-test showed that participants' JOLs were more strongly correlated with the decision to restudy for previously tested items ($M = -.77$, $SD = .35$) compared with restudied items ($M = -.53$, $SD = .54$), $t(37) = 3.18$, $p < .05$, *Cohen's d* = -.53. Thus, the relationship between restudy choices and JOLs was stronger for tested items.

CHAPTER 3: EXPERIMENT 2

In Experiment 1, a key finding was that taking an intervening test improves participants' discrimination between what they know and what they do not know (i.e., what they will or will not remember on a later test). Although this relationship was found for both initially restudied and tested items, it was reliably stronger for items subjected to a prior test. Additionally, it was observed that engaging in an initial test improved control processes over that of restudying items during the intervening phase. Items that were tested previously showed a stronger correlation between JOLs and restudy choices compared to initially restudied items, such that items that were given lower JOLs tended to be chosen for restudy and items given higher JOLs were not chosen.

One puzzling result in Experiment 1 was the lack of a reliable testing effect in recall performance for items that were chosen for restudy. One might expect that among those items chosen for additional restudy, those that were initially tested would show greater memory performance than items that were initially restudied. The results of Experiment 1 show that participants make more accurate discriminations between what they know and do not know in the test condition than in the restudy condition. When selecting items for restudy, participants were also more likely to choose the items they did not know in the intervening test condition compared to the intervening restudy condition. Therefore, one might predict that testing, by virtue of the improved monitoring and control processes, would benefit subsequent recall to a greater degree than that found

in the intervening restudy condition. The results of Experiment 1 did not support this prediction, however.

One possible explanation for this unexpected finding is that participants were afforded the opportunity to restudy the items that they chose before the final memory test. This additional exposure to the chosen items would likely boost recall regardless of whether items were restudied or tested during the intervening phase, thereby reducing the testing effect. Another potential problem with allowing participants to restudy the items that they chose is that the effects on final recall might change the relationships observed between average JOLs and recall (i.e., calibration) and between JOLs and whether or not items are recalled on the final test (i.e., resolution). Thus, Experiment 2 was conducted to using the same procedure as Experiment 1 except that participants were not allowed to restudy any items before the final memory test.

Method

Participants, Materials, Design, and Procedure

Thirty-six Colorado State University students participated for course credit. Experiment 2 utilized the same materials, design, and procedure as Experiment 1. However participants were not given the opportunity to restudy the items that they chose during the intervening phase before the final test. Instead, they proceeded directly to the 5 min distracter task immediately after the last list, and they then completed the final cued recall test.

Results and Discussion

Calibration Measures

JOLs and Recall. Predicted and actual recall performance is presented in Figure 3. Average item-by-item JOLs and average recall performance data were examined in a

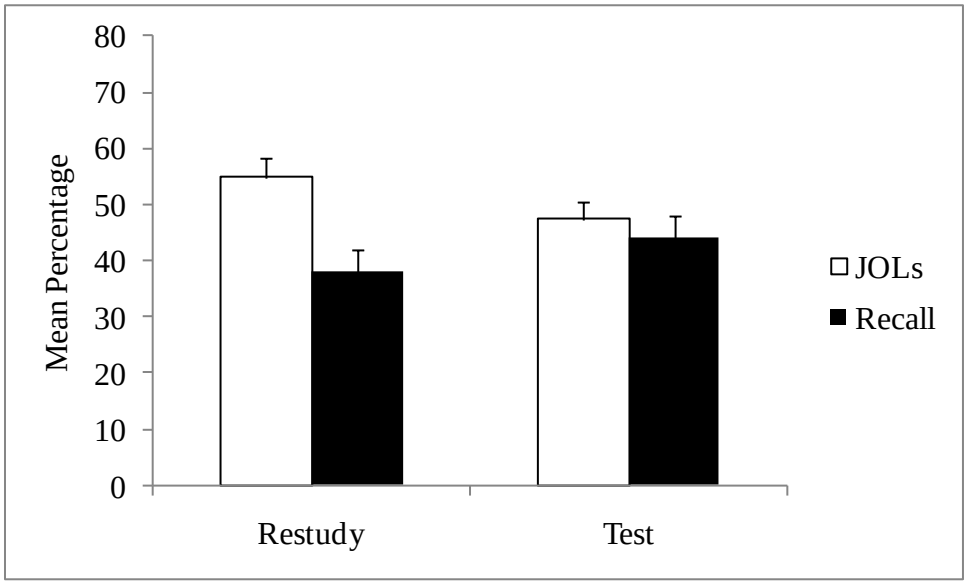

Figure 3. Experiment 2 mean JOL and recall percentages for restudied and tested items.

2 (Intervening Task: test vs. restudy) x 2 (Measure: recall vs. JOL) repeated-measures ANOVA. There was no main effect of Intervening Task, $F<1$. However, a main effect of Measure indicated that overall, JOLs (M = 51.05; SD = 21.04) were reliably higher than recall performance (M = 41.04; SD = 23.00), $F(1,35) = 7.39$, $p < .05$, MSE = 488.47, η_p^2 = .17. A significant Intervening Task x Measure interaction was also evident, $F(1,35)$ = 30.13, $p < .05$, MSE = 52.43, η_p^2 = .46.

Post-hoc tests conducted to examine the interaction showed a significant testing effect, with a reliably higher percentage of items recalled for previously tested items (M = 43.96, SD = 23.83) compared to restudied items (M = 38.13, SD = 22.07), $t(35) = -3.79$, $p < .01$, *Cohen's d* = .25. In addition, JOLs were reliably higher for restudied (M = 54.76, SD = 22.25) compared to tested items (M = 47.34, SD = 19.37), $t(35) = 3.23$, $p<.01$, *Cohen's d* = .36. Thus, although participants recalled more tested than restudied items, participants' JOLs did not reflect the benefits of testing on memory. The post-hoc tests also showed that participants' JOLs for restudied items reliably exceeded actual recall performance, indicating overconfidence, $t(35) = 3.70$, $p < .01$, *Cohen's d* = .75. However, participants' predictions were better calibrated with performance for previously tested items, as there was no reliable difference between JOLs and recall, $t(35) = 1.08$, $p > .01$. Note that these results differ from Experiment 1, in which participants' predictions were calibrated for restudied items and they were underconfident for tested items. This is likely because participants were not allowed to restudy any items before the final test, thereby reducing recall performance for both the test and restudy conditions as compared to Experiment 1. The JOLs in the present experiment were similar to those observed in Experiment 1, but recall declined. This yields the observed shift from calibrated predictions for restudied items and underconfidence in tested items in Experiment 1 to overconfidence in restudied items and calibrated predictions for tested items in Experiment 2.

JOLs and Recall for Items Chosen vs. Not Chosen for Restudy. Additional analyses were conducted to compare data for items that were selected to be restudied again before the final memory test to those that were not selected. Recall, however, that

participants in Experiment 2 were not actually given the opportunity to restudy the items they selected before the final test. Predicted and actual recall performance for items chosen versus not chosen for additional restudy is presented in Figure 4. To examine recall, a 2 (Intervening Task: test vs. restudy) x 2 (Restudy Choice: chosen for restudy vs. not chosen) repeated-measures ANOVA was conducted.

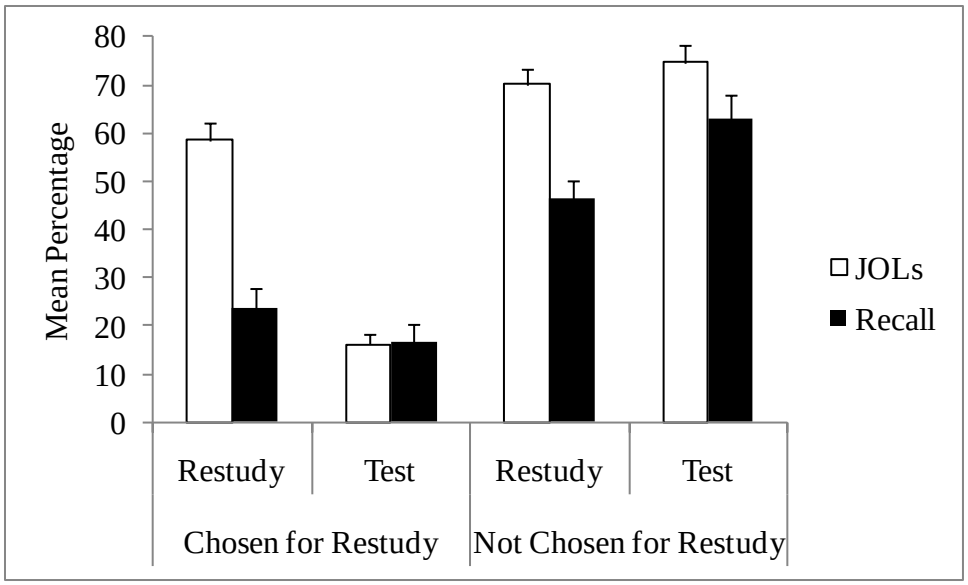

Figure 4. Experiment 2 mean JOL and recall percentages for restudied and tested items that were chosen versus not chosen for restudy.

A main effect of intervening task revealed a testing effect in which recall was higher overall for tested (M = 40.00, SD = 31.58) compared to restudied (M = 34.87, SD = 24.65) items, $F(1, 25) = 5.99$, $p < .05$, MSE = 114.39, η_p^2 = .19.There was also a main effect of restudy choice, such that recall was lower overall for items chosen for restudy (M= 20.19, SD = 21.71) compared to those not chosen (M = 54.69, SD = 25.55), $F(1, 25) = 112.28$, $p < .05$, MSE = 275.64, η_p^2 = .82. This suggests that participants' restudy choices were those items that they were more likely to forget. The interaction between

intervening task and restudy choice was also reliable, $F(1, 25) = 29.15$, $p < .05$, $MSE = 123.59$, $\eta_p^2 = .54$.

Post-hoc tests were conducted to further examine the interaction. For items tested during the intervening phase, recall was greater for items that they did not choose to restudy ($M = 53.71$, $SD = 24.08$) than items that were chosen for restudy ($M = 15.76$, $SD = 18.88$), $t(30) = -13.01$, $p < .01$, *Cohen's d* = 1.75. The same pattern was found for items that were restudied during the intervening phase, with greater recall for items that they did not choose to restudy ($M = 44.53$, $SD = 21.95$) than items that were chosen for restudy ($M = 24.16$, $SD = 22.91$), $t(26) = -4.51$, $p < .01$, *Cohen's d* = .91. When examining items not chosen for restudy, there was a significant testing effect in which recall of tested items ($M = 62.11$, $SD = 25.36$) was higher than recall of previously restudied items ($M = 46.40$, $SD = 22.00$), $t(31) = -5.35$, $p < .01$, *Cohen's d* = .66. However, there were no differences between previously restudied and tested items for the items chosen for restudy, $t(29) = 1.87$, $p > .01$. Thus, as observed in Experiment 1, there was not a reliable testing effect for those items that participants chose to restudy, counter to what one would expect if the observed advantage in monitoring and control processes for tested items over restudied items translates into a free recall advantage. Moreover, this lack of a testing effect does not appear to result from a confound of actually allowing participants to restudy chosen items, given that no such opportunity was provided in the present experiment.

A 2 (Intervening Task: test vs. restudy) x 2 (Restudy Choice: restudied vs. non-restudied) repeated-measures ANOVA was also conducted on JOLs made for restudied and tested items. There was a main effect of intervening task, in which higher JOLs were

made for items that were restudied during the intervening phase (M = 64.37, SD = 22.85) compared to tested (M = 45.39, SD = 33.42) during the intervening phase, $F(1,30)$ = 81.31, $p < .05$, MSE = 137.33, η_p^2 = .73. There was also a main effect of restudy choice in which higher JOLs were given for items not chosen for restudy (M = 72.40, SD = 21.48) compared to items chosen (M = 37.36, SD = 26.52), $F(1,30)$ = 166.73, $p < .05$, MSE = 228.29, η_p^2 = .85. A reliable Intervening Task x Restudy Choice interaction was also found, consistent with Experiment 1, $F(1,30)$ = 150.41, $p < .05$, MSE = 114.58, η_p^2 = .83.

Post-hoc tests were conducted to further examine the interaction. For test items, JOLs were higher for items not chosen for restudy (M = 74.70, SD = 20.25), compared to items chosen for restudy (M = 16.08, SD = 13.30), $t(30)$ = -14.72, $p < .01$, *Cohen's d* = 3.42. For items restudied during the intervening phase, the same pattern was found with higher JOLs for items not chosen for restudy (M = 67.99, SD = 20.25), compared to items chosen for restudy (M = 57.94, SD = 19.89), $t(32)$ = -3.87, $p < .01$, *Cohen's d* = .50 . In examining items that were chosen for restudy only, participants made higher JOLs for previously restudied items (M = 54.28, SD = 22.47), compared to tested items (M = 17.04, SD = 13.94), $t(34)$ = 8.89, $p < .01$, *Cohen's d* = 1.99. For items not chosen for restudy, there was no difference between previously restudied and tested items, $t(31)$ = -1.65, $p < .01$.

Resolution Measures: Gamma Correlations

JOLs and Final Recall. In the case of JOLs and recall, a positive gamma correlation indicated that participants provided higher JOLs when items were not recalled compared with items that were recalled. Gamma correlations (See Table 1) were calculated between item-by-item JOLs and recall performance on the final test. Separate

one-sample t-tests showed that mean gammas were different from zero for both restudied items, $t(35) = 7.14$, $p < .05$, and tested items, $t(35) = 21.82$, $p < .05$. A paired-samples t-test showed that gamma correlations were significantly higher for previously tested items ($M = .74$, $SD = .20$) compared with restudied items ($M = .35$, $SD = .30$), $t(35) = -7.30$, *Cohen's d* = 1.53. These findings are consistent with Experiment 1, lending additional evidence to support that testing when compared with restudying the items enables participants to better differentiate between items that they will or will not remember on a later test.

JOLs and Restudy Choices. Gamma correlations (See Table 1) were also calculated to examine the relationship between JOLs and the decision to restudy. These correlations were negative, suggesting that participants were more likely to choose to restudy items that had received low JOLs. One-sample t-tests indicated that gammas for both restudied items, $t(31) = -7.53$, $p < .05$, and tested items, $t(30) = -58.90$, $p < .05$, reliably differed from zero. A paired-samples t-test showed that participants' JOLs were more strongly correlated with the decision to restudy for previously tested items ($M = -.93$, $SD = .09$) compared with restudied items ($M = -.67$, $SD = .37$), $t(30) = 3.98$, *Cohen's d* = -.97. Thus, the relationship between restudy choices and JOLs was stronger for tested items, consistent with Experiment 1.

CHAPTER 4: EXPERIMENT 3

Experiments 1 and 2 both suggest that taking a prior memory test enhances both recall and metacognitive accuracy on an item-by-item basis (i.e., resolution). However, both experiments showed a testing effect in final recall of items that were not chosen for restudy, but a lack of a testing effect for items that were chosen for restudy. One possible explanation for these results is that there was an item selection artifact in the comparison of items chosen for restudy from the test condition compared to the initial restudy condition. Given that both the relationship between JOLs and actual recall performance and the relationship between JOLs and restudy choice was stronger for initially tested compared to restudied items, it is possible that the subset of items chosen for restudy in the two intervening task conditions differed with regard to item difficulty. For instance, it may be the case that in the intervening restudy condition, items chosen for restudy included both items that actually merit further attention and items that do not. However, in the intervening test condition, it may be the case that the majority of the items chosen for restudy actually merit further study (i.e., are the most difficult to remember). This would in turn put the intervening restudy condition items at an advantage for final recall.

Experiment 3 examined metacognitive control processes using a different measure of control. Specifically, participants were asked to indicate how much time they would like to allocate to restudying each item, instead of making a yes/no decision about whether or not to restudy the items. As such, the present experiment examines whether

the findings of Experiments 1 and 2 generalize to other control processes. Moreover, because this alternative measure of control involves a procedure in which all items are selected for restudy, it may reduce selection artifacts.

Experiment 3 also adopted a different approach to examining the benefits of control processes based on a study by Kornell and Metcalfe (2006), whereby restudy choices were honored for some participants and dishonored for other participants. Kornell and Metcalfe (2006) found that memory was better when participants were in control of their study choices (i.e., honored choices) than when they were not (i.e., dishonored choices). In the present experiment, some participants were allowed to restudy items for the amount of time they chose (i.e., honored condition), whereas other participants were given a standard amount of time to study each item, regardless of their time choices (i.e., dishonored condition). This method was used to examine whether the ability to control study processes is more beneficial for initially restudied or initially tested items.

Method

Participants

One hundred participants completed Experiment 3. They were Colorado State University students in freshman and sophomore level psychology courses who received course credit for their participation. Half of the participants were randomly assigned to the honored condition and the other half were randomly assigned to the dishonored condition. The experiment was completed in a single session lasting approximately 1 hr. Six participants, five in the honor condition and one in the dishonor condition, were

removed from the analysis because of failure to follow instructions, leaving a total of 93 participants (44 honored, 49 dishonored).

Materials, Design, and Procedure

Experiment 3 used the same materials, design, and procedure as Experiments 1 and 2 with a few changes. The number of lists was reduced from 10 lists of eight items to eight lists in order to keep the experiment within a one hour session, regardless of participant time choices. During the intervening phase, participants' provided a JOL for each item, then were asked to choose whether they would like 1, 3, or 5 seconds to restudy the item before the final memory test. Participants were not instructed as to how many items to choose for each time category, but they were instructed to choose *some* items for each amount of time. All of the items were re-presented during a restudy phase following the last list. For participants in the honored condition, each item was presented again for the amount of time that the participant chose for that item (e.g., 1, 3, or 5 seconds). For participants in the dishonored condition, all items were re-presented for 3 seconds during the restudy phase, regardless of participants' time choices.

Results and Discussion

Calibration Measures

JOLs and Recall. The predicted and actual recall performance was examined in a 2 (Intervening Task: test vs. restudy) x 2 (Measure: recall vs. JOL) x 2 (Control: honored vs. dishonored) mixed ANOVA. These data are shown in Figure 5. There was a main effect of measure such that overall recall (M = 59.93, SD = 26.48) was higher than overall predictions of performance (M = 54.09, SD = 21.17), F (1,91) = 5.37, MSE = 590.07 η_p^2

= .056, $p < .05$. The main effect of intervening task was not significant, $F(1,91) = 2.74$, $p > .05$. There was also no main effect of control, $F < 1$. There was, however, a reliable measure X intervening task interaction, $F(1,91) = 8.43$, $MSE = 111.80$ $\eta_p^2 = .085$, $p < .05$. The intervening task X control interaction and the three-way interaction between measure, intervening task, and control were not significant, $Fs < 1$.

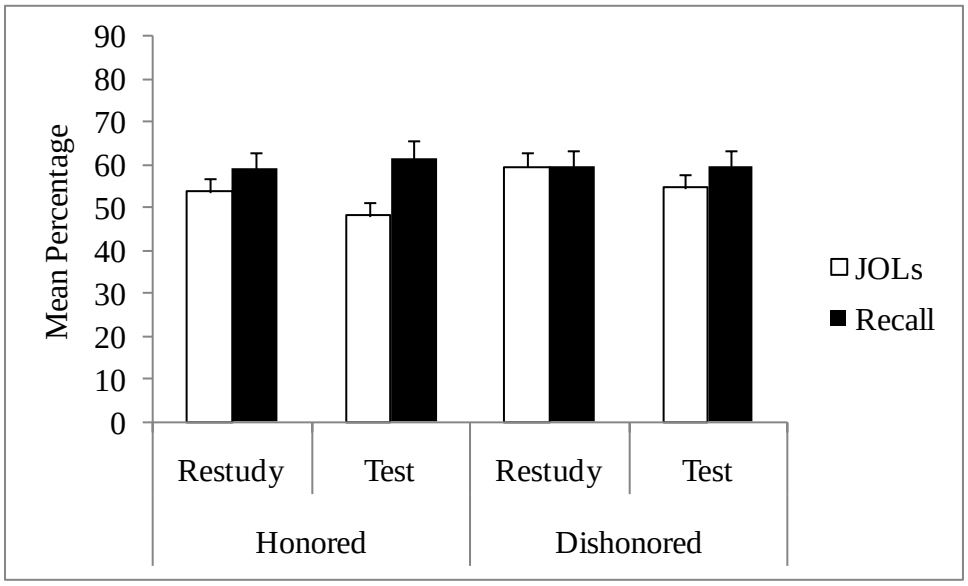

Figure 5. Experiment 3 mean JOL and recall percentages for restudied vs. tested items for which study time choices were honored or dishonored.

Post-hoc tests were conducted to investigate the Measure X Intervening Task interaction. These data, which are collapsed across the honored and dishonored conditions, are illustrated in Figure 6. The post-hoc tests showed that there was no difference in overall recall between previously restudied and tested items, $t(92) = -.712$, $p > .01$. However, it was revealed that JOLs were higher for restudied items ($M = 56.82$, $SD = 21.94$) compared to tested items ($M = 51.68$, $SD = 20.17$), a finding consistent with Experiments 1 and 2, $t(92) = 3.17$, $p < .01$, *Cohen's d* = .24. Recall of tested items ($M = 60.48$, $SD = 26.59$) was also higher than JOLs made for tested items ($M = 51.68$, $SD =$

20.17), indicating that participants were underconfident in their predictions of memory for tested items, $t(92) = -3.72$, $p < .01$, *Cohen's d* = .37. However, there was no difference between JOLs and recall for previously restudied items, $t(92) = -.81$, $p > .01$, suggesting that their predictions were calibrated with performance. This latter finding was consistent with Experiment 1.

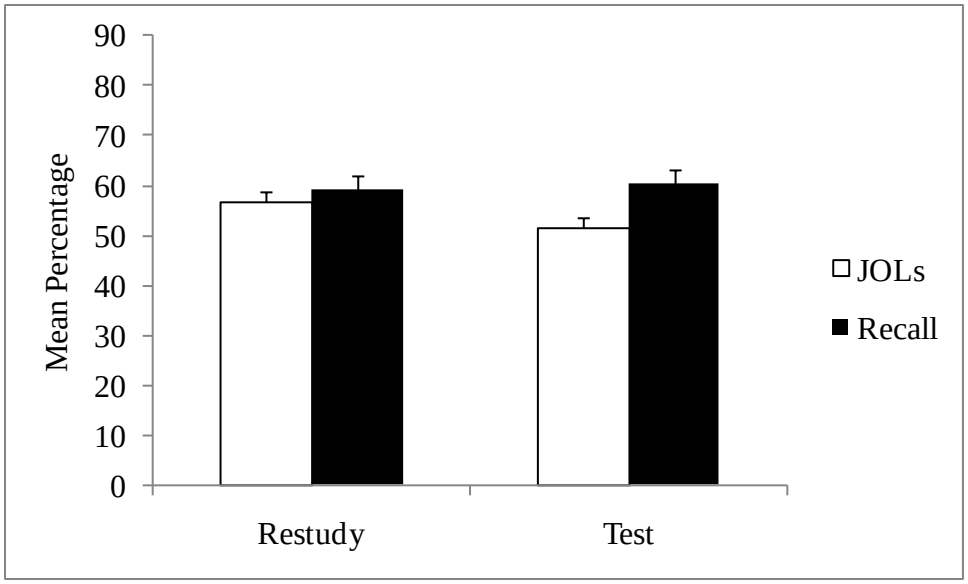

Figure 6. Experiment 3 mean JOL and recall percentages for restudied vs. tested items collapsed across honored and dishonored study time choices.

JOLs and Recall for Varying Restudy Time Choices. Predicted and actual recall performance for items chosen versus not chosen for restudy was examined in separate 2 (Intervening Task: test vs. restudy) x 3 (Time Choice: 1, 3, or 5s) x 2 (Control: honored vs. dishonored) mixed ANOVAs for JOL and recall data. In the examination of JOLs (See Figure 7), there was a main effect of intervening task such that JOLs were higher for restudied ($M = 55.16$, $SD = 21.94$) compared to tested items ($M = 49.93$, $SD = 20.17$), $F(1,66) = 8.10$, $MSE = 344.85$ $\eta_p^2 = .11$, $p < .05$. A main effect of time choice was also found such that JOLs decreased with increasing amount of time chosen, $F(1,66)$

= 89.59, *MSE* = 429.98 η_p^2 = .58, $p <$ .05. The interaction between intervening task and time choice was also reliable, suggesting that differences in JOLs for restudied and tested items depended on the amount of time chosen for restudy F (1,66) = 27.56, *MSE* = 234.71 η_p^2 = .30, $p < .05$ (See Figure 7).

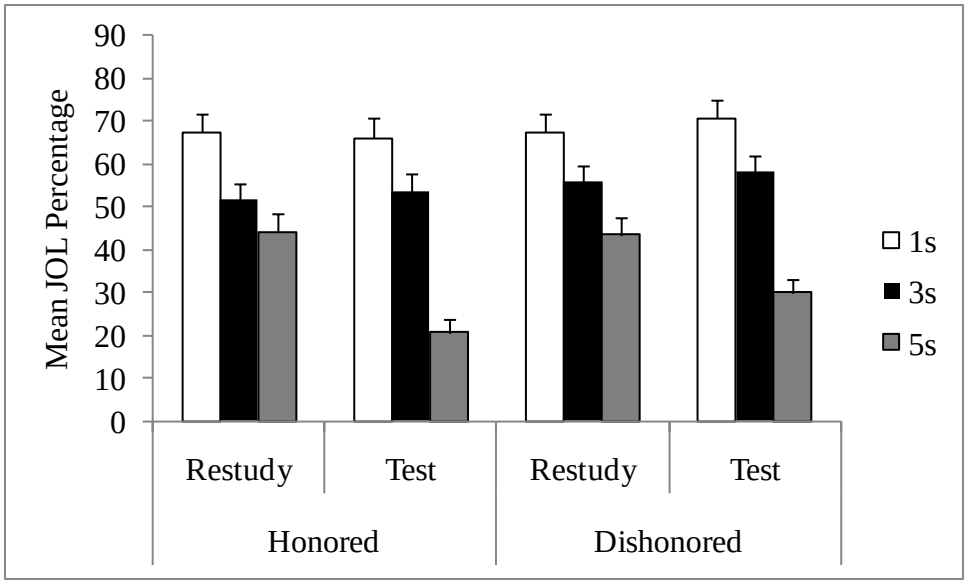

Figure 7. Experiment 3 mean JOLs for restudied versus tested items chosen for 1, 3, or 5s of restudy, separated by participants whose restudy choices were honored or dishonored.

Post-hoc tests were conducted to further examine the interaction. Nine comparisons were made to examine differences between the restudy and test conditions and between the three possible restudy time choices. To protect against Type I error, the alpha level was adjusted to .006. The post-hoc tests revealed that JOLs for previously restudied items (*M* = 45.66, *SD* = 25.52) were higher than JOLs for tested items (*M* = 28.06, *SD* = 22.36) for the items that were chosen to be restudied for 5s, $t(80) = -6.55$, $p <$ .006, *Cohen's d* = .73. However, there were no differences in JOLs for tested and restudied items for 1s and 3s items, $t(91) = .514$, $p > .006$ and $t(80) = .246$, $p > .006$, for the 1 and 3 s items, respectively. Presumably, participants chose items they perceived to be the most difficult to be included in the 5s category, which may provide an explanation

for why the only differences are found between those items. All other comparisons were significant at $p < .006$. These comparisons showed that for previously restudied items, JOLs for the 5s items were higher than JOLs for the 3s items, and JOLs for the 3s items were higher than JOLs for the 1s items. The same pattern emerged for tested items chosen for 1, 3, or 5s (See Figure 8).

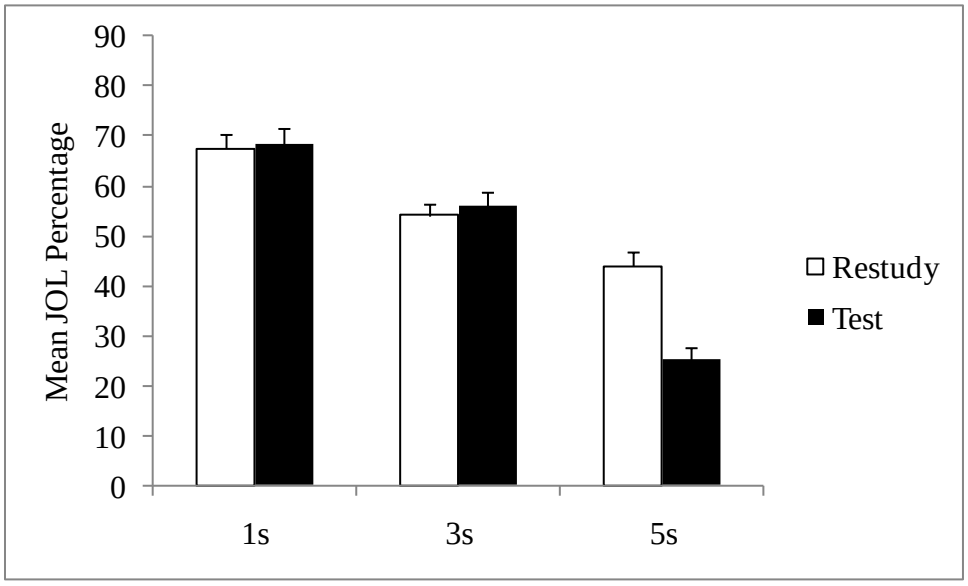

Figure 8. Experiment 3 mean JOLs for restudied versus tested items chosen for 1, 3, or 5s of restudy, collapsed across honored and dishonored choices.

A 2 (Intervening Task: test vs. restudy) x 3 (Time Choice: 1, 3, or 5s) x 2 (Control: honored vs. dishonored) mixed ANOVA was also conducted to examine recall data (See Figure 9). The main effect of intervening task was not significant, nor was the effect of control, $F(1,64) = 3.41$, $p > .05$ and $F < 1$, respectively. However, there was a reliable main effect of time choice such that lower recall occurred for items chosen for longer restudy times, $F(2, 128) = 60.53$, $p < .05$, $MSE = 419.72$, $\eta_p^2 = .49$. This suggests that items chosen for 5s were more difficult to recall. There was also a significant interaction between intervening task and time choice, suggesting that differences in recall

of restudied and tested items depended on the amount of time chosen for restudy, $F(2, 128) = 11.30$, $p < .05$, $MSE = 292.69$, $\eta_p^2 = .15$. The interactions between intervening task and control and between time choice and control were not significant, $Fs < 1$. The three-way interaction between intervening task, time choice, and control was also not significant, $F(2, 128) = 2.07$, $p > .05$.

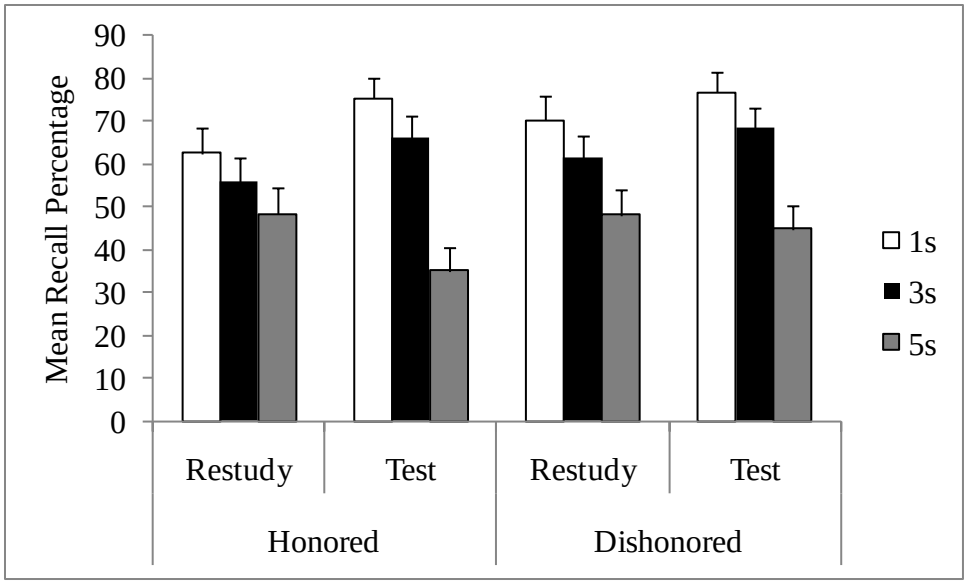

Figure 9. Experiment 3 mean recall for restudied versus tested items chosen for 1, 3, or 5s of restudy, separated by participants whose restudy choices were honored or dishonored.

Post-hoc tests were conducted to examine the interaction between intervening task and time choice (See Figure 10). The alpha level was adjusted to .006. The post-hoc tests showed a testing effect for items chosen for 1s and 3s, such that final recall of tested items ($M_{1s} = 77.78$, $SD_{1s} = 26.65$; $M_{3s} = 68.56$, $SD_{3s} = 28.52$) was higher than recall of previously restudied items ($M_{1s} = 66.82$, $SD_{1s} = 32.28$; $M_{3s} = 59.37$, $SD_{3s} = 30.34$), $t(76) = -4.19$, $p < .006$, *Cohen's d* = .37 and $t(91) = -3.98$, $p < .006$, *Cohen's d* = .31. For items chosen for 5s, there was a non-significant trend in the opposite direction in which recall was numerically higher for restudied items ($M = 51.59$, $SD = 32.02$) compared to tested

items (M = 43.12, SD = 31.44), $t(79) = 2.74$, $p = .008$, *Cohen's d* = .27. All other comparisons were reliable at $p < .006$. These comparisons showed that for both restudied and tested items, recall decreased with increasing restudy time choice (See Figure 10).

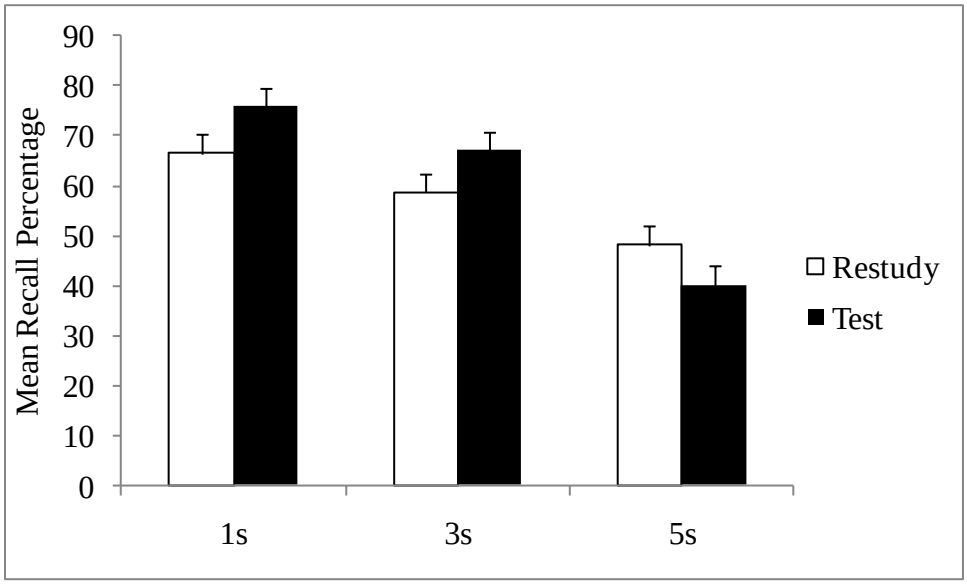

Figure 10. Experiment 3 mean recall for restudied versus tested items chosen for 1, 3, or 5s of restudy, collapsed across honored and dishonored choices.

Resolution Measures: Gamma Correlations

JOLs and Final Recall. Gamma correlations were calculated between item-by-item JOLs and recall performance on the final test (See Table 2). Separate one-sample t-tests showed that gamma correlations for both tested and restudied items were reliably different from zero, $t(89) = 22.34$ and $t(85) = 8.25$, $p < .05$, respectively. A paired-samples t-test also showed a stronger correlation for tested items (M = .69, SD = .29) compared to restudied items (M = .35, SD = .39), $t(83) = -6.29$, $p < .05$. These results are consistent with Experiments 1 and 2, suggesting that participants are better at differentiating between items that they will and will not remember on a later test for previously tested compared to restudied items.

JOLs and Restudy Time Choices. Gamma correlations were also calculated to examine the relationship between JOLs and the choice to restudy items for 1, 3, or 5s before the final memory test (See Table 2). A one-sample t-test indicated that the correlations were different from zero, t(85) = -7.45 and t(89) = -17.47 for restudied and tested items, respectively. A paired samples t-test revealed a stronger correlation for tested (*M* = -.64, *SD* = .36) compared to restudied (*M* = -.42, *SD* = .50) items, *t*(84) = 5.12, $p < .05$. These results coincide with those of Experiments 1 and 2, suggesting that in the test condition participants' are better at making relative choices about how to allocate their study time among items that they are in the restudy condition.

Table 2. Mean Gamma Correlations for Restudied and Tested Items

Comparison	Restudy	Test
JOLs and Recall		
Overall	.35 (.39)	.69 (.29)
1s items	.43 (.54)	.35 (.72)
3s items	.30 (.54)	.45 (.60)
5s items	.22 (.61)	.53 (.54)
JOLs and Restudy Choice	-.42 (.50)	-.64 (.36)

Note. Standard deviations in parentheses.

JOLs and Final Recall for Varying Restudy Time Choices. In addition to examining overall gamma correlations between JOLs and recall and between JOLs and restudy time choices, correlations were calculated to examine the relationship between

JOLs and recall for items chosen for either 1s, 3s, or 5s of restudy time (See Table 2). Separate one-sample t-tests indicated that the correlations for previously restudied items were significantly different from zero, $t(39) = 3.56$, $t(74) = 5.14$, and $t(61) = 3.03$, $p < .05$ for restudied items chosen for 1, 3, and 5s, respectively. Correlations for tested items were also different from zero, $t(42) = 3.73$, $t(65) = 6.68$, and $t(62) = 7.78$, $p < .05$ for tested items chosen for 1, 3, and 5s, respectively. Paired-sample t-tests revealed that for items chosen for 5s, Gamma correlations between JOLs and recall were stronger for tested ($M = .53$, $SD = .54$) compared to restudied items ($M = .22$, $SD = .61$), $t(51) = -2.89$, $p < .05$. However, gamma correlations for items chosen for 1s and 3s did not differ reliably for restudied and tested items, $t(26) = .44$ and $t(60) = -1.35$, respectively.

These results strengthen the argument that participants are better at choosing items that merit further study when they have been given a prior test on those items. This may also explain the differences in recall that are inconsistent with Experiments 1 and 2, in suggesting that the subset of items chosen for 5s in the test condition include a larger proportion of items that are actually less likely to be recalled later, whereas items chosen for 5s in the restudy condition may include a subset of items that are mixed in terms of how much additional study time is needed. Thus, items in the restudy condition may have an unfair advantage in recall over the items in the test condition, which may in turn eliminate or reverse the testing effect for these particular items.

CHAPTER 5: GENERAL DISCUSSION

The purpose of the current study was to examine how testing affects memory, and the accuracy of metacognitive monitoring and control, when compared to the effects of simply restudying material. The primary focus of this set of experiments was the testing effect, which suggests that retrieving items from memory or taking a test on them enhances subsequent recall (e.g., McDaniel et al., 2007; Roediger & Karpicke, 2006). In the current study, a metacognitive component was added to the standard testing effect procedure. Participants were asked to make predictions in the form of judgments of learning regarding their later recall performance for initially restudied and tested items (i.e., monitoring). They were also asked to make a decision about whether or not to restudy an item or how much additional time to spend restudying each item (i.e., control).

In Experiments 1 and 2, participants made predictions of memory performance and were then asked to choose whether or not they would like to restudy each item. In Experiment 1, participants were actually allowed to restudy the items that they chose during the intervening phase, whereas in Experiment 2 they were not afforded the opportunity to restudy any of the items before the final memory test. Despite this difference, consistent results were found across the two experiments. In both experiments, results suggested that participants' average judgments of learning (JOLs) were higher for restudied compared to tested items, suggesting greater overall confidence in memory for those items. In contrast to JOL data, patterns of actual recall demonstrated

that testing compared to restudying enhanced subsequent recall of information. In other words, although participants' overall predictions of memory for restudied items were greater than for tested items, they actually were more likely to recall tested items. Despite this mismatch between predictions and performance for restudied versus tested items, it was also discovered that for both restudied and tested items, judgments of learning were positively correlated with recall performance, such that relatively higher JOLs corresponded to successful recall on the final test. There was also a negative correlation between JOLs and restudy choice such that lower JOLs were associated with choosing items for restudy. Both of these correlations were stronger for tested compared to restudied items, suggesting that testing not only enhanced memory, but also enhanced metacognitive monitoring accuracy and influenced control processes.

In addition to the overall results in Experiments 1 and 2, the data were broken down and analyzed using participants' decisions to restudy items as a factor. Despite the difference between these two experiments in terms of allowing later restudy of chosen items (Experiment 1) or not allowing additional restudy (Experiment 2), results were consistent across experiments. The analysis of JOLs showed that participants made higher JOLs for restudied compared to tested items, but only for those items that were *not chosen* for additional restudy. There were no differences in JOLs for restudied compared to tested items for items that were chosen for later restudy. The analysis of recall similarly revealed a testing effect for items not chosen for additional restudy, whereas there were no differences for chosen items. Moreover, there was a greater mismatch between predicting recall and actual recall for items that were not chosen for restudy. Participants were more confident in the restudy condition for these items, but recalled

more of those items in the test condition. Gamma correlations between JOLs and recall suggest that participants were better at discriminating between items that they would and would not remember when those items had been subjected to a prior test. Additionally, Gamma correlations between JOLs and restudy choices suggest that participants' were more likely to choose to restudy those items for which they made lower JOLs when the items had been tested versus restudied.

One unexpected aspect of the data from Experiments 1 and 2 was that there was a testing effect for items *not* chosen for restudy, but no testing effect for items that were chosen for restudy. First, consider the finding that there was no testing effect for items chosen for restudy. Based on the evidence discussed above, it appears that the subset of items chosen for restudy in the test condition may have been different than the items chosen for restudy in the initially restudied condition. It is possible that participants chose a subset of tested items to restudy that had a greater need for restudy. It is also possible that within the subset of items chosen for restudy, the items from the intervening phase restudy condition may have included a set of items of varying difficulty levels for an individual, whereas the tested item set would likely include a greater proportion of items that were relatively more difficult. This possibility is suggested by the strong correlations between JOLs and both later recall and restudy choices in the test condition compared to the restudy condition. The selection of some items that merit further restudy and some items that do not in the restudy condition may put this set of items at an unfair advantage, thus eliminating the testing effect in the case of the items selected for restudy.

Next, consider the finding that there *was* a testing effect for items *not* chosen for restudy. Gamma correlations between JOLs and restudy choice might also suggest that

participants were better at selecting the items that they knew very well when those items had been tested, which in turn allowed them to *reject* those items for further restudy. The finding that there was a testing effect within the subset of items not chosen for restudy may therefore reflect a difference in the difficulty of this subset of items instead of, or in addition to an actual boost in memory performance.

Despite the unexpected results discussed above, most of the findings of Experiments 1 and 2 were replicated across experiments. The only inconsistent finding between the two experiments was with regard to overall calibration of JOLs and recall. This was measured by averaging across participants' individual item judgments and comparing that average to participants' average recall performance for restudied and tested items. In Experiment 1, JOLs and recall were calibrated for restudied items, but participants were underconfident in tested items. In Experiment 2, tested items were calibrated whereas participants were overconfident in their predictions for restudied items. However, inspection of the means for Experiments 1 and 2 suggests that this difference was driven by increased recall performance when restudying was allowed in Experiment 1 compared to when restudy was not allowed in Experiment 2.

Experiment 3 was conducted to extend the findings of Experiments 1 and 2 to another measure of control processes; namely, the allocation of restudy time (e.g., Mazzoni & Cornoldi, 1993; Metcalfe & Kornell, 2005; Kornell & Metcalfe, 2006). During the intervening phase of the experiment, participants were asked to make JOLs for restudied and tested items. For each item they were also asked to choose the amount of time they would like to spend restudying the item before the final test (1, 3, or 5 seconds). Overall results were consistent with Experiments 1 and 2 in that participants'

JOLs were higher on average for items restudied during the intervening phase compared to items that were tested. Calibration measures were consistent with Experiment 1 in which participants were given the opportunity to restudy items before the final test. In Experiment 3, participants' JOLs and recall were calibrated for restudied items, but indicated underconfidence for tested items.

Beyond the overall data discussed above, the JOL and recall data from Experiment 3 were also examined for differences based on the amount of time participants chose to restudy items. For recall, participants exhibited a testing effect for items chosen for 1s and 3s. However, a reverse effect was found for items chosen for 5 s, such that participants recalled more items in the restudy condition than in the test condition. Gamma correlations between JOLs and recall again suggested that participants were better at making relative predictions about what they would and would not recall later when they had been tested on those items previously. They also made restudy choices that showed a stronger negative correlation with the predictions made. In other words, participants made lower JOLs for items that were chosen for 5s versus 3s and 1s, and this relationship was stronger for previously tested items. As in Experiments 1 and 2, it could be suggested that the level of difficulty of the items that an individual selected for restudy in the test condition may have differed when compared with items restudied in the intervening phase. If participants were selecting a subset of more difficult items to restudy for 5 s in the test condition than in the restudy condition, the initially restudied items may have been at an unfair advantage in later recall. In order to examine this more closely, further studies would need to be conducted in which items difficulty is manipulated experimentally.

Experiment 3 also included a between subjects manipulation in which some participants' restudy choices were honored while others' choices were dishonored. In the honored condition, participants were allowed restudy each item for the amount of time they selected, whereas all items were presented at a standard 3 s rate in the dishonored condition. Kornell and Metcalfe (2006) utilized a similar procedure to examine the effects of control processes on memory performance. For each item, participants were asked to choose whether or not to restudy each item. In the honor condition, participants restudied the items they chose, whereas in the dishonor condition they were given the set of items they chose *not* to restudy. In several experiments, Kornell and Metcalfe demonstrated better final memory performance when participants' choices were honored than when they were dishonored. In other words, when participants were in control of which items they restudied, memory was enhanced by restudying. In Experiment 3 of the current study, however, there were no reliable differences in recall performance when restudy choices were honored versus dishonored. This is surprising given that many studies suggest that being in control of what is studied or how much time is allocated to studying improves memory (e.g., Atkinson, 1972; Mazzoni & Cornoldi, 1993; Nelson, 1994).

It is worth noting, however, that other research has suggested that differences in the conditions under which participants make their study choices can affect the choices that they make. For instance, Thiede and Dunlosky (1999) found that when participants were given a difficult learning goal (e.g., remembering 24/30 items) they tended to choose to restudy easier items, whereas when they were given an easy learning goal (e.g., remembering 6/30 items), they tended to choose difficult items. Kornell and Metcalfe

(2006) reported similar findings in that participants tended to choose to restudy items that had been presented to them, but were not fully learned, instead of choosing new items that they had not seen previously (i.e., the most difficult items because they have not been learned). These findings were explained using the Region of Proximal Learning (RPL) framework to suggest how participants made their restudy choices (e.g., Kornell & Metcalfe, 2006; Metcalfe & Kornell, 2005). This theory suggests that during the monitoring process participants rejected the items they had learned fully and then selected items from a subset of less-well learned items. Because they had had prior exposure to these items, the items fell within the individual's RPL. Furthermore, when participants selected items from the subset of partially learned information, they tended to choose the easiest items within the list of unlearned items (Kornell & Metcalfe, 2006). This is in contrast with the Discrepancy Reduction model (DR), which suggests that participants would have a tendency to choose to restudy the most difficult items in a learning set.

In short, some research suggests that item selection strategies rely in part on the circumstances surrounding the learning goals. It is therefore possible that procedural differences between the current study and prior research may explain differences as a function of honored and dishonored restudy choices. In particular, Kornell and Metcalfe (2006) instructed participants to choose only half of the items for restudy. Participants were able to track the number of items they had chosen because there was a running total on the computer screen so that they could accurately choose half of the items as they went through each items one-by-one. In Experiments 1 and 2 of the current study, participants were told that they should choose to restudy *some* items and not others, but

they were not asked to choose a specific number of items. In Experiment 3, participants were asked to choose *some* items for each time interval, but they were not restricted to a specific number of items in each category. It may be the case that in the current study, participants tended to choose more difficult items in the test conditions because there were few restrictions placed on what they could and could not restudy. In other words, they were given less difficult learning goals. It is also possible that participants may have adopted different strategies for making choices about items that were restudied during the intervening phase versus those that were tested.

Conclusions and Future Directions

Overall, all three experiments make strong suggestions about how testing affects both memory and metacognition. First, Experiments 1 and 2 show an overall testing effect in which previously tested items were better remembered on the final memory test compared to restudied items. Another common finding across all three experiments is that the average JOL for restudied items tended to be higher than the average JOL for tested items, suggesting that participants had greater overall confidence in memory for restudied items. Although participants' average predictions suggested that they were not aware of the benefits of testing, Gamma correlations in all three experiments suggested that metacognitive monitoring on a relative item-by-item basis (e.g., resolution) was better for tested than restudied items when comparing JOLs for items recalled on the final test relative to items that were not recalled. Put another way, testing improved participants' abilities to differentiate between what they knew and what they did not know. Furthermore, Gamma correlations between JOLs and restudy choices, whether regarding the decision to restudy an item or not or deciding how much time to allocate to

an item, were stronger for tested items that initially restudied items. This finding suggests that not only did JOLs correspond better to items that would or would not be remembered later in the test condition, but participants were also better able to engage in control processes that would focus their attention on the items that merited additional study.

Another important point to note is that throughout all three experiments medium and large effects sizes were evident, suggesting that the results bear practical significance in addition to statistical significance. Small to medium effect sizes were also found for results that were not statistically significant, suggesting that trends in the data may merit further investigation in future research. For example, in both Experiments 1 and 2 there was a non-significant trend among items chosen for restudy in which greater recall occurred on the final memory test for restudied items compared to tested items. Cohen's *d* for these comparisons was .17 and .32 for Experiments 1 and 2, respectively. In Experiment 3 this trend approached significance ($p = .008$) for items chosen for 5s of restudy time, with the probability value adjusted conservatively to $p < .006$. Cohen's *d* for this comparison was .27. The trends in the data, along with their corresponding effect sizes, suggest that the restudy advantage for items chosen for restudy may be meaningful, and may well reveal important principles with regards to the affects of testing on memory and the effectiveness of metacognitive monitoring and control.

Of particular interest is the finding that the testing effect disappeared or reversed when items chosen for restudy or items chosen for the longest amount of restudy time were examined. This begs the question, "Does engaging in control processes for tested items compared to restudied items actually benefit memory?" Recall data in the current study might suggest that it does not. However, differences in the correspondence between

predictions, recall, and restudy choices (i.e., as measured by Gamma correlations) suggest that something more complex might explain why testing effects did not occur when comparing items chosen for restudy or for longer restudy times. If participants tended to choose more difficult items to restudy when they had been tested, those specific items may have required more effort in order encode them fully and protect against forgetting. It may be the case that while utilizing testing as a general study strategy is effective in enhancing overall learning, relying on monitoring of those items to make decisions for future study decisions may not be the best strategy. Furthermore, if a person does engage in restudy based on monitoring learning after testing themselves, it may not be enough to simply re-present the item for another, single study opportunity. Perhaps re-presenting the items multiple times or engaging in a more elaborative, meaning-based processing task for those items may be more effective in boosting memory performance. Engaging in a better post-monitoring strategy for those items may overcome any learning problems created by greater item difficulty.

A more basic question that should be investigated further would be *why* does testing improve the accuracy of metacognitive monitoring and control? One suggestion is that participants may use the fluency or ease with which they process restudied and tested items as a basis for predictions subsequent recall. For example, there are many anecdotal examples of students using such information while preparing for a test which leads them to erroneously believe that they know the information better than they actually know it. It is not uncommon that students will report that they read over their notes, felt like they "knew" the information and then were quite surprised when they performed poorly on a test. Several studies within the metacognitive literature have suggested that fluency plays

an important role in making judgments about future recall (e.g., Begg et al., 1989, Benjamin, Bjork, & Schwartz, 1998; Kelley & Lindsay, 1993; Koriat & Ma'ayan, 2005). Likewise, participants in the current experiments may have given higher JOLs to restudied items compared to tested items overall because they processed those items more fluently. While reliance on this type of information may lead to errors in overall judgments, it appears to improve relative judgments on an item-by-item basis. Prior research supports this notion by suggesting that participants give lower JOLs to items that are perceived to be more difficult to learn or retrieve compared to items perceived as easy to learn or retrieve (Benjamin et al., 1998; Rhodes & Castel, 2008; Serra & Dunlosky, 2005).

As a measure of retrieval fluency, Benjamin et al. (1998) recorded participants' response latencies (reaction time) in answering general knowledge questions. After answering the question, participants predicted the likelihood of recalling the answer on a later test. Results showed that participants gave the highest JOLs to items that were answered in the shortest amount of time. However, the best recall performance was found for items with the longest response latencies. Several studies within the testing effect literature have similarly shown that as the difficulty of retrieval during an intervening test increases, the advantage for later retrieval also increases (Carpenter, 2009; Carpenter & DeLosh, 2006; Glover, 1989). For example, Carpenter and DeLosh (2006) manipulated retrieval difficulty by providing participants with varying numbers of letter cues during the initial cued recall test. They reasoned that with fewer cues available (e.g., *c*_ _ _ _ rather than *cab*_ _ for the target word *cabin*), retrieval should be more difficult. Carpenter and DeLosh reported that recall on a later memory test was inversely related to the

number of cues providing on an initial test, such that recall was best when just one letter cue was provided. In a similar study conducted recently by Littrell, Rhodes, and DeLosh, participants followed a similar procedure in which they were given items that varied in level of processing difficulty during the intervening task. Specifically, pairs of words were either re-presented intact with the cue word followed by a target word (i.e., a restudy opportunity), or they were presented with the cue word followed by a blank, one letter of the target, or 3 letters of the target (i.e., tests of varying difficulty). Participants were also asked to make a JOL for each of these items. Results showed that while participants' predictions decreased as processing became more difficult, their subsequent recall increased with more difficult processing tasks. Furthermore, gamma correlations between JOLs and recall became stronger as retrieval fluency decreased. These results suggest that testing improves resolution when the act of retrieval is less fluent. In comparison with restudying an item, a task in which processing is very fluent because the item is re-presented intact, it might be suggested that testing provides a better index of later retrieval and thus enhances the accuracy of predictions of subsequent recall.

Educational Implications

Perhaps the most significant aspect of this research lies in its applicability to student learning. More and more studies are beginning to examine testing in classroom simulations, web-based courses (e.g., Butler & Roediger, 2007; McDaniel, Anderson, Derbish, & Morissette, 2007), and actual classrooms (e.g., Carpenter, 2009). Several studies have demonstrated that testing serves to improve learning and protect against forgetting. However, none of these studies have directly examined how testing affects metacognitive monitoring and control. The current experiments provide strong evidence

that testing improves relative metacognitive judgments (resolution) and influences control processes. The evidence presented here suggests that testing not only provides a direct benefit to memory, but it also shows that testing boosts monitoring and control processes. Thus, testing could indirectly benefit memory if improvements in monitoring and control lead to better follow-up study practices.

A possible limitation of this work, however, could exist in its generalizablility to learning in a classroom setting. One way to extend this work would be to examine the questions posed in the current study using materials and an experimental design that more closely resembles student learning in a classroom setting. For example, in the current set of experiments, participants studied lists of pairs of unrelated words. Although these materials are commonly used in basic research involving list-learning paradigms, they may not generalize to the learning circumstances that a student faces. It may be beneficial to test monitoring and control of tested and restudied information using materials such as passages of text, vocabulary words and their definitions, or foreign language words and their English equivalents. Additionally, another way to extend this work would be to use a between-subjects design in which participants restudy an entire set of material and take a test on a separate set of material. In the current study, a within-subjects design was used in which participants restudied and took tests on items within a single list, rather than separate lists. In educational scenarios, one might argue that students are unlikely to study some materials and test themselves on other materials, as is the case in a within-subjects design. Instead, they may adopt a strategy of either restudying their notes or testing themselves with practice quizzes or flash cards, as is better approximated in a between-subjects design. It is important to note, though, that prior findings in the testing

effect literature have been replicated consistently across various types of materials, laboratory and classroom settings, and differing experimental designs (e.g., McDaniel et al., 2007). Therefore, it would be expected that the current findings would also be replicated under the circumstances described above.

Should the present results prove to be reliable and generalize to educational settings, they would have direct implications for student learning. As an instructor, the testing effect literature suggests that having students engaging in frequent test-taking should improve learning and retention of that information above and beyond simply studying. Adding to this suggestion, the current study indicates that students may also use test-taking to improve monitoring accuracy and to inform their decisions about what information merits additional study. Therefore, the current research suggests that instructors should not only encourage students to test themselves as a means to enhance learning, but they should also inform students that test-taking can give them metacognitive feedback that will improve their assessments learning and the subsequent study strategies that they engage in to further improve learning.

REFERENCES

Atkinson, R. C. (1972). Optimizing the learning of a second-language vocabulary. *Journal of Experimental Psychology*, *96,* 124-129.

Agarwal, P. K., Karpicke, J. D., Kang, S. K., Roediger, H. L., & McDermott, K. B. (2007). Examining the testing effect with open- and closed-book tests. *Applied Cognitive Psychology,* www.interscience.wiley.com.

Bangert-Drowns, R. L., Kulik, J. A., & Kulik, C. C. (1991). Effects of frequent classroom testing. *Journal of Educational Research, 85,* 89-99.

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

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

Brown, R. & McNeill, D. (1966). The "tip of the tongue" phenomenon. *Journal of Verbal Language and Verbal Behavior, 5,* 325-337.

Butler, A. C., Karpicke, J., & Roediger, H. L. (2007). The effect of type and timing of feedback on learning from multiple-choice tests. *Journal of Experimental Psychology: Applied, 13,* 273-281.

Butler, A. C., Karpicke, J., & Roediger, H. L. (2008). Correcting a metacognitive error: Feedback increases retention of low-confidence correct responses. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34,* 918-928.

Butler, A. C., & Roediger, H. L. (2007). Testing improves long-term retention in a simulated classroom setting. *European Journal of Cognitive Psychology, 19,* 514-527.

Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: The benefits of elaborative retrieval. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 35,* 1535-1569.

Carpenter, S. K., & DeLosh, E. L. (2006). Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. *Memory & Cognition, 34*(2), 268-276.

Carpenter, S. K., Pashler, H., Cepeda, N. J. (2008). Using Tests to enhance 8th grade students' retention of U.S. History Facts. *Applied Cognitive Psychology*, www.interscience.wiley.com.

Carpenter, S. K., Pashler, H., & Vul, E. (2006). What types of learning are enhanced by a cued recall test? *Psychonomic Bulletin & Review, 13,* 826-830.

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

Cull, W. L. (2000). Untangling the benefits of multiple study opportunities and repeated testing for cued recall. *Applied Cognitive Psychology, 14,* 215-235.

Dunlosky, J., & Hertzog, C. (1998). Training programs to improve learning in later adulthood: Helping older adults educate themselves. In D. J. Hacker (Ed.), *Metacognition in educational theory and practice* (pp. 249-275). Mahwah, NJ: Erlbaum.

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

Dunlosky, J., & Thiede, K. W. (2004). Causes and constraints of the shift-to-easier-materials effect in control of study. *Memory and Cognition, 32*, 779-788.

Flavell, J. (1976). Metacognitive aspects of problem-solving. In L. Resnick (Ed.), The Nature of Intelligence. Hillsdale, NJ: Erlbaum Assoc.

Gardiner, J. M., & Richardson-Klavehn, A. (2000). Remembering and knowing. In E. Tulving & F. I. M. Craik (Eds.), *The Oxford handbook of memory* (pp. 229-244). London: Oxford University Press.

Gates, A.I. (1917). Recitation as a factor in memorizing. *Archives of Psychology, 6*(40).

Goodman, L. A. & Kruskal, W. H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association, 49,* 732-764.

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

Hart, J. T. (1965). Memory and the feeling-of-knowing experience. *Journal of Educational Psychology, 56,* 208-216.

Kang, S. H. K., McDermott, K. B., & Roediger, H.L., III. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology, 19,* 528-558.

Karpicke, J.D. & Roediger, H.L., III (2007). Repeated retrieval during learning is key to long-term retention. *Journal of Memory and Language, 57,* 151-162.

Karpicke, J. D. & Roediger, H. L., III (2008). The critical importance of retrieval for learning. *Science, 319,* 966-968.

Kelley, C. M., & Lindsay, D. S. (1993). Remembering mistaken for knowing: Ease of retrieval as a basis for confidence in answers to general knowledge questions. *Journal of Memory and Language, 32,* 1-24.

Koriat, A. (2007). Metacognition and consciousness. In P. D. Zelazo, M. Moscovitch, & E. Thompson (Eds.), *The Cambridge Handbook of Consciousness* (pp. 289-325). Cambridge, UK: Cambridge University Press.

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

Kornell, N. & Metcalfe, J. (2006). Study efficacy and the region of proximal learning framework. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 32,* 609-622.

Kornell, N., & Son, L. K. (2009). Learners' choices and beliefs about self-testing. *Memory, 17,* 493-501.

Kuo, T., & Hirshman, E. (1996). Investigations of the testing effect. *American Journal of Psychology, 109,* 451-464.

Kucera & Francis, W. N. (1967). *Computational Analysis of Present-Day American English.* Providence: Brown University Press.

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,* 464-470.

Mazzoni, G. & Cornoldi, C. (1993). Strategies in study time allocation: Why is study time sometimes not effective? *Journal of Experimental Psychology: General, 122,* 47-60.

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

Mazzoni, G. & Nelson, T. O. (1995). Judgments of learning are affected by the kind of encoding in ways that cannot be attributed to the level of recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 1263-1274.

McCabe, D. P. & Soderstrom, N. C. (in press). Recollection-based prospective metamemory judgments are more accurate than those based on confidence: Judgments of remembering and knowing (JORKs). *Journal of Experimental Psychology: General.*

McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology, 19,* 494-513.

McDaniel, M. A., & Fisher, R. P. (1991). Tests and test feedback as learning sources. *Contemporary Educational Psychology, 16,* 192-201.

McDaniel, M. A., & Masson, M. E. (1985). Altering memory representations through retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11,* 371-385.

McDaniel, M. A., Roediger, H. L., III, & McDermott, K. B. (2007). Generalizing test-enhanced learning from the laboratory to the classroom. *Psychonomic Bulletin & Review, 14,* 200-206.

Metcalfe, J., & Kornell, N. (2005). A region of proximal learning model of study time allocation. *Journal of Memory and Language, 52,* 463-477.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). The University of South Florida word association, rhyme, and word fragment norms. http://w3.usf.edu/FreeAssociation.

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

Nelson, T. O. (1996). Consciousness and metacognition. *American Psychologist, 51,* 102-116.

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

Nelson, T. O., Dunlosky, J., Graf, A., & Narens, L. (1994). Utilization of metacognitive judgments in the allocation of study time during multitrial learning. *Psychological Science, 5,* 207-213.

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In. G. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (pp. 125-173). New York: Academic Press.

Nelson, T., Narens, L., & Dunlosky, J. (2004). A Revised methodology for research on metamemory: Pre-judgment recall and monitoring (PRAM). *Psychological Methods*, *9*, 53-69.

Rhodes, M. G., & Castel, A. D. (2008). Memory predictions are influenced by perceptual information: Evidence for metacognitive illusions. *Journal of Experimental Psychology: General, 137,* 615-625.

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

Roediger, H. L., & Karpicke, J. D. (2006a). The power of testing memory: basic research and Implications for educational practice. *Perspectives on Psychological Science, 1*(3), 181-210.

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

Serra, M. J. & Dunlosky, J. (2005). Does retrieval fluency contribute to the underconfidence-with-practice effect? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 1258-1266.

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

Spitzer, H. F. (1939). Studies in retention. *Journal of Educational Psychology, 30,* 641-656.

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

Thiede, K. W., & Dunlosky, J. (1999). Toward a general model of self-regulated study: An analysis of selection of items for study and self-paced study time. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25,* 1024-1037.

Thompson, C. P., Wenger, S. K., & Bartling, C. A.(1978). How recall facilitates subsequent recall: A reappraisal. *Journal of Experimental Psychology: Human Learning, & Memory, 4,* 210-221.

Tulving, E. (1967). The effects of presentation and recall of material in free-recall learning. *Journal of Verbal Learning and Verbal Behavior, 6,* 175-184.

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,* 1527-1538.

Wilson, M. D. (1988). The MRC psycholinguistic database: Machine readable dictionary, version 2. *Behavioral Research Methods, Instruments, and Computers, 20,* 6-11.