THE DIRECT EFFECT OF MONITORING OF LEARNING ON MEMORY PERFORMANCE: HOW IS IT INFLUENCED BY RETENTION INTERVAL OR JUDGMENT INSTRUCTIONS?

by

AMBER ELIZABETH WITHERBY

Bachelor of Science, 2013
Colorado State University
Fort Collins, Colorado

Submitted to the Graduate Faculty of the
College of Science and Engineering
Texas Christian University
in partial fulfillment of the requirements
for the degree of

Master of Science

May 2016
THE DIRECT EFFECT OF MONITORING OF LEARNING ON MEMORY PERFORMANCE: HOW IS IT INFLUENCED BY RETENTION INTERVAL OR JUDGMENT INSTRUCTIONS

by

Amber Elizabeth Witherby

Thesis approved:

Major Professor

For The College of Science and Engineering
ACKNOWLEDGEMENTS

First, I would like to thank my advisor Dr. Uma Tauber. Her continual guidance and support has been invaluable. Her feedback has been instrumental in my development as a cognitive researcher. I would also like to acknowledge my committee members Dr. Naomi Ekas and Dr. Timothy Barth, both of which provided important feedback for the development of this project.

I would also like to thank members of my cohort, Randi Proffitt Leyva, Sarah Theisen, and Shannon Conrad, as well as my fiancé, Marcos Gonzalez, whose love and encouragement have helped my overcome many hurdles.

Finally, I would like to thank my parents for their unwavering support and for pushing me to reach my goals.
# TABLE OF CONTENTS

Acknowledgments.................................................................................................................. ii

List of Figures ........................................................................................................................ iv

List of Tables ......................................................................................................................... v

Introduction............................................................................................................................. 1

What is Metacognition? .......................................................................................................... 3

How Monitoring Influences Memory ................................................................................... 6

The Present Experiments ....................................................................................................... 11

Experiment 1 .......................................................................................................................... 12

Experiment 2 .......................................................................................................................... 18

Experiment 3 .......................................................................................................................... 21

General Discussion ................................................................................................................ 25

Appendix ................................................................................................................................ 29

References ............................................................................................................................. 30

Vita

Abstract
LIST OF FIGURES

1. Theoretical model of how monitoring influences learning.................................7
2. Mean proportion correctly recalled in Experiment 1.............................................16
3. Mean proportion correctly recalled in Experiment 2.............................................20
4. Mean proportion correctly recalled in Experiment 3.............................................24
LIST OF TABLES

1. Judgment of Learning Magnitude in all experiments .......................................................... 17

2. Mean gamma correlations in all experiments ..................................................................... 18
Student learning and achievement in undergraduate classes is of critical importance. With the job market becoming increasingly competitive, improving student learning is even more imperative because success in the classroom is likely to contribute to success in life. Fortunately, ample research in cognitive psychology has investigated strategies that students can implement to improve their learning, and cognitive psychologists continue to investigate this issue (for a thorough review of common learning techniques, see Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013). For example, to enhance learning Dunlosky et al. (2013) recommend that students implement self-testing (dubbed the testing effect) and distributed study.

The testing effect is the robust phenomenon in which people remember more information after testing themselves on it compared to rereading it (Karpicke & Roediger, 2008). Researchers have investigated this effect for a number of years, dating back to the early 20th century (e.g., Abbott, 1909; Gates, 1917; Thorndike, 1914). Further, it has been demonstrated using a range of materials (e.g., Allen, Mahler, & Estes, 1969; Helder & Shaughnessy, 2008; Hogan & Kintsch, 1971; Putnam & Roediger, 2013), including educationally relevant materials (e.g., Carrier & Pashler, 1992; McDaniel & Fisher, 1991; Roediger & Karpicke, 2006). For instance, Roediger and Karpicke (2006) had participants read a prose passage in an initial study session. Then, in an intermediate session participants either reread the passage (i.e., the restudy group) or recalled everything they could remember from the passage (i.e., the test group). Finally, participants took a final free recall test that took place after 5 min, 2 days, or 1 week.
Results revealed that memory performance was slightly elevated after restudy for the immediate (5 min) test, but it was greatly enhanced after testing for both the 2 day and 1 week delays. These results demonstrate that testing is more beneficial than restudying for long term retention of information, which is an academic gold standard.

Another learning technique that has shown robust benefits for memory is distributed study. As the name implies, the distributed study effect is the finding that memory performance is enhanced when study sessions are spaced out over time (i.e., distributed), rather than massed into a single session (i.e., cramming; see Carpenter, Cepeda, Rohrer, Kang, & Pashler, 2012; Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006; and Benjamin & Tullis, 2010 for recent reviews). For example, consider Bahrick’s (1979) influential paper on distributed study. During an initial phase, participants were provided with 50 English-Spanish translations and were required to study them until they were correctly recalled once on a cued recall test (on which they were given the English word and produced the Spanish equivalent). During the second phase participants restudied the translations six times, but the interval between study sessions differed. Specifically, participants had a 0 day (i.e., massed study), 1 day, or 30 day interval between study sessions. At the beginning of each session, participants took a cued recall test for the translations, and then restudied the entire list. Participants in all groups returned 30 days after the final study session to take the final cued recall test. Consistent with the testing effect, performance during the study sessions was superior for the participants who had massed study (i.e., no interval) compared to participants who had distributed study (i.e., 30 day interval). Importantly, on the final test participants’ memory performance was greatest following 30 day intervals between study sessions relative to both the 1 day and 0 day groups.
Even though the previously discussed techniques are well established, many students choose to use less effective strategies, such as rereading (Carrier, 2003; Hartwig & Dunlosky, 2012). One possibility is that students may choose less effective strategies because self-testing and distributed study can be more effortful and time consuming. For example, distributed study requires planning ahead of time in order to schedule study sessions. Likewise, testing can require the construction of testing materials such as flash cards. Fortunately, new research in metacognition has demonstrated that the act of simply making judgments about the material you are trying to learn (i.e., monitoring your learning) may be a fruitful strategy to improve memory performance (e.g., Soderstrom, Clark, Halamish, & Bjork, 2015). This finding is promising for student learning because this technique is very easy to implement while studying. Thus, the present experiments further investigated the effect of monitoring one’s learning on memory performance. Before further elaboration on this research, a brief explanation of metacognition is necessary.

**What is Metacognition?**

Metacognition can be conceptualized as thinking about your thinking (Flavell, 1979). One of the largest fields of metacognition, and the focus of the present research, is metamemory. Metamemory is commonly defined as thoughts about one’s own memory (Nelson & Narens, 1990). The key components of metamemory are metacognitive knowledge, monitoring, and control (for a review, see Dunlosky & Metcalfe, 2009). To illustrate these concepts consider the following example. A student is trying to learn a list of English-Spanish translations for class. He may have a belief that rereading all of the translations is the best way to remember them. This belief is part of his metacognitive knowledge (i.e., beliefs about how different factors influence memory). Thus, he decides
to start by rereading all of the translations. By doing so he is exerting metacognitive control, which is any act that alters the course of the study session (e.g., initiating study, changing study strategies, ending study). After rereading all of the translations he can then compare his current state of learning to his desired state of learning (via metacognitive monitoring). Doing this, he may determine that he has not learned many of the translations. Accordingly, he may decide to change his strategy and make flash cards to test himself. Thus, he uses the information obtained through monitoring to inform his study decisions (i.e., control). As he continues to study, he may decide that he has reached his desired state of learning (via monitoring), and he ends the study session (via control). Finally, perhaps he has learned that self-testing is a more effective strategy than rereading and thus, updates his metacognitive knowledge. As demonstrated in this example, metacognitive knowledge, monitoring, and control are highly interrelated and continuously rely on one another throughout a study session.

In Nelson and Narens’s (1990) seminal paper on metamemory, the authors put forward a model describing how monitoring and control work together to influence memory performance. Specifically, they separated cognitive processes into two interrelated levels: the object-level and the meta-level. The object-level can be thought of as the learning activity one is trying to complete, whereas the meta-level is one’s thoughts about that learning activity. With the use of monitoring and control these levels communicate with each other to guide study. Through monitoring, information from the object-level is used to inform the meta-level. This can be thought of as feedback about how the learning activity is going. With this information, the meta-level can then modify the object level (via control), by continuing,
changing, or ending the learning activity. Metamemory research typically focuses on how these two processes interact and how their interaction influences memory performance.

The primary way researchers evaluate participants’ monitoring of learning is by eliciting judgments about their learning. There are many different judgments that have been investigated (e.g., ease of learning [EOL] judgments, feeling of knowing [FOK] judgments, confidence judgments; see Dunlosky & Metcalfe, 2009, for a review), and a commonly used one is the judgment of learning (JOL). In a typical JOL experiment, participants study a series of items (e.g., word pairs, images, sentences) for a later memory test. After studying each item the participant is asked to make a judgment of the likelihood that they will remember that item on an upcoming memory test (usually on a scale of 0%-100%). JOLs are typically elicited immediately following the presentation of each item or following a delay (Dunlosky & Nelson, 1994; see Rhodes & Tauber, 2011 for a review). Finally, participants take a memory test. To illustrate, Koriat, Bjork, Sheffer, & Bar (2004) had participants study a list of paired associates (i.e., word pairs) that were either related (e.g., castle – king) or unrelated (e.g., book – dog). Participants were instructed to remember the word pairs such that they would be able to recall the second word of the pair (e.g., king) when given the first (e.g., castle) on a memory test. After studying each pair, participants judged the likelihood (on a scale of 0%-100%) that they would remember that word pair on a later memory test. Importantly, the final cued recall test (e.g., castle - ?) occurred immediately, 1 day, or 1 week following study. Participants’ JOLs were higher for related word pairs than unrelated word pairs; however, their JOLs did not vary based on the retention interval (RI). This suggests that participants’ JOLs were sensitive to item difficulty, although
they did not take forgetting into account (but see Experiments 2 and 3b where JOLs differed as a function of RI using a within-participants design; see also Rawson, Dunlosky, & McDonald, 2002; Tauber & Rhodes, 2012).

Theoretical perspectives on monitoring of learning originally suggested that JOLs are made based on the direct strength of the memory trace (Hart, 1965). That is, people could directly monitor the strength of a memory trace, and thus they could accurately predict whether a given item would be remembered. However, ample research has demonstrated that participants may be relying on other information, causing their judgments to be inaccurate under some circumstances (e.g., Begg, Duft, Lalonde, Melnick, & Sanvito, 1989; Benjamin, Bjork, & Schwartz, 1998; Rhodes & Castel, 2008, 2009). As an alternative, contemporary theories of monitoring of learning (e.g., the cue-utilization approach) suggest that JOLs are inferential such that participants rely on cues from the study environment when making their judgments (e.g., Koriat, 1997). That is, when studying different items participants take into account any distinguishable features about that item (i.e., cues) and use this information to predict how these cues will influence later memory performance. For example, when studying a list of words that vary in concreteness, participants take this cue into account and typically assign higher JOLs to concrete words (e.g., table) than abstract words (e.g., justice) (e.g., Begg et al., 1989; Hertzog, Dunlosky, & Kidder, 2003; Tauber & Rhodes, 2012, Witherby & Tauber, in prep).

**How Monitoring Influences Memory**

The accuracy of JOLs is important because (as previously highlighted) monitoring and control work together in a continuous feedback loop to update metacognitive knowledge to achieve a specific learning goal (Nelson & Narens, 1990). Thus, it is important that JOLs
accurately assess the current state of learning so that students will make good study decisions. As illustrated in Figure 1 (dashed lines), monitoring can influence memory performance indirectly by influencing study decisions (i.e., control; e.g., Dunlosky & Hertzog, 1997; Mitchum, Kelley, & Fox, 2016; Tauber & Rhodes, 2012; Thiede, 1999). For instance, Dunlosky and Hertzog (1997) had participants study a series of word pairs, make a JOL for each, and select half of the items for restudy. Participants’ item selections were identical to those selected by a computer algorithm, which chose the items with the lowest JOLs for restudy. Furthermore, memory performance benefited from restudying the selected items, demonstrating that monitoring can improve memory performance by informing study decisions.

*Figure 1. A theoretical model of how monitoring of learning can influence memory performance. Dashed lines represent the established indirect effect of monitoring on memory performance. The solid line represents the proposed direct effect.*
Research has also demonstrated that monitoring of learning (as measured via JOLs) may have a direct effect on memory (e.g., Begg et al., 1989; Dougherty, Scheck, Nelson, & Narens, 2005; Mitchum et. al., 2016; Putnam & Roediger, 2013; Soderstrom et al., 2015; Zechmeister & Shaughnessy, 1980; see the solid line of Figure 1). For instance, Zechmeister and Shaughnessy (1980) presented participants with a list of words, with each word presented one-at-a-time for 6 seconds. Participants made JOLs for half of the items, and were allotted 6 seconds to make their judgment. After studying all items participants completed math problems for 1 minute, and then took a final free recall test. Memory performance was significantly higher for judged items (i.e., items given JOLs) compared to not-judged items. Although promising for establishing the direct effect of monitoring of learning on memory performance, the results of this study should be interpreted with caution. In particular, study time was not equated for judged and not-judged items (a similar confound occurred in Dougherty et al., 2005). In fact, the judged items had potentially double study duration relative to the non-judged items. As a result, in these experiments any memory benefit of judged items over not-judged items could simply have been due to increased study time.

Soderstrom et al. (2015) designed their experiments to evaluate the influence of JOLs on memory while eliminating the study time confound. Specifically, they had participants study a list of related word pairs (e.g., loaf-bread) for a memory test. Half of the participants made JOLs for each pair and the other half did not. Importantly, study time was equated for judged and not-judged items. After studying all items, participants completed a 3-minute distractor task, which was followed by a cued recall test. On the test, participants were presented with the first word of each pair (e.g., loaf - ?) and were instructed to recall the second word of the pair (e.g., bread). Memory performance for the judged items was
elevated relative to the not-judged items. Moreover, the memory enhancement was approximately 10%-15%, which could translate to a letter grade improvement for students in undergraduate courses. Even so, when the same methodology was used with unrelated word pairs, there was no memorial benefit after making JOLs. The authors explained this by suggesting that the act of making JOLs for related items strengthened the cue (i.e., associative relatedness) that the test relied on (dubbed here the cue-strengthening hypothesis). For the unrelated items, there was essentially nothing to strengthen, and as such making JOLs for those items did not benefit memory performance.

Relatedly, researchers have recently evaluated how a different monitoring judgment - remember/know (R/K) judgments - can influence memory. Specifically, Naveh-Benjamin and Kilb (2012) had young and older adult participants study four lists of word pairs. For some items, participants were instructed to remember each word (i.e., item memory), for others participants were instructed to remember the word pair as a whole (i.e., associative memory). After studying each list, participants took a recognition memory test on which they were presented with either an individual word (i.e., item recognition) or a word pair (i.e., associative recognition) and were instructed to respond “old” if they had seen that item before or “new” if they had not. Additionally, half of the participants were instructed to make a R/K judgment following their old/new decision. To do so, participants were instructed to say “remember” if they remembered the specific context from which they studied the word or “know” if they knew they had seen the item, but could not remember the specific context. Participants were also allowed to answer with “guess” if they guessed on their old/new decision. The results demonstrated that the act of making a R/K judgment improved older adults’ associative memory (but not item memory). Further, the
requirement to make these judgments did not affect young adults’ memory at all. This provides further evidence that the act of making a monitoring judgment can influence memory, but in some circumstances it may not.

As mentioned previously, making monitoring judgments does not inevitably improve memory performance (e.g., memory for unrelated word pairs did not benefit from making JOLs; Soderstrom et al., 2015). In fact other research comparing judgment groups to no-judgment groups have not found a memorial benefit from making judgments (e.g., Kelemen & Weaver, 1997; Mitchum et. al., 2016; Tauber & Rhodes, 2012). In their experiment, Tauber and Rhodes (2012) had participants study a list of concrete (e.g., table) and abstract (e.g., justice) words. Participants were randomly assigned to one of three judgment groups: judgment of retention (JOR) group, JOL group, or restudy only group. In the JOR group participants were instructed to estimate how long (0-60 minutes) they would be able to remember each word. In the JOL group participants made a JOL for each item. These judgments (i.e., JOR or JOL) were made immediately after studying each item. In the restudy only group, participants did not make any monitoring judgment. After a 3-minute distractor task all participants took a free recall test. Although memory performance was superior for concrete words than abstract words, there were no differences between judgment groups. Consistent with Soderstrom et al. (2015), this finding demonstrates that the memorial benefit of making a JOL is not an inevitable outcome. Specifically, there must be a cue that can be strengthened in order for the process of making judgments to improve memory performance. Thus, it is important to determine when making monitoring judgments will benefit memory and when it will not.
Although it has been demonstrated that making JOLs can improve memory performance on an immediate test (e.g., Soderstrom et al., 2015), it is unclear whether this memorial benefit persists over longer retention intervals. If there is a memory enhancement after a longer delay, then this effect could have important educational implications. Making JOLs is quick and does not require extensive training, and as such could be easily implemented during study. Further, there are also important empirical implications because numerous experiments require participants to make JOLs without a no-judgment control group. If the act of monitoring learning (via JOLs) alters the memorial strength of a given item, then it is difficult for research without a no-judgment control group to establish a relationship between a given independent variable and memory performance if they require participants to make JOLs. In this case, making a JOL may be a confounding variable. On the other hand, if the memorial benefit of making JOLs does not persist over a longer delay, this will also be informative for updating theory, because it provides additional information on the effects of monitoring of learning on memory. Understanding when and how monitoring influences memory performance is essential for designing and interpreting research.

**The Present Experiments**

Our primary interest was to investigate how monitoring learning (via JOLs) directly influences memory. In Experiments 1 and 2 we aimed to replicate and extend the findings of Soderstrom et al. (2015). Importantly, these are the first experiments to examine this effect with a longer (2-day) retention interval (RI). Although it is intriguing that making JOLs improves memory on an immediate test, students should ideally start preparing for a test at least a few days in advance. Therefore, the memorial benefit of making JOLs is much more educationally relevant if it is observed after a 2-day delay.
Research evaluating memory performance over a delay typically observes a significant amount of forgetting relative to memory performance on immediate tests. For example, consider Koriat et al. (2004) discussed previously. After just 1 day, memory performance decreased by roughly 20%. However, researchers have demonstrated that forgetting can be slowed by using effective study methods such as self-testing (e.g., Roediger & Karpicke, 2006). Thus, if making JOLs substantially improves memory performance, participants who make JOLs should outperform those who do not even after a 2-day RI.

Additionally, in Experiment 3 we evaluated whether the instructions used to elicit JOLs influenced the magnitude of this effect. Specifically, participants were provided with instructions emphasizing the importance of making accurate JOLs and were given practice making them prior to beginning the experiment. If making JOLs enhances the memorial strength of the items, then it reasons that participants who thought more about their JOLs would show an even larger benefit due to the additional attention paid to forming their JOLs.

**Experiment 1**

The primary goals of Experiment 1 were to follow-up on the findings of Soderstrom et al. (2015) and evaluate whether the memorial benefit of making JOLs holds up over a 2-day RI. As such, participants were presented with a list of strongly related word pairs (e.g., seed – plant) to study for a cued recall test. Half of the participants made a JOL for each pair, whereas the other half made no judgments. Additionally, half of the participants took a cued recall test after a short 3-minute RI, whereas half took the test after a long 2-day RI. We expected participants in the short RI group (3-min) who made JOLs to show better memory performance than participants in
the short RI group who do not make JOLs, as this was a direct replication of Soderstrom et al. (2015). Further, this effect was also expected to maintain over the 2-day RI.

Method

Participants and Design

A 2 (judgment group: JOL vs. no-JOL) x 2 (retention interval: short vs. long) between-participants design was used. To estimate the number of participants needed in each group, a power analysis was performed, using G*Power (Faul, Erdfelder, Lang, & Buchner, 2007). Using the effect size, $d = .79$ (from Soderstrom et al., 2015), power at .90, and alpha level of .05, it was estimated that 35 participants would be required for each group. However, Soderstrom et al. (2015) did not have a 2-day RI between study and test. Therefore, we increased the estimate to 40 participants per group in order to detect the effect for our delay groups. As such, 160 ($n = 40$ in each group) undergraduate students from Texas Christian University participated for partial course credit in psychology courses.

Materials and Procedure

Pilot data were collected to estimate memory performance for participants in the long delay groups. To do so, participants ($n = 20$) studied 50 related word pairs (e.g., castle – king), presented one-at-a-time for 8 s and no JOLs were made (i.e., no-JOL group). After a 2 day RI, recall on the cued recall test (e.g., castle - ?) was relatively high ($M = .59$, $SD = .12$). Therefore, an additional 10 items were added to the study phase to prevent ceiling effects in the immediate test groups. Thus, participants studied 60 related word pairs obtained from the Nelson, McEvoy, and Schreiber (2004) free association norms. The mean associative strength for the word pairs was .33. That is, when prompted with each cue (e.g., castle - ?), 33% of participants responded with the target (i.e., king) during normative data collection.
The cues and targets did not differ in length ($M = 5.00$ letters and $M = 4.95$ letters respectively), word frequency ($\text{Log}_F\text{req}_{\text{Hal}}; M = 9.36$ and $M = 9.68$ respectively), or number of syllables ($M = 1.45$ and $M = 1.31$ respectively), $t_s \leq 1.42$.

Participants were tested in small groups (4 – 6 participants), though tasks were completed individually. Each word pair was presented for 8 s in the center of a computer screen, and item order was randomized anew for each participant. Participants were randomly assigned to one of four groups: no-JOL short RI, no-JOL long RI, JOL short RI, or JOL long RI. In both JOL groups, participants were prompted to make a JOL after 4 s of the presentation time had elapsed, and had the remaining 4 s to enter their judgment. In the no-JOL groups, participants studied the word for the entire 8 s with no JOL prompt. Importantly, this procedure equated study time for all groups (Soderstrom et al., 2015). Participants in the JOL groups were instructed to make their JOLs on a scale from 0-100% indicating the likelihood that they would be able to recall the second word of the pair (e.g., king) when presented with the first (castle - ?) on a later cued recall test. A judgment of 0% indicated that the participant would not recall the second word of the pair, whereas a judgment of 100% indicated that the participant would absolutely recall the second word of the pair. Participants in the short RI groups engaged in a 3 min distractor task during which they attempted to recall the states that make up the United States. Participants in the long RI groups returned to the lab after 2 days. Following the RI, participants in all groups took a final cued recall test, on which they were presented with the first word of each word pair (e.g., castle - ?) and asked to recall the second word (e.g., king). The test was self-paced and the order of items was randomized anew for each participant.
Results

Our primary interest was the proportion correctly recalled on the final test. As such analyses of memory performance are reported first. Although not central to the current hypotheses, for completeness analyses on monitoring accuracy (JOL magnitude and resolution) are reported next.

Memory Performance

Memory performance was scored by calculating the proportion of items participants correctly recalled on the cued recall test. Small spelling errors that did not change the meaning of the words (e.g., pluralizing a word) were counted as correct. As evident in Figure 2 (left panel), participants who made JOLs correctly recalled more word pairs than did participants who did not make JOLs after a short RI, replicating Soderstrom et al. (2015). However, the requirement to make JOLs did not influence memory performance after a 2 day RI. These observations were confirmed by a 2 (judgment group: no-JOL vs. JOL) x 2 (retention interval: short vs. long) between-participants analysis of variance (ANOVA). Specifically, there was a main effect of judgment group, $F(1, 156) = 8.57, p = .004, \eta_p^2 = .05$, demonstrating that participants in the JOL groups recalled a higher proportion of words than did participants in the no-JOL groups. Further, there was a main effect of RI, $F(1, 156) = 52.71, p < .001, \eta_p^2 = .25$, demonstrating that recall was greater following a short RI relative to a long RI. These effect were qualified by a significant two-way interaction, $F(1, 156) = 4.26, p = .041, \eta_p^2 = .03$. Follow-up tests using Tukey’s LSD revealed that for the short RI group, recall was greater for participants who made JOLs relative to those who did not, $p = .001, d = .81$. However, after 2 days memory performance was equivalent for participants who did and did not make JOLs, $p = .54$. 
These results may be taken as evidence that the act of making JOLs does not improve memory performance after a long RI. However, upon inspection of the data, we found that roughly half of the participants in the JOL groups were not consistently making JOLs. In fact, 42.5% of participants in the JOL groups failed to make JOLs for 10 or more items. Thus, it is possible that the lack of a memory benefit for the JOL group over the no-JOL group after 2 days was due to this artifact. To explore this, participants who were missing 10 or more JOLs were removed from the JOL groups and analyses were conducted on the remaining sample (126 participants: short RI JOL \(n = 22\) and long RI JOL \(n = 24\)). The observed effects maintained, except the interaction was no longer significant, \(p = .17\) (see right panel of Figure 2). Specifically, recall was greater in the JOL groups relative to the no JOL groups, \(F(1, 122) = 8.02, p = .005, \eta_p^2 = .06\), and after a short RI relative to a long RI,
Although there was no interaction, the effect was still stronger in the short RI group relative to the long RI group (\(d_s\) of .69 and .22 respectively).

**Magnitude of Judgments of Learning**

When analyzing JOL magnitude, all participants who made at least one JOL were included. As expected, JOLs did not differ by RI, \(t(62) = 1.18, p = .25\) (see Table 1). This replicates previous research demonstrating that JOLs do not differ when RI is manipulated between-participant (e.g., Koriat et al., 2004).

**Table 1**

*Mean Judgments of Learning*

<table>
<thead>
<tr>
<th></th>
<th>3-min RI</th>
<th>2-day RI</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Experiment 1</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>JOL group</td>
<td>68.06 (3.11)</td>
<td>62.67 (2.75)</td>
</tr>
<tr>
<td><strong>Experiment 2</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>JOL group</td>
<td>70.45 (1.85)</td>
<td>66.20 (2.67)</td>
</tr>
<tr>
<td><strong>Experiment 3</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>JOL group</td>
<td>66.10 (2.92)</td>
<td>66.77 (2.98)</td>
</tr>
<tr>
<td>JOL-emphasis group</td>
<td>60.53 (2.95)</td>
<td>61.45 (2.77)</td>
</tr>
</tbody>
</table>

*Note.* Standard errors of the mean are in parentheses. RI = retention interval. JOL = Judgment of Learning. The JOL-emphasis group was only included in Experiment 3.

**Resolution**

Resolution provides information about how accurate participants are in predicting which items they will and will not remember. To measure resolution, a Goodman-Kruskal gamma correlation (Nelson, 1984) was computed between JOLs and memory performance for each participant in the JOL groups. The correlations were then averaged across
participants for the JOL short RI group and JOL long RI group separately. A one-sample \( t \)-test comparing participants’ average gamma correlations to zero (i.e., chance performance), revealed that correlations were significant greater than chance, \( t(60) = 3.35, p = .001, d = .43. \)

Further, an independent samples \( t \)-test comparing gamma correlations between groups revealed that participants in the short RI group showed better resolution than did participants in the long RI group, \( t(59) = 2.89, p = .005, d = .70 \) (see Table 2). The positive correlations indicate that participants gave high JOLs to items they remembered and low JOLs to items they forgot. Thus, participants in the short RI group were better than participants in the long RI group at determining which items they would and would not remember.

Table 2

Mean Gamma Correlations

<table>
<thead>
<tr>
<th></th>
<th>3-min RI</th>
<th>2-day RI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experiment 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>JOL group</td>
<td>.33 (.07)</td>
<td>.04 (.05)</td>
</tr>
<tr>
<td>Experiment 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>JOL group</td>
<td>.20 (.06)</td>
<td>.20 (.04)</td>
</tr>
<tr>
<td>Experiment 3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>JOL group</td>
<td>.27 (.04)</td>
<td>.27 (.03)</td>
</tr>
<tr>
<td>JOL-emphasis group</td>
<td>.26 (.06)</td>
<td>.26 (.04)</td>
</tr>
</tbody>
</table>

Note. Standard errors of the mean are in parentheses. RI = retention interval. JOL = Judgment of Learning. The JOL-emphasis group was only included in Experiment 3.

Experiment 2

Experiment 1 replicated Soderstrom et al. (2015) such that JOLs enhanced memory performance on an immediate test when compared to a group that did not make JOLs. As important, this effect was observed in the 2-day RI groups; however, it was only observed
with 21% of the overall sample excluded due to missing JOL data. It seems likely that this high rate of missing data may be due to participants learning that they could do nothing (i.e., not enter JOLs) and the program would still progress. Indeed, even after excluding participants, a majority of the remaining participants were still missing 3 or more JOLs. Excluding participants from analyses post-hoc is not ideal because the obtained effects may be only observed in the subset of the sample and as such may not represent the population. As a result, Experiment 2 served as a critical replication of Experiment 1, while implementing methodology to reduce missing JOL data. Specifically, participants reported their JOLs verbally to a research assistant. This procedure ensured that participants attended to each item and made a JOL for each.

**Participants and Design**

Following the same power analysis from Experiment 1, 160 Texas Christian University undergraduate students participated for partial course credit. A 2 (judgment group: JOL vs. no-JOL) x 2 (retention interval: short vs. long) between-participants design was used. As in Experiment 1, participants were randomly assigned to one of four groups: no-JOL short RI, no-JOL long RI, JOL short RI, or JOL long RI \( (n = 40 \text{ in each group}) \).

**Materials and Procedure**

The materials and procedure were identical to Experiment 1 with the following exceptions. Participants were run individually and verbally reported their JOLs to a research assistant. The research assistant entered each JOL into the computer program.

**Results**

As with Experiment 1, results of memory performance are presented first, followed by analyses of JOL magnitude and resolution respectively.
Memory Performance

As evident in Figure 3, the requirement to make JOLs improved memory performance on the cued recall test after both a 3-min RI and a 2-day RI, $F(1, 156) = 19.34$, $p < .001$, $\eta^2_p = .11$. Additionally, there was a main effect of RI, $F(1, 156) = 97.88$, $p < .001$, $\eta^2_p = .39$, because recall was greater following the short RI relative to the long RI. The interaction was not significant, $F < 1$. Thus, making JOLs enhanced memory performance irrespective of the RI.

Figure 3. Mean proportion correctly recalled in Experiment 2. RI = retention interval. JOL = judgment of learning. The short RI consisted of a 3-minute distractor task whereas the long RI consisted of a 2 day delay between study and test. Error bars represent one standard error of the mean.
Magnitude of Judgments of Learning

As in Experiment 1, JOLs did not differ based on RI, $t(78) = 1.32, p = .19$ (see Table 1).

Resolution

A one-sample $t$-test comparing all participants’ gamma correlations to chance performance was significant, $t(75) = 6.27, p < .001, d = .72$, demonstrating that participants were significantly better than chance at predicting what they would and would not remember. Gamma correlations were positive and did not differ between groups, $t < 1$ (see Table 2). Thus, both groups did an equivalent job at predicting which items would and would not be recalled.

Experiment 3

In Experiments 1 and 2, making JOLs (compared to not making JOLs) enhanced memory performance not only after a short 3-min RI, but also after a longer 2-day RI. In fact, making JOLs improved memory performance on the final test by roughly 10%. Experiment 3 was designed to replicate and extend these outcomes. Specifically, we evaluated whether including additional instructions emphasizing the importance of constructing accurate JOLs and giving participants practice making JOLs would increase the magnitude of the JOL effect on memory over and above the benefit elicited by the standard JOL instructions. If the memorial benefit of making JOLs is due to enhanced attention to the cues used to make JOLs (Soderstrom et al., 2015), then placing emphasis on them should further increase attention to these cues. Put another way, because JOLs enhance memory then it follows that instructing participants to focus on their judgments should lead to an even larger memorial benefit. To that end, the methodology in Experiment 3 was the same as Experiment 2, with the addition of JOL-emphasis groups. In these groups, participants received additional instructions
emphasizing the importance of constructing accurate JOLs, as well as practice constructing JOLs prior to the experiment. Of most interest, participants in the JOL-emphasis groups were expected to show better memory performance than participants in the JOL groups, who were expected to show better memory performance than participants in the no-JOL groups on the final test. Further, these effects were expected to hold over the 2-day RI.

Method

Participants and Design

A 3 (judgment group: no-JOL, JOL, JOL-emphasis) x 2 (retention interval: short vs. long) between-participants design was used. Based on the power analysis from Experiment 1 that estimated the need for 40 participants per group, 235 undergraduate students from Texas Christian University participated for partial course credit in psychology courses. Participants were randomly assigned to one of six groups: no-JOL short RI (n = 40), no-JOL long RI (n = 40), JOL short RI (n = 40), JOL long RI (n = 38), JOL-emphasis short RI (n = 40), or JOL-emphasis long RI (n = 37).

Materials and Procedure

The materials used in Experiment 3 were the same as those used in the previous experiments. The procedure for the no-JOL and JOL groups was identical to Experiment 2. The JOL-emphasis groups were treated similar to the JOL groups, except that participants were given additional instructions emphasizing that they should make their JOLs as accurate as possible (see Appendix for a comparison of the JOL group and JOL-emphasis group instructions). Additionally, participants in the JOL-emphasis groups were given practice constructing JOLs for 2 items prior to the beginning the experiment. Participants were told that the practice items were to get them used to the procedure, and
that they would not need to remember them for the final test. The practice items were presented exactly as the study items were in the experiment. After completing the study phase participants in all groups had a RI of either 3 min or 2 days. Following the RI all participants engaged in a cued-recall test as in Experiment 2.

**Results**

Memory performance is discussed first, followed by JOL magnitude and resolution.

**Memory Performance**

As evident from Figure 4, participants in the JOL groups recalled a greater proportion of items relative to participants in the no-JOL groups, which was evident after a short or long RI. Memory performance for the JOL and JOL-emphasis groups was roughly equivalent. These observations were confirmed by a 3 (judgment group: no-JOL, JOL, vs. JOL-emphasis) x 2 (retention interval: short vs. long) between-participants ANOVA. Specifically, the results revealed a main effect of judgment group, \( F(2, 229) = 11.03, p < .001, \eta^2_p = .09 \). Follow-up analyses using Tukey’s LSD revealed that participants in the JOL group recalled more items than did participants in the no-JOL group, \( p < .001, d = .47 \). Likewise, participants in the JOL-emphasis group recalled more items than did participants in the no-JOL group, \( p < .001, d = .56 \). The JOL and JOL-emphasis groups did not differ, \( p = .48 \).

There was also a main effect for RI, \( F(1, 229) = 101.73, p < .001, \eta^2_p = .31 \), demonstrating that participants recalled more items following a 3-min RI relative to a 2-day RI. The interaction was not significant, \( F < 1 \).
Figure 4. Mean proportion correctly recalled in Experiment 3. RI = retention interval. JOL = judgment of learning. The short RI consisted of a 3-minute distractor task whereas the long RI consisted of a 2 day delay between study and test. Error bars represent one standard error of the mean.

Magnitude of Judgments of Learning

As in the previous experiments, participants’ JOLs did not differ based on RI. Participants in the JOL groups provided higher JOLs than did participants in the JOL-emphasis groups (see Table 1). These results were supported by a 2 (judgment group: JOL vs. JOL-emphasis) x 2 (RI: short vs. long) between-participants ANOVA. Specifically, there was a main effect of judgment group, $F(1, 151) = 4.96, p = .027, \eta^2_p = .03$. The main effect of RI and the interaction were not significant, $F$s < 1.

Resolution

A one-sample t-test comparing participants’ gamma correlations to chance performance was significant, $t(151) = 13.88, p < .001, d = 1.13$, demonstrating that participants were significantly better than chance at predicting which items they would and would not remember. Gamma correlations were positive and did not differ by judgment
A 2 (judgment group: JOL vs. JOL-emphasis) x 2 (RI: short vs. long) between-participants ANOVA was conducted and neither the main effects of judgment group or RI nor the interaction were significant, $F_s < 1$.

**General Discussion**

The present research was the first to investigate whether making JOLs improves memory performance over an extended retention interval (RI). In three experiments memory performance on a test that occurred 2 days after study was superior in groups that made JOLs relative to those that did not. In Experiment 1, the JOL effect on memory was attenuated for the 2-day RI groups relative to the other experiments, which was likely a result of participants failing to follow the JOL instructions. This hypothesis was supported in Experiments 2 and 3, where a larger memory effect was observed between the JOL and no-JOL groups after the 2-day RI when participants verbally reported their JOLs to a research assistant. The JOL effect on memory was also observed when comparing the JOL-emphasis groups to the no-JOL groups in Experiment 3. Additionally, having participants make JOLs resulted in better memory performance compared to participants who did not make JOLs in our 3-min RI groups (cf. Begg et al., 1989; Dougherty et. al., 2005; Soderstrom et al., 2015; Zechmeister & Shaughnessy, 1980). Thus, the memorial benefit of making JOLs is a robust phenomenon that was replicated in multiple experiments and that maintained with different sets of instructions regarding how to make JOLs.

The reported experiments provide support for the cue-strengthening hypothesis (Soderstrom et al., 2015). Specifically, the act of making a JOL potentially strengthened the cue that the test relied on (i.e., relatedness) leading to enhanced memory for participants who made JOLs relative to those who did not. Even so, the act of making JOLs does not always
improve memory performance (e.g., unrelated items; Soderstrom et al., 2015). For example, making JOLs does not benefit memory when learning single words without a salient cue to help predict future memory (e.g., Tauber & Rhodes, 2012). As such, researchers should continue to investigate what types of information benefit from making JOLs (and other monitoring judgments). For instance, making JOLs may benefit memory when learning English – Spanish word translations. These items are related and provide a cue for JOLs to rely on (Koriat, 2007), and as such making JOLs may strengthen memory for these items.

Importantly, the JOL groups (including the JOL emphasis groups in Experiment 3) outperformed the no-JOL groups on the final memory test by roughly 10%. Such a substantial memory benefit could have important educational implications. Specifically, this could translate to a whole letter-grade increase if it extends to students learning material for classes. To more closely examine this issue an important direction for future research will be to evaluate the influence of making JOLs on memory for educationally relevant materials (e.g., key-term definitions, prose passages, foreign language translations) and in educational contexts (e.g., when studying information for a test). If making JOLs enhances memory for these materials and in these contexts, this could be a technique that students can adopt when studying for classes. Making JOLs does not take an extensive amount of time, and it does not require elaborate instructions or practice for people to benefit from their effects. Indeed, contrary to our hypotheses, additional instructions about how to make JOLs and practice making them led to equivalent memory performance for participants without these additions (Experiment 3). Thus, this technique has the potential to be easily implemented and provide substantial learning gains.
It is noteworthy that emphasizing the importance of making accurate JOLs and giving participants practice constructing them (i.e., the JOL-emphasis group in Experiment 3) did not enhance memory performance relative to participants who had standard JOL instructions and no practice. Though challenging to interpret null effects, there are a few possibilities why this may have occurred. First, memory performance was relatively high in the JOL group, especially after the 3-min RI ($M_{\text{recall}} = 84\%$). Thus, there may have been a ceiling effect on the percent correctly recalled, which would have constrained our ability to detect enhanced memory in the JOL-emphasis groups. However, in the 2-day RI groups there was room for improvement ($M_{\text{recall}} = 67\%$), so it is unlikely that ceiling effects contributed to the null effect between the JOL and JOL-emphasis groups after the longer RI. Another possibility is that the short amount of time participants had to make JOLs (i.e., 4 seconds) was not long enough to allow them to thoroughly evaluate the likelihood of recalling an item on the future test (above what they were already doing in the JOL group). For example, the instructions for the JOL-emphasis group encouraged participants to think about each word pair and try to come up with any reasons they may be more or less likely to remember it. It is possible that 4 seconds was not enough time for participants to make personal connections to each item. Researchers should investigate whether the time constraint attenuated a potential effect of the JOL-emphasis manipulations by allowing participants more time to construct their JOLs. Doing so may lead to an even larger increase in performance on a final test.

In all experiments, participants’ JOLs (including the JOL-emphasis group in Experiment 3) did not differ between the 3-min and 2-day RIs. This is consistent with previous research (e.g., Koriat et al., 2004) demonstrating that JOLs are generally equivalent when RI is manipulated between-participants. This is also consistent with
Koriat’s (1997) cue-utilization approach which suggests that JOLs are inferential. When participants are asked to make JOLs, and they are unaware of any differences between them and another participant (e.g., differing RIs), they have nothing to compare their judgment to and as such are unable to discriminate between different levels of the variable (e.g., short vs. long RI). Additionally, with the exception of Experiment 1, monitoring resolution did not differ between retention interval groups (or judgment group in Experiment 3). It is unclear why participants in Experiment 1 were more accurate in the short RI group relative to the long RI group. One possibility is that participants in the long RI group were not paying much attention to their JOLs, and were making them inconsistently, which resulted in lower accuracy. Moreover, gamma correlations are a somewhat noisy measure, which may have resulted in the discrepancy in Experiment 1. Even so, monitoring resolution was relatively equivalent between RI groups.

To conclude, for decades researchers have elicited monitoring judgments (e.g., JOLs) from their participants often without considering the effect that these judgments may have on memory. In the current experiments, we replicated the recent finding that JOLs improve memory performance after a short RI. More importantly, these were the first experiments to evaluate this effect following a long, 2-day RI. In three experiments participants recalled more items after a short and long RI when making JOLs relative to participants who did not make JOLs. Additionally, people do not need extensive practice to benefit from making JOLs. Thus, this is a relatively simple technique that has both important empirical and educational implications.
APPENDIX

**JOL Group Instructions**

After you study a word pair for 4 seconds, we would like you to make a judgment of how likely you are to remember the second word of the pair when given the first. To do so, a box will appear on the screen while the word pair is still visible, and you will make your judgment anticipating a memory test that will occur in about 10 minutes (2 days). You will make your judgment on a scale from 0% to 100%. A judgment of 0% will indicate that you are certain you will not recall the second word of the pair, whereas a judgment of 100% will indicate that you will absolutely recall the second word of the pair. You will have 4 seconds to make your judgment, and the word pair will remain on the screen until that 4 seconds has elapsed. You may use any whole number between 0 - 100. Please use the entire range of the scale. Once you have verbally reported your judgment continue to study the word pair.

**JOL-emphasis Group Instructions**

After you study a word pair for 4 seconds, we would like you to make a judgment of how likely you are to remember the second word of the pair when given the first. To do so, a box will appear on the screen while the word pair is still visible, and you will make your judgment anticipating a memory test that will occur in about 10 minutes (2 days). You will make your judgment on a scale from 0% to 100%. A judgment of 0% will indicate that you are certain you will not recall the second word of the pair, whereas a judgment of 100% will indicate that you will absolutely recall the second word of the pair.

It is important that you make your judgment as accurate as possible. When thinking about your judgment, take into consideration anything about the word-pair that will make you more or less likely to recall it. For example, if you see the pair “circus – clown,” you may have a very distinct memory of a time you saw a clown at a circus and thus may give is a high judgment. On the other hand, if you see the pair “chalk – board,” there may be nothing that jumps out to you as being very memorable. Therefore, you might give this pair a low judgment. **The key when you make your judgments is to be sure to thoroughly evaluate the likelihood that you will recall the second word of each pair after 10 minutes (2 days).** You will have 4 seconds to make your judgment, and the word pair will remain on the screen until that 4 seconds has elapsed. You may use any whole number between 0 - 100. Please use the entire range of the scale.


VITA

Amber Elizabeth Witherby was born December 20, 1990, in Denver, Colorado. She is the daughter of Barry Witherby and Laurie Winslow. She graduated as Valedictorian from Fort Lupton High School in 2009, and earned a Bachelor of Science from Colorado State University, graduating Cum Laude and with departmental honors in 2013.

In August 2014, she enrolled in the Experimental Psychology graduate program at Texas Christian University. While working on her Masters of Science Degree she served as a graduate teaching assistant for the university.
ABSTRACT

THE DIRECT EFFECT OF MONITORING OF LEARNING ON MEMORY PERFORMANCE: HOW IS IT INFLUENCED BY RETENTION INTERVAL OR JUDGMENT INSTRUCTIONS?

By Amber Elizabeth Witherby, B.S., 2013
Department of Psychology
Texas Christian University

Thesis Advisor: Dr. Uma K. Tauber, Assistant Professor of Psychology

New research has demonstrated that making Judgments of Learning (JOLs) can improve memory performance (e.g., Soderstrom et al., 2015). In the present research we further investigated this effect. In three experiments, participants studied a list of related word pairs. Half the participants made a JOL for each pair, whereas the other half did not. After studying all pairs, participants took a cued-recall test. Importantly, the retention interval between study and test was manipulated (either 3-min or 2-days). In Experiment 3 we included a JOL-emphasis group in which participants received instructions emphasizing the importance of making accurate JOLs and practice constructing them. Results replicated those of Soderstrom et al. (2015) such that participants in the JOL condition recalled more word pairs relative to participants in the no-JOL condition. Moreover, this outcome was extended to a long retention interval. Instructions used to elicit JOLs did not increase the magnitude of the effect.