DISSERTATION

FRAMING METAMEMORY JUDGMENTS: JUDGMENTS OF RETENTION INTERVAL (JORIs)

Submitted by

Sarah K. Tauber

Department of Psychology

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, CO

Summer 2010

COLORADO STATE UNIVERSITY

May 17, 2010

WE HEREBY RECOMMEND THAT THE DISSERTATION PREPARED UNDER OUR SUPERVISION BY SARAH K. TAUBER ENTITLED FRAMING METAMEMOY JUDGMENTS: JUDGMENTS OF RETENTION INTERVAL (JORIs) BE ACCEPTED AS FULFILLING IN PART REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY.

Committee on Graduate Work

David McCabe

Kurt Kraiger

Dawn Rickey

Advisor: Matthew Rhodes

Department Chair: Ernest Chavez

ABSTRACT OF DISSERTATION

FRAMING METAMEMORY JUDGMENTS: JUDGMENTS OF RETENTION INTERVAL (JORIs)

Prior research has shown that participants' predictions of memory performance are not sensitive to the time between study and test. However, this work has largely relied in one metacognitive measure, Judgments of Learning (JOLs), to assess such awareness. Thus, in three experiments I explored a new metacognitive measure, Judgments of Retention Interval (JORIs), in which participants determine how long (in minutes) information will be remembered. Results demonstrated that the metacognitive measure itself influences assessments of monitoring and control. For example participants chose to restudy more items when JORIs were made, compared with fewer restudy choices from participants who made JOLs (Experiment 2). However, participants demonstrated difficulty incorporating information about a retention interval into their judgments regardless of the type of judgment made (i.e., JOLs or JORIs). Results are considered within existing theoretical frameworks. I suggest that the metacognitive measure needs to be considered in order to accurately assess metacognitive awareness, and additional work is needed to assess metacognitive awareness of RI.

Sarah Katherine Tauber
Department of Psychology
Colorado State University
Fort Collins, CO 80523
Summer 2010

TABLE OF CONTENTS

Framing Metamemory Judgments: Judgments of Retention Interval (JORIs)

A student studying for a pending exam must make several decisions about the best way to prepare, such as when to start and stop studying. The student's awareness of how much information is already learned and how much is yet to be learned is critical in making optimal study choices. Students frequently question when the next exam will be administered, suggesting that they are attempting to determine when to begin preparing for the next exam. Such an observation invokes a number of questions. Do students account for the interval between the presentation of material and the test to determine how and when to study? More generally, do people make accurate memory predictions based on an anticipated retention interval (RI)? In the experiments that follow, I examined this question as well as the possibility that current metamemory measures do not fully capture all of the information a rememberer might consider when making predictions.

Metacognition

Metacognition generally refers to awareness of one's own cognitive processes. More specifically, metamemory falls under the umbrella of metacognition, and includes control and monitoring processes associated with awareness of one's own memory (Koriat, 2007; Metcalfe, 2000; Nelson & Narens, 1994). Nelson and Narens (1994) proposed a framework for metacognition which distinguishes between monitoring and control. For example, a student preparing for an upcoming exam will be better prepared if she is able to accurately assess (i.e., monitor) what information has been learned.

Metacognitive monitoring then informs the self-regulation of learning (i.e., control) such as the decision to stop studying because a student has determined the content to be understood (e.g., Metcalfe & Kornell, 2005; Nelson, 1996, Nelson & Leonesio; 1988; Thiede & Dunlosky, 1999).

Metacognitive control can thus be thought of as applying metacognitive awareness to regulate and change behavior. Therefore, metacognitive monitoring and control processes have been posited to have a direct causal role in human behavior (but see Koriat, Ma'ayan, & Nussinson, 2006, for an alternative perspective). To the degree that monitoring is poor, corresponding control processes will not operate optimally. Monitoring has been examined with a number of measures, which I review in the next section.

Metamemory Measurement

Several measures have been devised to assess metamemory accuracy (Nelson & Narens, 1994) and these can be classified as prospective or retrospective in nature (Busey, Tunnicliff, Loftus, & Loftus, 2000; Dougherty, Scheck, Nelson, & Narens, 2005). Prospective judgments are predictions of future performance, typically made during learning, whereas retrospective judgments are assessments of prior performance made during retrieval (e.g., judgments of confidence in the accuracy of a memory). The most common prospective measure is the item-by-item Judgment of Learning (JOL) in which participants make a scale prediction indicating how likely he/she is to later remember an item (e.g., I am 80% likely to remember this word)[1].

[1] Judgments of Learning (JOLs) are not the sole measure of metamemory accuracy. Nelson and Narens (1994) offer a widely used organization of metacognitive

measures involving classification into four phases of learning: before acquisition, during acquisition, during retrieval, and after retrieval. Additional metacognitive measures include: ease-of-learning (EOL), self-paced study, study termination, feeling-of-knowing (FOK), restudy choice, and confidence.

These predictions are then compared with later memory performance allowing a direct assessment of the correspondence between memory predictions and memory performance (i.e., *calibration*) and an assessment of the degree to which JOLs distinguish between what is and what is not remembered (i.e., *resolution*).

JOLs are often accurate, yet a number of discrepancies between predictions and actual memory performance have been demonstrated (e.g., Carroll, Nelson, & Kirwan, 1997; Koriat & Bjork, 2005; Koriat, Bjork, Sheffer, & Bar, 2004; Rhodes & Castel, 2008; Tauber & Rhodes, 2010; Tauber & Rhodes, in press). People often base memory predictions on information or cues that are salient during study but that are not necessarily diagnostic of later memory performance (e.g., font size). For example, Rhodes and Castel (2008) reported that participants provided higher JOLs for words presented in a large font compared with a smaller font, indicating that participants expected to better remember words studied in a large font. However, results showed no difference in recall as a function of font size. Thus, participants made predictions using the accessible cue of font size at study, which was unrelated to memory performance.

While factors relating to study or test conditions that influence JOLs have been examined, few studies have considered the metacognitive measure (i.e., the use of JOLs) itself. Accordingly, a primary goal of this dissertation is to examine JOLs in comparison to a new measure of metacognition, JORIs (judgments of retention interval). Little prior work has examined whether the particular measure used has an impact on metamemory

judgments (but see Finn, 2008). Such issues are important as it is essential to determine whether metacognitive measures are capturing metacognitive awareness or reflect artifacts produced by the specific scale used. Other areas of psychology, such as judgment and decision making (JDM), have investigated question framing to address related issues.

Framing Effects

Data from the decision making literature suggests that the framing of a question can have a substantial impact on decisions (see Schwarz, 1999, for a review). For example, people are more likely to decline a choice if it is framed in terms of failure (e.g., if you go through with this surgery there is a 15% chance you could die) than in terms of success (e.g., if you go through with this surgery there is a 85% chance you will survive) (e.g., McNeil, Pauker, Saks, & Tversky, 1982; Tversky & Kahneman, 1981). Additionally, the framing of a judgment may change the information that participants attend to when making a decision. For example, Shafir (1993) had participants make a decision framed in terms of acceptance or rejection. Specifically, participants were asked to render a verdict in a simulated child custody case based on whether a parent should be *awarded* or *denied* custody. Participants were presented with a choice between parent A, an average parent, and parent B, who had both outstanding positive qualities (e.g., above average income, close relationship with the child) and salient negative qualities (e.g., lots of work related travel). Shafir demonstrated that framing questions in terms of acceptance or rejection directly influenced later decisions. Specifically, when asked which parent to *award* custody, participants were sensitive to the positive features of a decision and were likely to award custody to parent B. However, when deciding which parent to *reject*

participants were sensitive to negative features and were more likely to reject parent B. Thus, the framing of a question influences the information decision makers attend to.

Could metamemory judgments be similarly influenced by framing? Finn (2008) explored this question by asking participants to provide memory predictions for a list of words. For each word participants judged either how likely they were to *remember* each word or how likely they were to *forget* each word. Participants in both conditions also selected items for restudy prior to a memory test. Overall, there was a stronger correspondence between predictions and recall (i.e., participants were better calibrated) when making memory predictions of how likely each item was to be *forgotten*, compared with making predictions of how likely each item was to be *remembered.* That is, predictions made in the forget condition were more conservative than predictions made in the remember condition. Further, participants chose to restudy more items under the forget context compared with the remember context, demonstrating the influence of framing on later study decisions. Finn (2008) suggested that the forget frame made memory failure more salient compared with the remember frame, which reduced predictions of later recall and also influenced control choices. However, to date, this is the only study to explore framing effects in metacognitive judgments. Framing effects might also be relevant to work examining participants' awareness of the impact of a retention interval (RIs) on memory performance.

Metamemory & Retention Interval (RI)

Koriat et al. (2004) examined the degree to which JOLs are sensitive to a retention interval (RI). They presented participants with word pairs that were either related (e.g., CAT-KITTEN) or unrelated (e.g., TABLE-MONKEY) and solicited JOLs

based on one of three RIs (i.e., immediate recall, recall in 1 day, or recall in 1 week). Participants were then given a cued-recall test following the RI that was specified during the study phase. Results demonstrated that JOLs and recall were higher for related compared with unrelated pairs. However, while recall was negatively related to the RI, JOLs remained constant. That is, participants' predictions were sensitive to item-relatedness but entirely insensitive to RI. Koriat et al. (2004) extended this finding to a 1 year RI (Experiment 4C) reporting that when participants were asked to make predictions for either immediate recall or recall in 1 year, estimates for the 1 year RI did not differ from predictions for immediate recall. Thus, even when given an extreme RI, predictions did not reflect the decline in memory performance that would be expected with a 1 year RI.

Carroll et al. (1997) similarly asked participants to make predictions for one of two RIs: 2 weeks or 6 weeks. Memory predictions were not reliably lower for the 6-week RI compared with the 2-week RI while, as expected, recall performance was significantly lower at the 6-week RI. In conjunction with Koriat et al.'s (2004) observations, people do not appear to be sensitive to the impact of RI on memory performance. That is, people provide predictions suggesting that they will remember about the same amount of information after 2 days, 1 week, or even 1 year. However, do these data reflect the method of judgment (i.e., JOL) or a metacognitive deficit demonstrating unawareness of the impact of a long RI?

Overview of Current Research

These experiments were intended to address two specific questions. First, are people aware of the negative influence an intervening RI can have on memory

performance (Experiment 1, Experiment 3)? Prior work (Carroll, et al., 1997; Koriat et al., 2004) has indicated that participants are entirely unaware of the impact of a RI on memory performance. However, it may be that a different metacognitive measure is necessary for participants to be able to demonstrate sensitivity to RI. JOLs are typically made using a percentage scale such that average predictions and average performance can be directly compared. However, some cues might be sufficiently difficult to attend to that the scale itself may need to be modified to make the cue more salient. Thus, the second primary question was what impact does the metacognitive judgment itself have on monitoring and control processes (Pilot Study 1, Pilot Study 2, and Experiment 2)? Perhaps participants are better able incorporate time (e.g., the interval between study and test) into their predictions when predictions are made with a different judgment. That is, participants may be better able to anticipate future memory performance when predictions are made in units of time. I examined this issue by asking participants to make judgments of retention interval (JORIs) indicating how long, in minutes, they expected to be able to remember information. I anticipated that participants might be better able to anticipate future memory performance if asked to make predictions in terms of minutes compared with percentages. Additionally, the type of judgment made during study might produce different study choices. I examined this issue by asking participants to either make JORIs or JOLs followed by a choice to restudy words prior to test. I anticipated that participants who made JORIs would make better study choices than participants who made JOLs.

Pilot Data: Overview

In order to begin this line of inquiry two sets of pilot data were collected. Pilot Study 1 explored participants' metamemory awareness when making judgments of retention interval (JORIs). Specifically, participants were provided with a continuous minute scale (0 – 60 minutes) and asked to indicate how long they would remember individual items. In Pilot Study 2 participants made JORIs on a continuous minute scale (0 – 60 minutes) and for words which differed in the ease with which they could be recalled.

Pilot Study 1

Participants in Pilot Study 1 studied a list of words that they were to remember for an upcoming test and made JORIs for each word on a continuous minute scale (0 – 60 minutes). Based on the framing literature, it was expected that framing the metamemory scale in this way would lead participants to provide lower JORIs than would be anticipated from the existing JOL literature (e.g., Finn, 2008; Shafir, 1993; Tversky & Kahneman, 1981). This would contrast sharply with prior demonstrations that participants expect to remember words for days, months, or even a year when making a JOL (Carroll et al., 1997; Koriat et al., 2004). Thus, Pilot Study 1 was expected to demonstrate that participants would provide much more realistic memory predictions when JORIs were measured.

Method

Participants

Forty individuals from Colorado State University participated (*M* age = 19.45, *SD* = 1.30) and received extra credit in a psychology course.

Materials & Procedure

After providing informed consent, participants were presented with 30 nouns equated on word frequency (*M* = 37.54, *SD* = 12.24: MRC Psycholinguist Database, 1987) and word length (*M* = 4.77, *SD* = 1.07). Two versions of the study list with differing word orders were created in order to account for item order-effects. No significant differences were found in recall based on item-order ($t < 1$) and so this will not be discussed further. Participants were asked to make judgments of retention interval (JORIs) on a min scale (0 – 60 min) where a judgment of 0 indicates not being able to remember the item immediately at all and a judgment of 60 indicates the ability to remember an item in 60 min (1 hour). The first two and last two words of the 30-item word lists were treated as primacy and recency buffers and were excluded from all analyses reported. Words were presented for 4 s each followed by 5 s to write down the JORI. A 500 ms inter-stimulus-interval (ISI) was included before the presentation of the next study item following the JORI. Following a 3 min filler task (writing down states of the United States) participants were allotted 3 min to write down as many words as they could remember from the study list (i.e., free-recall).

Results & Discussion

JORIs and Recall

As expected, participants provided shorter JORIs (M = 15.12 min, SD = 11.79 min) than would be anticipated from the JOL and RI literature (Carroll et al., 1997; Koriat et al., 2004). Thus, on average, participants predicted they would be able to remember a word for about 15 min after it had been presented. Further, a one-sample t-test with a significance level of .05 revealed that participants provided durations that were reliably greater than 0, $t(39) = 8.11$, $p < .001$. Additionally, participants recalled approximately one-third of the items studied (M = 33.44%, SD = 11.91%). These trends demonstrate that when participants are permitted to predict how long they will be able to remember words the durations specified are relatively modest, particularly compared with prior work using standard JOLs (Carroll et al., 1997; Koriat et al., 2004).

Resolution

Measures of resolution capture how well people predict performance on an item-by-item level. Conceptually, resolution examines the degree to which predictions distinguish between items that will and will not be remembered. Resolution is a measure of relative accuracy and is typically measured with a within-subjects Goodman-Kruskal Gamma correlation (Nelson, 1984; but see Masson & Rotello, 2009, for alternatives) by calculating an average Gamma correlation for each participant and then averaging across all participants. Gamma correlations do not require predictions and performance to reside on the same scales only that there be ordinal level data. For Pilot Study 1, a positive Gamma correlation between predictions and performance would indicate that high JORIs were given for remembered words and low JORIs were given for words that were not

remembered. A negative correlation would indicate the opposite; high JORIs given for words that were not remembered and low JORIs given for words that were remembered.

A one-sample t-test indicated that average Gamma correlations ($G = .28$, $SD = .34$) were significantly greater than 0, $t(39) = 5.28$, $p < .001$. Thus, on an item-by-item level, participants were significantly better than chance at determining whether they would be able to remember each word.

Pilot Study 2

Pilot Study 1 demonstrated that, on average, participants predicted they would be able to remember a list of words for approximately 15 minutes, in contrast to prior reports of more extravagant predictions (Carroll et al., 1997; Koriat et al., 2004). To extend this line of research it was necessary to determine if participants' JORIs were sensitive to other manipulations, such as item difficulty. Thus, Pilot Study 2 was designed to determine whether the JORIs that participants provide are similarly sensitive to item difficulty, as previously reported for JOLs (e.g., Koriat et al., 2004; Koriat & Ma'ayan, 2005).

Pilot Study 2 used a within-subjects design where the manipulation of primary interest was item difficulty. Prior work has consistently shown that abstract words are more difficult to remember than concrete words (e.g., Paivio, 1966). Thus, for this experiment, item difficulty was manipulated by having participants study abstract and concrete words[2]. Based on prior work (e.g., Koriat et al., 2004; Koriat & Ma'ayan, 2005)

[2]A novel aspect of Pilot Study 2 is the use of concreteness as a manipulation of item difficulty. Item difficulty has more typically been manipulated by employing related (e.g., CAT-KITTEN) and unrelated word pairs (e.g., DECK-GLASS) (e.g., Koriat, Bjork, Sheffer, & Bar, 2004).

it was expected that item difficulty would impact recall such that participants would recall more concrete words than abstract words. In addition, it was expected that predictions would reflect the difference in recall such that JORIs would be lower for abstract compared with concrete words. This pattern of results would replicate trends demonstrated previously with JOLs (e.g., Koriat et al., 2004; Koriat & Ma'ayan, 2005).

Method

Participants

Thirty-four Colorado State University students (*M* age = 18.94, *SD* = 2.09) were tested and received extra credit in a psychology course.

Design

A single within-subjects design (Item Type: concrete, abstract) was employed and JORIs and recall were measured.

Materials & Procedure

Materials consisted of 30 words that varied on concreteness (MRC Psycholinguist Database, 1987; Range = 100-700; Concrete words: *M* = 631.46, *SD* = 9.43; Abstract words: *M* = 258.23, *SD* = 10.90). Abstract and concrete words were additionally controlled for frequency and length (Concrete words: MRC Psycholinguist Database, 1987; *M* frequency = 37.62, *SD* = 28.47, *M* length = 5.85, *SD* = 0.48; Abstract words: MRC Psycholinguist Database, 1987; *M* frequency = 44.08, *SD* = 50.18, *M* length = 5.85, *SD* = 1.21). The study list consisted of 15 concrete words (e.g., *rabbit*) and 15 abstract words (e.g., *value*) that were randomly intermixed. Two versions were created with differing word orders to account for item order-effects. No significant differences were

found in recall based on item-order ($t < 1$) so this will not be further discussed. Otherwise, the procedure used was identical to the procedure used in Pilot Study 1.

Results & Discussion

JORIs and Recall

Results demonstrated that participants recalled significantly more concrete words ($M = 39.28$, $SD = 18.82$) than abstract words ($M = 25.21$, $SD = 13.50$), $t(33) = 4.11$, $p < .001$; *Cohen's d* = .87. Moreover, JORIs (see Figure 1) were sensitive to this difference in recall. Specifically, JORIs for concrete words ($M = 17.94$, $SD = 10.67$) were significantly longer than JORIs for abstract words ($M = 12.74$, $SD = 9.06$), $t(33) = 4.91$, $p < .001$; *Cohen's d* = .53. This pattern demonstrates that the JORIs participants provided were sensitive to item difficulty such that shorter JORIs were provided for items (i.e., abstract words) that were less likely to be remembered.

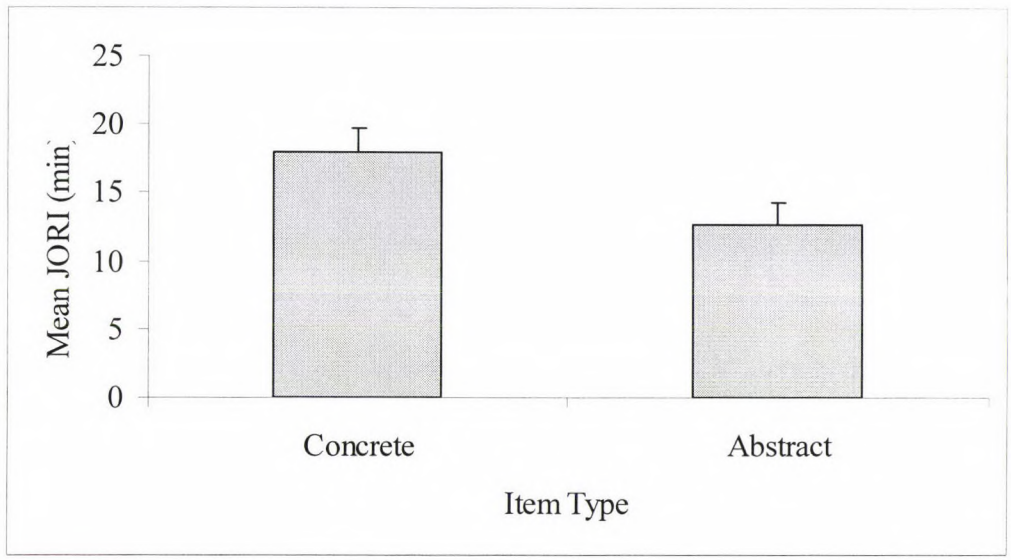

Figure 1. Pilot Study 2 JORI Data.

Resolution

A one-sample t-test revealed that average Gamma correlations ($G = .34$, $SD = .38$) were significantly greater than 0, $t(31) = 5.12$, $p < .001$. Thus, when including item difficulty as a cue, participants' JORIs were sensitive to later recall.

Summary: Pilot Data

The pilot studies demonstrated that participants provided much shorter durations than would be anticipated from the JOL and RI literature (Carroll et al., 1997; Koriat et al., 2004). Further, Pilot Study 2 suggested that participants' JORIs are sensitive to a variable that influences later memory performance.

Dissertation Experiments

In order to continue this line of inquiry three experiments were conducted. Experiment 1 examined the importance of the JORI scale provided and varied the RI. Experiment 2 explored how providing JORIs influences a standard metamemory control measure (i.e., restudy selection) and also compared JORIs and JOLs. Finally, Experiment 3 tested whether participants' JOLs were influenced by different framings of the RI. To summarize, I expected to replicate standard patterns of monitoring and control previously evident for JOLs (e.g., Finn, 2008) with JORIs; for example, the relationship between JORIs and recall should be similar to the relationship between JOLs and recall. Of particular interest, participants' JORIs might also be more sensitive to RI such that, in contrast to prior work with JOLs (e.g., Koriat, et al., 2004), JORIs would differ for a long compared with short RI. Finally, I also examined whether participants' JOLs might become sensitive to RI when a more specific temporal value was used to specify the RI.

Experiment 1: Scale Comparison & RI

Experiment 1 was designed to further explore the importance of the JORI scale while simultaneously examining the sensitivity of JORIs to RI. Thus, in Experiment 1, participants made JORIs in reference to a 30 second or 20 minute RI to permit an evaluation of whether predictions would vary based on the interval specified (See Appendix A and Table 1A for data on a 10 min RI Condition).

In order to fully assess whether the distribution of judgments would change based on the scale used half of the participants made JORIs on a binary scale, indicating whether they would able to remember each word for the *RI or longer* or *less than the RI.* For participants using a binary scale, half made their judgments anticipating a 30 sec RI while the other half made their judgments anticipating a 20 min RI. For example, participants in the 30 sec RI condition determined if each word would be remembered for either "30 sec or longer" or "less than 30 sec".

The use of a binary scale may diminish sensitivity to the units of the judgment by employing larger units of time (e.g., Yes, I will remember this longer than the RI) rather than individual units of time (e.g., I will remember this word for 5 min, but another word for only 1 min). Eliciting a judgment in aggregate units rather than individual units should reduce salient RI cues. Dunning, Heath, and Suls (2004) offer a related example with programmers employed by Microsoft. If employees are asked to generate how much time it will take to complete a given task, a typical answer is "about a month". However, if that month is parsed into usable amounts of time (e.g., there are about 22 working days in the average month, what 22 things are you going to be able to accomplish in that time?) employees are better able to understand that they are underestimating that amount

of time it will take to complete the project. It is possible that JORIs function in a similar manner, such that continuous scale JORIs (0 - 60 min) allow time to be broken into individual units, changing the manner in which predictions are made and perhaps increasing participants' appreciation of an intervening RI. However, other previous work suggests that improvements may result in resolution when limiting the JORI scale to fewer points (Benjamin & Diaz, 2008) compared with a continuous 0-60 min scale.

The other half of the participants in Experiment 1 used a continuous scale (0 – 60 min) for JORIs, identical to Pilot Study 1. Of the participants using a continuous scale, half were asked to make their judgments while anticipating a 30 sec RI while the other half made their judgments while anticipating a 20 min RI. Thus, in Experiment 1 (see Table 1 for an overview) participants made judgments on a binary or continuous scale for a 30 sec or 20 min RI. Participants in all conditions waited the designated RI prior to recall (i.e., across conditions, half of the participants had a 30 sec RI and half had a 20 min RI).

Table 1 *Design of Experiment 1*

Judgment Condition	Retention Interval (RI)	Recall
Binary JORI	30 sec	3 min Free-recall
Binary JORI	20 min	3 min Free-recall
Continuous JORI	30 sec	3 min Free-recall
Continuous JORI	20 min	3 min Free-recall

Based on prior work, it was expected that participants would be poor at making predictions if given a binary scale, perhaps due to diminished scale sensitivity (e.g.,

Dunning, et al., 2004). In other words, the frequency of judgments that a word would be remembered for more than the RI (e.g., remember words for longer than 30 sec/20 min) and less than the RI (e.g., remember words for 30 sec/20 min or less) was expected to be equivalent for the 30 sec and 20 min RI conditions. Drawing from Pilot Study 1, minute scale judgments were expected to be more diagnostic than binary judgments (i.e., due to increased scale sensitivity). That is, if participants in the continuous JORI condition are sensitive to RI, they may make shorter predictions when anticipating a 20 min RI compared with participants anticipating a 30 sec RI. Finally, participants were expected to demonstrate equivalent levels of recall for binary judgments above and below the RI such that participants would remember about the same percentage of items given shorter JORIs compared with longer JORIs.

Method

Participants

Sixty-four Colorado State University students (16 per condition) were tested (*M* age = 19.08, *SD* = 1.62) and received extra credit in psychology courses for participating.

Design

A 2 (RI Condition: 30 sec, 20 min) x 2 (Type of Judgment: Continuous, Binary) mixed-factor design was employed with Measure manipulated within-subjects and RI Condition and Type of Judgment manipulated between-subjects.

Materials & Procedure

Participants studied the same words as in Pilot Study 1 and made JORIs for each word anticipating a memory test on the entire list following either 30 sec or 20 min RI. One of two types of JORIs was elicited. Half of the participants made JORIs on a

continuous min scale (0 – 60 min) and half made JORIs on a binary scale. In addition, half of the participants in each condition were tested after a 30 sec RI and half were tested after a 20 min RI. For the 30 sec RI condition making binary judgments, participants indicated whether they would be able to remember a word for 30 sec or less or for more than 30 sec. For the 20 min RI condition making binary judgments, participants indicated if they would be able to remember a given word for 20 min or less or for more than 20 min. For all participants given a 30 sec RI between study and test, the interval was filled by subtracting numbers for 30 sec. For all participants given the 20 min RI, the interval was filled by completing math problems for 5 min, followed by 5 min listing U. S. states, then 5 min completing a new set of math problems, and finally 5 min listing major U.S. cities. Following the RI, memory was tested via free recall (3 min). See Table 1 for an overview of Experiment 1.

Experiment 1 Results

Recall

Average recall performance is presented in Table 2. A 2 (Type of Judgment: Continuous, Binary) x 2 (RI Condition: 30 sec, 20 min) between-subjects ANOVA comparing recall performance in the four conditions (i.e., Binary 30 sec RI, Binary 20 min RI, continuous 30 sec RI, and continuous 20 min RI) was conducted. Overall, participants in the 30 sec RI conditions recalled significantly more words than participants in the 20 min RI conditions, $F(1, 60) = 12.97$, $p = .001$, $\eta^2_p = .18$. No reliable differences were found between the Continuous scale and the Binary scale conditions, $F < 1$. Additionally, the Type of Judgment did not interact with RI Condition, $F(1, 60) = 2.31$, $p = .13$, $\eta^2_p = .04$.

Table 2 *Average Recall Performance for Experiment 1*

	30 sec RI	20 min RI
Continuous Judgment	29.56 (8.50)	20.88 (6.97)
Binary Judgment	33.94 (13.42)	24.25 (10.74)
Overall	31.75 (11.27)	22.56 (9.07)

*Note. SD*s in parentheses.

In order to further compare the continuous and binary conditions, recall performance was categorized by JORI such that two averages were created for each condition: recall performance for items with JORIs at or below the RI and recall performance for items with JORIs longer than the RI (see Figure 2). For completeness, I report categorized recall for all conditions, but it should be noted that 10 participants in the continuous 30 sec condition did not provide JORIs less than 30 sec. Thus, recall performance for the 30 sec continuous condition should be interpreted with caution. A 2 (Type of Judgment: Continuous, Binary) x 2 (RI Condition: 30 sec, 20 min) x 2 (Categorized Recall: more than the RI, less than the RI) mixed-factor ANOVA was conducted. Overall, recall was reliably greater when conditionalized by JORIs that were longer than the RI ($M = 38.09$, $SD = 22.87$) compared with recall conditionalized by JORIs that were less than the RI ($M = 19.20$, $SD = 16.08$), $F(1, 45) = 21.17$, $p < .0001$, $\eta^2_p = .32$. No other reliable main effects or interactions were supported. In sum, across conditions, longer JORIs were associated with higher levels of recall. However, soliciting JORIs on a binary or continuous scale had no impact on later recall.

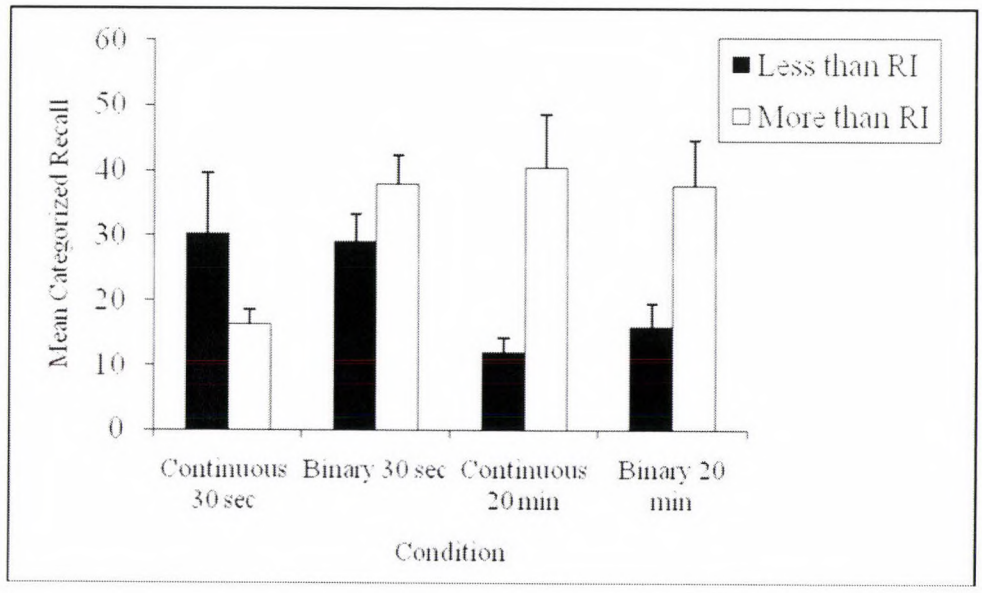

Figure 2. Categorized recall data for Experiment 1.

JORIs

The magnitude of continuous JORIs was examined via an independent samples t-test comparing JORIs by RI (i.e., 30 sec, 20 min). Contrary to expectations, no reliable differences were found between the continuous 30 sec condition (M = 25.08 min, SD = 12.19) and the continuous 20 min condition, (M = 21.27 min, SD = 11.43), $t < 1$. An independent samples t-test comparing JORIs from Experiment 1 with overall JORIs (collapsed across 30 sec and 20 min RI conditions) from Pilot Study 1 revealed that the JORIs in Experiment 1 (M = 23.18 min, SD = 11.78) were reliably longer than the JORIs from Pilot Study 1, (M = 15.12 min, SD = 11.79), $t(70) = 2.88$, $p = .005$, *Cohen's d* = .68. It appears that providing information about the expected RI (Experiment 1) produced longer JORIs than when no information was provided about the RI (Pilot Study 1). However, it should be noted that the JORIs in both experiments were still shorter than what might be expected from the JOL literature.

I next explored the frequency of judgments above and below the given RI in order to compare JORIs for the two RI conditions and the two judgment conditions. JORIs for both continuous scale conditions were thus categorized into two conditions (above or below the respective RI) in order to compare these conditions with the binary scale conditions (see Figure 3).

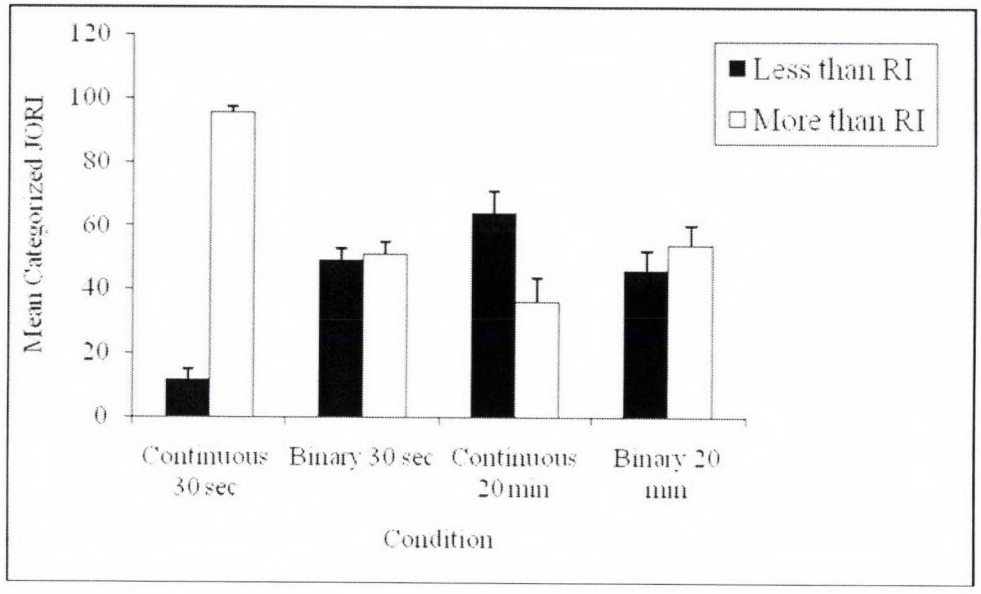

Figure 3. Categorized JORIs for all conditions in Experiment 1.

A 2 (Type of Judgment: Continuous, Binary) x 2 (RI Condition: 30 sec, 20 min) x 2 (Categorized JORI: more than RI, less than RI) mixed-factor ANOVA was conducted on the percentage of JORIs above and below the mean. Similar to the recall data, these data should be interpreted with caution as only 6 participants in the continuous 30 sec condition provided JORIs that fell less than the RI and these JORIs were consistently zero. The main effect of Categorized JORI was supported, $F(1, 50) = 5.34$, $p = .03$, $\eta^2_p = .10$, such that there were reliably more JORIs that were greater than the respective RIs ($M = 60.48$, $SD = 28.94$) compared with a lower percentage of JORIs that were shorter than the RIs ($M = 49.78$, $SD = 26.07$). However, the main effects of Type of Judgment, $F(1,$

50) = 1.65, p = .21, p = .21, η^2_p = .03, and RI Condition, $F(1, 50) = 1.30$, p = .26, η^2_p = .03, were not reliable.

These main effects were qualified by a few reliable interactions. First, Categorized JORI reliably interacted with RI Condition, $F(1, 50) = 16.10$, $p < .0001$, η^2_p = .24. In particular, significantly more JORIs were greater than the RI for the 30 sec RI compared with the 20 min RI condition, $t(62) = 3.21$, p = .002, *Cohen's d* = .81, whereas significantly more JORIs were less than the RI for the 20 min RI compared with the 30 sec condition, $t(52) = 2.35$, p = .02, *Cohen's d* = .65 (see Figure 4).

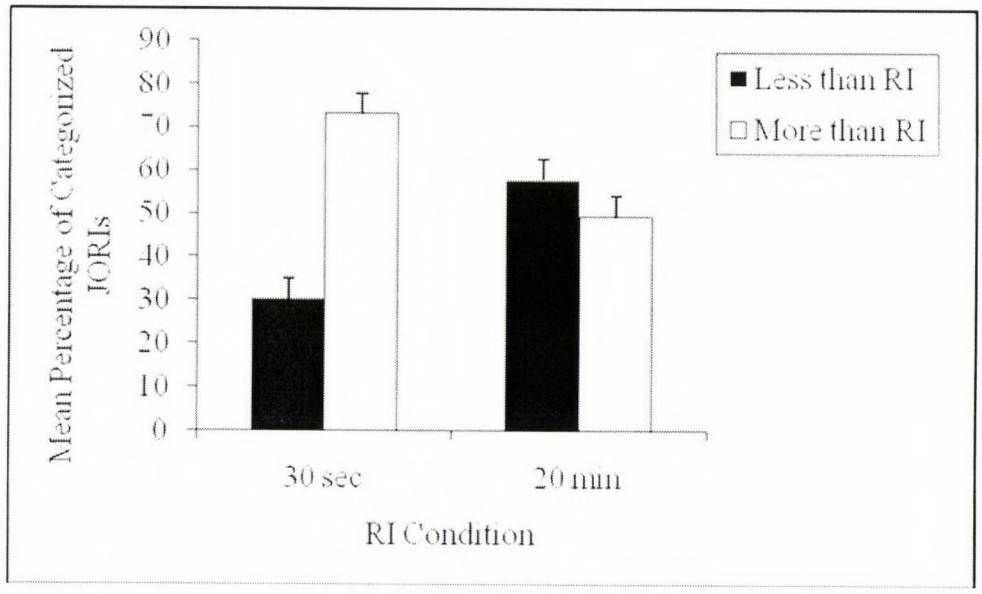

Figure 4. Categorized JORIs by RI Condition (collapsed across judgment scale conditions) in Experiment 1.

Also, Categorized JORI, Type of Judgment, and RI Condition reliably interacted, $F(1, 50) = 23.44$, $p < .0001$, η^2_p = .32. Specifically, for the Binary Conditions, Categorized JORI did not reliably interact with RI Condition ($F < 1$) indicating that the percentage of JORIs above and below the RIs did not vary when participants made their judgments on a binary scale. However, Categorized JORI did reliably interact with RI

Condition for the Continuous Conditions, $F(1, 17) = 32.83$, $p < .0001$, $\eta^2_p = .66$. In particular, the percentage of JORIs below the RI were significantly greater in the 20 min RI Condition, $t(17) = 5.73$, $p < .0001$, *Cohen's d* = .56, compared with the 30 sec RI Condition, whereas the percentage of JORIs above the RI were significantly greater in the 30 sec RI condition, $t(17) = 5.73$, $p < .0001$, *Cohen's d* = .58, compared with the 20 min RI Condition. Further, in the 30 sec RI Condition reliably more JORIs were greater than the 30 sec RI, $t(17) = 5.15$, $p < .0001$, *Cohen's d* = .87, compared fewer JORIs that were below the 30 sec RI, whereas in the 20 min RI Condition reliably more JORIs were less than the 20 min RI, $t(17) = 2.62$, $p = .02$, *Cohen's d* = .31, compared with fewer JORIs that exceeded the 20 min RI. No other interactions were evident.

Resolution

A JORI-recall Gamma correlation (*G*) was calculated for each of the four JORI conditions (i.e., Binary 30 sec RI, Binary 20 min RI, continuous 30 sec RI, continuous 20 min RI). Four one-sample t-tests were conducted comparing each *G* (see Table 3) against chance (zero). *G* correlations were reliably greater than chance for the Continuous 30 sec condition: $t(15) = 2.88$, $p = .01$; the Continuous 20 min condition: $t(15) = 6.23$, $p < .0001$; and the Binary 20 min condition: $t(13) = 3.61$, $p = .003$. However, *G* correlations were not reliably different from chance for the Binary 30 sec condition, $t(15) = 1.38$, $p = .19$. A 2 (Type of Judgment: JOL, JORI) x 2 (RI Condition: 30 sec, 20 min) between-subjects ANOVA on Gamma correlations indicated that *G* correlations were reliably greater for the 20 min RI ($M = .467$, $SD = .401$) compared with the 30 sec RI ($M = .185$, $SD = .385$) condition, $F(1, 58) = 7.70$, $p = .007$, $\eta^2_p = .12$. These data are in contrast with prior reports (i.e., Koriat et al, 2004, Experiment 1) which have demonstrated no impact of RI on

resolution. This discrepancy between the current data and prior work may be a result of differing RIs, as Koriat, et al. (2004) employed immediate, 1 day, and 1 week intervals whereas in the current data RIs of 30 sec and 20 min were used. Type of Judgment did not influence *G* correlations, nor did Type of Judgment interact with RI Condition (*F*s < 1).

Table 3 *Resolution data for Experiment 1*

	Gamma (*G*)
Continuous 30 sec*	.208 (.290)
Binary 30 sec	.162 (.471)
Continuous 20 min*	.435 (.279)
Binary 20 min*	.498 (.516)

*Note. SD*s provided in parentheses.
* significantly greater than zero.

Experiment 1 Discussion

To summarize, recall performance was reliably lower in the long RI conditions (i.e., Continuous 20 min and Binary 20 min) compared with the shorter RI conditions (i.e., Continuous 30 sec and Binary 30 sec). Additionally, memory performance was greater for words given longer JORIs compared with shorter JORIs. While the 30 sec RI conditions resulted in reliably better recall compared with the 20 min RI conditions there were no differences in JORIs based on RI. These data are consistent with prior research (e.g., Koriat et al., 2004) indicating that participants may be insensitive to RI. However, participants' judgments on a minute scale were reasonable time estimates of memory performance in contrast with prior JOL and RI research (e.g., Koriat, et al., 2004), replicating Pilot Study 1.

Further, as expected, there was more variability in the distribution of JORIs when made on a continuous scale compared with binary scale JORIs. That is, JORIs were equally distributed around the RI when made on a binary scale regardless of whether participants were anticipating a 30 sec or 20 min interval. This indicates that participants were not very effective at incorporating the RI into their judgments when the scale was condensed into a two-alternative choice. In contrast, participants' judgments varied in the continuous conditions were better adjusted for the RI. Participants who were expecting a 30 sec RI provided more judgments that exceeded 30 sec, whereas participants who were expecting a 20 min RI provided more judgments below 20 min. Thus, it appears that participants were better able to attend to RI when judgments were made on a 0-60 min scale. It should again be noted that these data should be interpreted with caution as only 6 participants in the continuous 30 sec condition provided JORIs that could be categorized above and below the 30 sec RI. Regardless, what is not apparent from the experiments thus far is what impact JORIs might have on control processes. This issue was explored in Experiment 2.

Experiment 2: Control - Restudy Selection

Metacognitive control can be thought of as how metacognitive knowledge is used to regulate and change behavior (Koriat, 2007). Control is typically measured via study termination, self-paced study, or the selection of items for restudy (Nelson & Narens, 1994). Of particular interest in Experiment 2 is restudy choice whereby participants are offered the opportunity to select items to restudy during encoding.

The discrepancy-reduction model (e.g., Thiede & Dunlosky, 1999) posits that goals are initially set for a standard of learning (termed the norm of study) which is then

compared with the current state of knowledge (but see Ariel, Dunlosky, & Bailey, 2009; Metcalfe & Kornell, 2005, for alternatives). The degree of separation between the desired goal (e.g., to learn the given information) and state of learning (e.g., how much of the information is learned) is the discrepancy that needs to be reduced or eliminated in order to perform optimally. Thus, participants will seek to restudy those items with the greatest discrepancy. As empirical support for this model, Thiede and Dunlosky conducted several experiments examining the interaction between metacognitive control and monitoring. For example, goal status was manipulated (Experiment 1) such that participants were either given an easy goal or a challenging goal. All participants generated JOLs and were allowed to select items for a later restudy opportunity. JOLs were negatively correlated with item restudy selection under the difficult goal context such that items were less likely to be selected for restudy if given a high JOL. This suggests that when given a challenging goal, participants seek to restudy items they deem less well learned.[3]

It is important to determine how JORIs impact control in order to fully assess this new metamemory measure. Specifically, in Experiment 2, I sought to determine what influence eliciting JORIs had on restudy selection. Finn (2008) reported that framing metamemory judgments in terms of forgetting led more items to be selected for restudy compared with framing metamemory judgments in terms of remembering. This was explored in Experiment 2 to determine whether soliciting JORIs would similarly lead to more frequent restudy choices than JOLs.

[3] It should be noted that an easy goal changes the relationship between JOLs and restudy.

Participants were assigned to one of three conditions in Experiment 2. One-third of the participants made a JOL, one-third made JORIs, and the final third made no metamemory monitoring judgment. In all three conditions participants determined which items they would like to restudy. For participants in the JOL and JORI conditions the restudy decision immediately followed each judgment. Participants studied concrete and abstract items (as in Pilot Study 2) to examine the impact of item difficulty on restudy selection. Finally, participants in the JOL and JORI conditions completed a qualitative post-experiment survey (see Appendix B) after completing the study-test session. The post-experiment survey focused on what information participants used to make their JORIs/JOLs in order to gather self-report information about the bases of metacognitive judgments and restudy selection.

It was expected that participants would choose to restudy more abstract than concrete words regardless of condition (i.e., JOL, JORI, Restudy Only). This pattern would demonstrate awareness that abstract words are less memorable than concrete words and thus warrant further study (cf. Thiede & Dunlosky, 1999). In addition, it was expected that participants would provide lower JORIs for abstract compared with concrete words (see Pilot Study 2) and, similarly, that participants would provide lower JOLs for abstract compared with concrete words (e.g., Koriat et al., 2004; Koriat & Ma'ayan, 2005). Finally, I anticipated that participants would remember more concrete than abstract words regardless of condition (e.g., Pilot Study 2). With regard to restudy selection, I anticipated that participants making JORIs might choose to restudy more words overall in comparison with participants making JOLs and participants making only restudy selections. This prediction is based on prior data indicating that judgment framing

can influence metacognitive control choices (Finn, 2008). Specifically, following a forget framed JOL, Finn observed that participants chose more items for restudy compared with participants who made JOLs framed in the more standard remember context.

In terms of resolution, positive correlations were expected between judgments and recall (i.e., items assigned higher values of JOLS or JORIs would be more likely to be remembered). An inverse relationship was expected between restudy selection and recall for all conditions such that participants would select to restudy more words that were less likely to be remembered (i.e., abstract words), with these more challenging words resulting in lower overall recall. As well, a negative correlation should be evident between restudy selection and JORIs/JOLs, such that items more often selected for restudy would have received lower JORIs/JOLs. These patterns would replicate trends previously demonstrated with JOLs (e.g., Nelson, Dunlosky, Graf, & Narens, 1994; Dunlosky & Thiede, 2004; Thiede & Dunlosky, 1999).

Method

Participants

One-hundred and twenty Colorado State University students (*M* age = 19.38, *SD* = 1.66) participated (40 in each condition) in exchange for course credit.

Design

A 2 (Item Type: concrete, abstract) x 3 (Type of Judgment: JOL, JORI, Restudy Only) mixed-factor design was employed with Item Type manipulated within-subjects and Condition manipulated between-subjects.

Materials

The 30-item word list used in Pilot Study 2 was used. As in Pilot Study 2, half of the words were concrete (e.g., *rabbit*) and half were abstract (e.g., *value*). Two randomized versions of the word list were created so that all participants learned the same words, but half learned them in a different order. No significant differences were found in recall based on item-order ($t < 1$) and so this will not be discussed further.

Procedure

Participants studied words presented one-at-a-time for 4 s. In two conditions, immediately after the presentation of each word, participants made a prediction. One-third of the participants made JORIs indicating how long they would be able to remember each word on a min scale (0- 60 min). Another one-third of the participants made standard JOLs predicting the likelihood of recalling each item on a later memory test on a scale from 0 (*not likely at all*) to 100 (*very likely*). Following typical methodology (e.g., Finn, 2008; Rhodes & Castel, 2009) restudy choices immediately followed each JORI or JOL. In the restudy only condition, participants made the same restudy selection decision with this judgment preceded by study and no intervening monitoring judgment. Participants in all conditions circled "YES" or "NO" to indicate the desire to restudy a word. It should be noted that participants were not actually offered a restudy opportunity prior to free recall (e.g., Finn 2008). Participants then completed a 3 min filler task (listing states of the United States of America) and were given a 3 min free-recall test. Finally, a portion of participants in the JOL and JORI conditions completed a post-experiment survey indicating what they considered when making their predictions (see Appendix B). Questions on the first side of the survey were open-ended and the questions

on the back of the survey were structured as a checklist. Participants were directed to complete the first side of the survey prior to completing the checklist on the back.

Experiment 2 Results

Recall

The mean percentage of items recalled (see Table 4) was examined in a 2 (Item Type: Abstract, Concrete) x 3 (Type of Judgment: JOL, JORI, Restudy Only) mixed-factor ANOVA. Results showed that recall was significantly lower for abstract compared with concrete words, $F(1, 117) = 140.97$, $p < .0001$, $\eta^2_p = .55$. There was no main effect Type of Judgment nor was there an Item Type x Type of Judgment interaction (Fs < 1). Thus, participants in all conditions exhibited better memory for concrete compared with abstract words.

Table 4 *Experiment 2 Percent recall data by Item Type*

	Abstract	Concrete	Overall
JOL Condition*	24.93 (11.83)	42.05 (14.88)	33.55 (11.29)
JORI Condition*	22.78 (13.21)	42.53 (16.15)	32.68 (10.36)
Restudy Only Condition*	21.28 (14.52)	41.48 (21.26)	31.40 (16.11)
Overall*	22.99 (13.20)	42.02 (17.50)	

*Note. SD*s provided in parentheses. * Significant difference between Abstract and concrete recall performance, $p < .05$.

Metamemory Judgments

Due to the differing scales for JOLs (measured in percentages) and JORIs (measured in minutes) monitoring judgments were analyzed separately. A paired-samples t-test comparing JOLs by Item Type (i.e., concrete, abstract) indicated that participants provided significantly higher JOLs (measured in percentages) for concrete words (M =

61.92, SD = 15.14) compared with abstract words (M = 49.62, SD = 16.44), $t(39) = 8.71$, $p < .0001$, *Cohen's d* = .78. Likewise, participants provided significantly longer JORIs (measured in minutes) for concrete words (M = 25.78 min, SD = 12.49) compared with abstract words (M = 18.12 min, SD = 9.34), $t(39) = 6.70$, $p < .0001$, *Cohen's d* = .69. Thus, participants in both judgment conditions expected concrete words to be more memorable than abstract words.

Restudy Selection

The mean proportion of items selected for restudy (Table 5) was analyzed in 2 (Item Type: Abstract, Concrete) x 3 (Type of Judgment: JOL, JORI, Restudy Only) mixed-factor ANOVA. Overall, significantly more abstract words were selected for restudy than concrete words, $F(1, 117) = 188.27$, $p < .0001$, $\eta^2_p = .62$. A main effect of Type of Judgment was also present, $F(2, 117) = 3.71$, $p < .05$, $\eta^2_p = .06$. In particular, participants in the JORI condition selected more items for restudy compared with the JOL condition, $t(78) = 2.03$, $p < .05$, *Cohen's d* = .45, and the Restudy Only condition, $t(78) = 2.21$, $p < .05$, *Cohen's d* = .49 (see Figure 5). However, there was no difference in the proportion of items selected for restudy between the JOL and Restudy Only conditions, $t < 1$. Finally, Item Type did not interact with Type of Judgment, $F < 1$. Thus, more abstract than concrete items were selected for restudy across conditions. Critically, participants also chose to restudy more items overall when making JORIs compared with the JOL and Restudy Only conditions.

Table 5 *Experiment 2 Percentage of Items selected for Restudy*

	Abstract	Concrete
JOL Condition*	51.54 (33.38)	26.54 (30.62)
JORI Condition*	65.96 (24.07)	36.73 (28.65)
Restudy Only Condition*	53.10 (24.61)	22.33 (19.14)
Overall*	56.87 (28.21)	28.53 (27.08)

*Note. SD*s provided in parentheses. * Significant difference in the proportion selected for restudy between abstract and concrete items, $p < 05$.

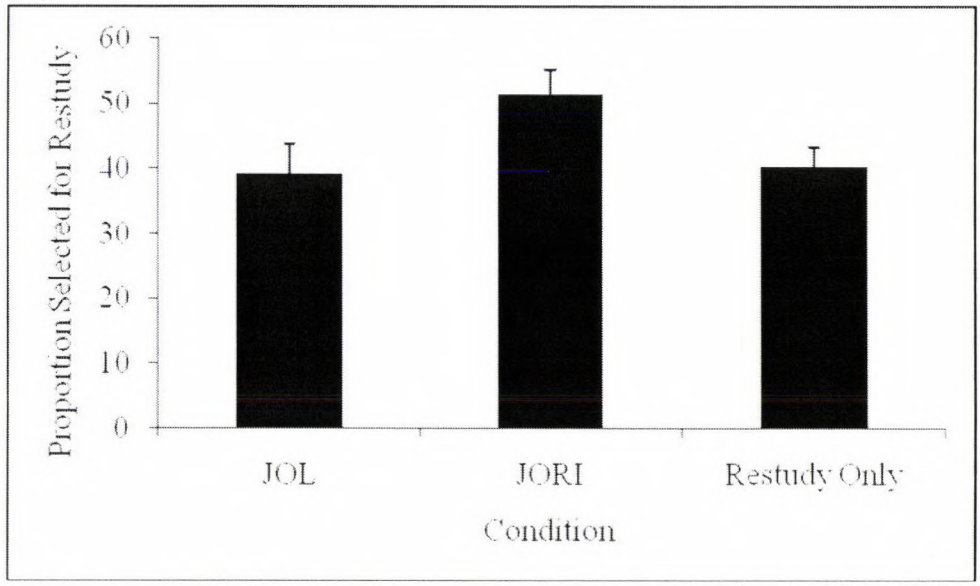

Figure 5. Proportion selected for restudy for Experiment 2.

Calibration

Calibration can only be examined for the JOL condition because the two measures were assessed using the same scale (i.e., percentage). A 2 (Item Type: Abstract, Concrete) x 2 (Measure: JOL, recall) repeated-measures ANOVA was conducted on the mean predicted and actual recall performance for the JOL condition (see Figure 6). Overall, concrete words ($M = 51.98$, $SD = 11.42$) received higher JOLs and were more

likely to be recalled than abstract words ($M = 37.27$, $SD = 10.04$), $F(1, 39) = 91.98$, $p < .00001$, $\eta^2_p = .07$. In addition, JOLs ($M = 55.77$, $SD = 15.16$) exceeded recall performance ($M = 33.49$, $SD = 11.37$), $F(1, 39) = 56.78$, $p < .00001$, $\eta^2_p = .59$, demonstrating overconfidence. Finally, Item Type interacted with Measure, $F(1, 39) = 4.79$, $p < .05$, $\eta^2_p = .11$. Follow-up tests indicated that while JOLs reliably exceeded recall for abstract words, $t(39) = 7.65$, $p < .0001$, *Cohen's d* = 1.73, and concrete words, $t(39) = 6.45$, $p < .0001$, *Cohen's d* = 1.32, this discrepancy was greater for the abstract words. Thus, participants were more poorly calibrated for abstract compared with concrete items.

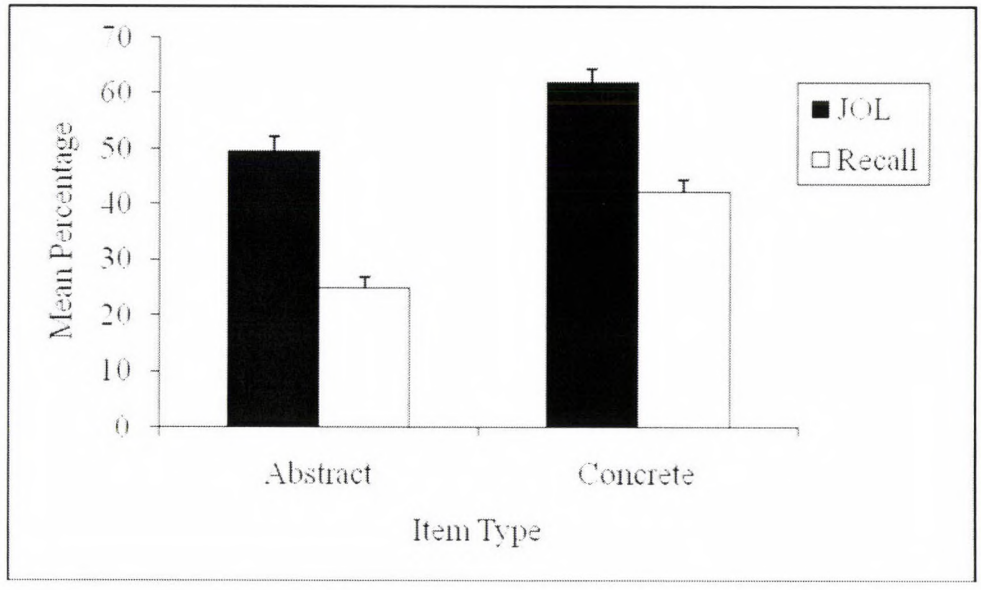

Figure 6. Calibration Data for the JOL Condition in Experiment 2.

Resolution

Mean gamma correlations for the JOL and JORI conditions can be found in Table 6. Seven Gamma correlations were calculated, three for the JORI condition (i.e., JORI-Restudy, JORI-recall, Restudy-recall), three for the JOL condition (i.e., JOL-Restudy, JOL-recall, Restudy-recall) and one for the Restudy only condition (i.e., Restudy-recall).

Seven one-sample t-tests were conducted comparing each resulting Gamma correlation against chance (zero).

Judgment-Recall Gammas. As expected, a positive relationship was found between judgments and recall for the JOL and JORI conditions. Each Judgment-Recall *G* correlation was reliably greater than chance (JOL Condition, $t(39) = 8.09$, $p < .0001$; JORI Condition, $t(39) = 7.50$, $p < .0001$). An independent samples t-test comparing Judgment-Recall *G* for the JOL and JORI condition demonstrated that *G* did not reliably differ between the JOL and JORI conditions, $t < 1$.

Judgment-Restudy Gammas. An inverse relationship (i.e., words with low JOLs associated with more restudy) was found between Judgments and Restudy with means reliably different from zero for the JOL condition, $t(32) = 46.78$, $p < .0001$, and JORI condition, $t(38) = 31.62$, $p < .0001$. An independent samples t-test comparing Judgment-Restudy *G* between the JOL and JORI conditions indicated that there was no reliable difference between conditions, $t(70) = 1.31$, $p = .19$.

Restudy-Recall Gammas. An inverse relationship was found between restudy selection and recall for each condition and each *G* was reliably less than zero (JOL Condition, $t(32) = 4.58$, $p < .0001$; JORI Condition, $t(38) = 9.54$, $p < .0001$; and Restudy Only Condition, $M = -.458$, $SD = .550$, $t(36) = 5.06$, $p < .001$). A one-way ANOVA comparing Restudy-Recall *G* for the 3 conditions indicated that there was no reliable difference between conditions, $F < 1$.

Table 6 *Experiment 2 Resolution data for the JOL and JORI condition*

	Judgment-Recall	Judgment-Restudy	Restudy-Recall
JOL Condition	.365(.054)*	-.908 (.024)*	-.406 (.107)*
JORI Condition	.359 (.057)*	-.863 (.032)*	-.518 (.065)*

*Note. SD*s provided in parentheses. * Significantly different from chance, $p < .05$.

Post-Experiment Survey

A portion of the participants in the JOL ($n = 30$) and JORI ($n = 24$) conditions completed a post-experiment survey intended to gather information about what participants used as bases for their judgments. A careful assessment of the open-ended questions led to no clear, discernable differences between judgment conditions (see Appendix C for sample open ended-responses). In general, participants reported using the following as bases for their predictions: personal theories about memory, item difficulty, and memory strategies employed.

I then explored the frequency of 'yes' responses to the question: Do you think your memory predictions were accurate? An independent samples t-test revealed no differences between the JOL condition ($M = 26.67$, $SD = 44.98$) and JORI condition, $M = 16.67$, $SD = 38.07$, $t < 1$. The frequency of 'yes' responses to the following question was then examined: If you did this task again would you make different predictions? An independent samples t-test revealed no differences between the JOL condition ($M = 70.00$, $SD = 46.61$) and JORI condition, $M = 87.50$, $SD = 33.78$, $t(52) = 1.54$, $p = .13$. Aggregate responses to the checklist (Appendix B) are presented in Table 7. First, the frequency of 'yes' responses was compared between the JOL and JORI conditions to determine if participants in one condition selected more items on the checklist. An

independent samples t-test revealed no reliable differences between the JOL condition (M = 6.28, SD = 2.20) and JORI condition, M = 5.58, SD = 1.79, $t(51) = 1.24$, $p = .22$. Next, 13 independent samples t-tests were conducted comparing each question between the JOL and JORI conditions. Due to the large number of tests, a Bonferroni correction was employed with the p value set at .004 for each analysis. No reliable differences were found between Judgment Condition for any of the 13 t-tests (see Table 7). The three most commonly reported bases for judgments were strategy-based memory techniques. Specifically, participants indicated whether they had thought of the following things: "if the words reminded you of something else, something relevant to your life", "if you thought the words would be easy or difficult to remember", and "if you tried to relate the words to other words you were studying". The three lowest rated bases for judgments (excluding "other" which was the lowest rated basis for judgments) were factors about the experimental context or outside information. Specifically, participants indicated that they did not often consider: "how you think other people would make predictions", "how long the words were presented for", and "the time between studying and testing".

Table 7 *Percent of Participants self-reporting each Basis for Predictions on the Post-Experiment Survey*

Question	JOL	JORI	Inferential Statistics
words reminded you of something else relevant to your life	93.10 (4.79)	87.50 (6.90)	$t < 1$
tried to related the words to other words you were learning	86.21 (6.52)	62.50 (10.10)	$t(52) = 1.75$, $p = .09$
thought the words would be easy or difficult to remember	82.76 (7.14)	87.50 (6.90)	$t < 1$
whether you would choose to restudy the words or not	62.07 (9.17)	79.17 (8.47)	$t(52) = 1.51$, $p = .14$
the number of words you were learning overall	62.07 (9.17)	29.17 (9.48)	$t(52) = 2.34$, $p = .02$
tried to repeat the words in order to remember them	58.62 (9.31)	45.83 (10.39)	$t(52)=1.03$, $p = .31$
how much you paid attention to the words	55.17 (9.40)	50.00 (10.43)	$t < 1$
the order of words (e.g., word 1, word 2, word 3, etc.)	44.83 (9.40)	41.67 (10.28)	$t < 1$
the other memory predictions you had made	31.03 (8.74)	29.17 (9.48)	$t < 1$
the amount of time between studying and testing	31.03 (8.74)	16.67 (7.77)	$t(52) = 1.13$, $p = .26$
how long the words were presented for	17.24 (7.14)	16.67 (7.77)	$t < 1$
how you think other people would make memory predictions	3.45 (3.45)	8.33 (5.76)	$t < 1$
other: _____	0.00 (0.00)	4.17 (4.17)	$t(52) = 1.10$, $p = .28$

Note. JOL condition ($n = 30$) and JORI condition ($n = 24$). *SE*s provided in parentheses.

Experiment 2 Discussion

Participants in Experiment 2 recalled significantly more concrete words than abstract words regardless of the judgment condition (i.e., JOL, JORI, Restudy Only). Additionally, higher JORIs and JOLs were provided for concrete words compared with abstract words (see also Pilot Study 2). Participants in all conditions chose to restudy

significantly more abstract words than concrete words. Critically, the framing of the metamemory judgment (i.e., JORI vs. JOL) influenced restudy selection. In particular, participants who made JORIs selected more items for restudy compared with participants who made JOLs and participants who only made restudy choices. Consistent with prior research (i.e., Finn, 2008), these data suggest that the framing of metacognitive judgments modifies study choices (I discuss this further in the General Discussion). If the framing of metacognitive judgments can be influential, perhaps the framing of the RI can be similarly influential. This notion was explored in Experiment 3 to determine if re-framing information about the RI itself would influence JOLs.

Experiment 3: Re-framed JOLs

The experiments reported thus far are not intended to support the abandonment of JOLs as a measure of metamemory. Rather, the primary argument is that metamemory awareness is a function of how a particular judgment is framed (cf. Finn, 2008). Thus, Experiment 3 extends the framing effects demonstrated in the decision making (Shafir, 1993; Tversky & Kahneman, 1981) and metacognitive literatures (e.g., Finn, 2008) to the framing of the specific RI. These literatures have demonstrated that differential framing (e.g., success vs. failure) can have a direct impact on later choices. Thus, framing may influence sensitivity to a manipulation such as RI.

In Experiment 3, participants made JOLs instead of JORIs. A standard JOL scale (0-100%) was used, replicating the JOL procedure from Experiment 2. JOLs were made anticipating one of two RIs (i.e., 5 min or 2 days). Of these RIs, three conditions existed such that participants were told to expect an RI of 5 min, 2 days, or 2,880 min (i.e., the equivalent of 2 days on a minute scale). Thus, comparisons can be made between

predictions for the 2 day and 2,880 min conditions holding the RI constant and only varying how the RI is presented to participants. Participants then waited the designated RI (5 min or 2 day RI) and recall was assessed.

It was expected that when the long RI was provided in minutes (i.e., 2,880 minutes), JOLs would be sensitive to RI, as compared with when the RI was given in larger units (i.e., 2 days). Specifically, JOLs were expected to be more conservative for the 2,880 min RI compared with the 2 day RI (cf., Finn, 2008). This pattern of results would demonstrate that reframing the RI (into 2,880 minutes instead of 2 days) can draw participants' attention to the interval, thus producing more realistic JOLs as compared to the standard framing (i.e., 2 days). An anchoring hypothesis (e.g., LeBoeuf & Shafir, 2009) would alternatively predict that JOLs would be higher for the 2,880 min RI compared with the other conditions.

Method

Participants

Eighty-four Colorado State University students (28 participants in each Framed RI condition; *M* age = 18.79, *SD* = 1.07) participated for course credit.

Design

Experiment 3 employed a 2 (RI: 5 min, 2 day) x 2 (Measure: JOL, recall) x 3 (Framed RI: 5 min, 2 days, or 2,880 min) mixed-factor design with Measure manipulated within-subjects and RI and Framed RI manipulated between-subjects.

Materials

Participants were presented with a 30 unrelated word pairs [equated on word frequency (*M* = 87.41, *SD* = 5.25: MRC Psycholinguist Database, 1987) and word length

($M = 5.25$, $SD = 1.23$)] in order to avoid floor effects in recall after the 2-day delay (e.g., GLACIER-SWEET). Word pairs were presented in a fixed-random order for each participant (excluding two primacy and two recency buffers). Two versions of the study list presented in a different order were created in order to account for item order-effects. No significant differences were found in recall based on item-order ($t < 1$) so this will not be discussed further.

Procedure

After providing informed consent, participants were presented with the 30 word pairs. Word pairs were presented one-at-a-time for 4 s each and participants were instructed to study each pair of words such that they would later be able to remember the second word of the pair (i.e., the target) if given the first word of the pair (i.e., the cue). Next, participants were presented with each cue for 5 s and wrote down their JOL using a standard JOL scale (0-100%) predicting the likelihood of later being able to recall the second word of the pair. A 500 ms inter-stimulus-interval (ISI) was included before the presentation of the next word pair following the JOL. Participants made JOLs anticipating one of three RIs: 5 min, 2 days, or 2,880 min (the equivalent of 2 days on a minute scale). Half of the participants received a 5 min filler task (writing down states of the United States) and the other half of participants returned to the lab 2 days after the study session. During the test phase each cue was presented in one of two fixed-randomized orders for 3 s followed by a 500 ms ISI prior to the presentation of the next item. Participants wrote down the second word of each pair.

Experiment 3 Results

Recall

A one-way ANOVA comparing recall performance between the 3 RI framing conditions (i.e., 5 min, 2 days, 2,880 min; see Figure 7) was conducted. As expected, recall performance reliably differed based on the RI framing condition, $F(2, 83) = 35.85$, $p = .0001$. Follow-up tests showed that recall performance was significantly better at the 5 min delay ($M = 58.14$, $SD = 18.25$, $p = .00001$) compared with the 2 day condition ($M = 24.14$, $SD = 15.23$), $t(54) = 7.35$, $p < .00001$, *Cohen's d* = 2.02, and the 2,880 min condition ($M = 26.64$, $SD = 15.76$), $t(54) = 6.91$, $p < .00001$, *Cohen's d* = 1.85. No significant differences were found between the 2 day and 2,880 min RI conditions, $t < 1$. Thus, the RI manipulation influenced memory performance with the 2-day producing poorer overall memory than the 5 min delay.

JOLs

A one-way ANOVA comparing JOLs between the three RI framing conditions (i.e., 5 min, 2 days, 2,880 min; see Figure 7) indicated that, contrary to expectations, the RI framing condition did not significantly impact JOLs ($F < 1$). Thus, the framing manipulation had a negligible influence on JOLs despite RI influencing memory performance.

Calibration

A 2 (Measure: JOL, recall) x 3 (RI framing: 5 min, 2 day, 2,880 min) mixed-factor ANOVA on mean JOLs and mean recall performance indicated that, overall, JOLs exceeded recall, $F(1, 81) = 125.48$, $\eta^2_p = .61$. A main effect of RI framing condition was also evident, $F(2, 81) = 20.30$, $\eta^2_p = .33$, such that combined JOLs and recall performance

for the 5 min RI condition were greater than both the 2 day and 2,880 min RI conditions. Critically, these main effects were qualified by a reliable Measure x RI framing interaction, $F(2, 81) = 16.53$, $\eta^2_p = .29$. Specifically, no significant difference was found between JOLs and recall performance for the 5 min condition, $t(27) = 1.49$, $p = .15$, *Cohen's d* = .40, while significantly higher JOLs compared with recall performance was found for both the 2 day, $t(27) = 10.46$, $p < .00001$, *Cohen's d* = 2.62, and 2,880 min, $t(27) = 9.45$, $p < .00001$, *Cohen's d* = 2.14, conditions. Thus, participants demonstrated the best calibration in the 5 min RI condition and were overconfident in the 2 day and 2,880 min RI conditions.

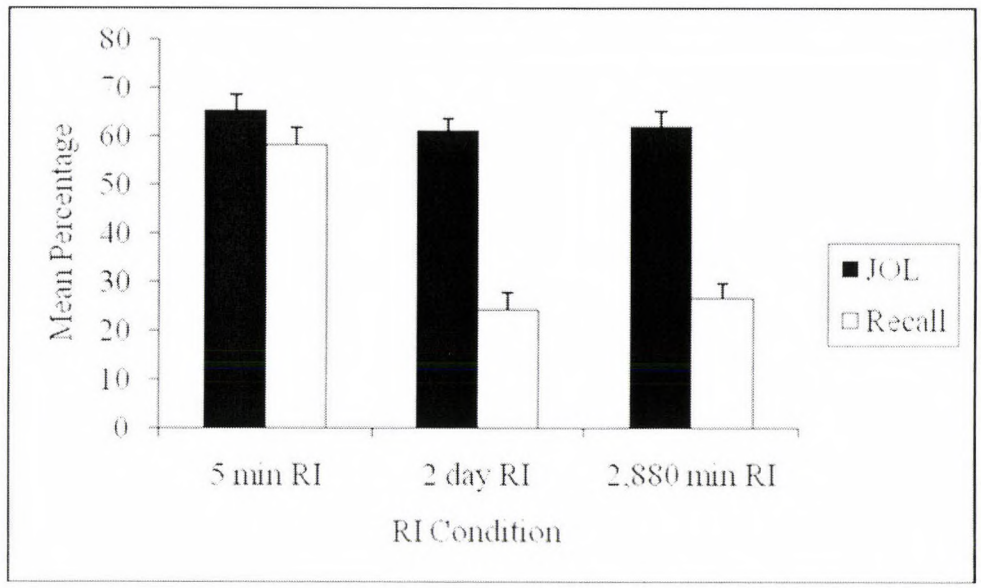

Figure 7. Calibration Data for Experiment 3.

Resolution

Gamma (*G*) correlations between JOLs and recall (see Table 8) were calculated for each RI framing condition. Three one-sample t-tests were conducted comparing each *G* correlation against chance (zero). *G* correlations for the 5 min RI did not reliably exceed zero ($t < 1$), whereas *G* for the 2 day condition approached significance, $t(27) =$

1.69, $p = .10$. Gamma correlations for the 2,880 min RI condition reliably exceeded zero, $t(27) = 2.69$, $p = .012$. A one-way ANOVA comparing *G* correlations between the three RI framing conditions indicated that there were no significant differences in *G* correlations between the three RI framing conditions ($F < 1$).

Table 8 *Experiment 3 Gamma (G) correlations*

	JOL-Recall *G*
2 day RI*	.115 (.361)
2,880 min RI*	.206 (.421)
5 min	.054 (.448)

*Note. SD*s provided in parentheses.
* significantly greater than zero, $p < .05$.

Experiment 3 Discussion

As expected, recall performance was significantly poorer after the long delay (i.e., 2 day and 2,880 min conditions) compared with the 5 min delay. However, contrary to expectations, JOLs were not lower when the 2 day RI was provided in minutes (i.e., the 2,880 min RI condition). In other words, calibration was equally poor for the 2 day and 2,880 min RI conditions. Thus, reframing the RI from an hour to a minute scale did not effectively improve calibration for long RIs. These data replicate prior work (Koriat et al., 2004) and indicate how challenging it is to make participants sensitive to RI.

General Discussion

Several novel patterns were found with the current set of experiments and each will be discussed in turn. I will first discuss participants' ability to adjust predictions based on time, and then consider the role of framing in study choices. Next, the current work will be interpreted within existing metacognitive frameworks. Finally, I will discuss limitations and extensions of this work and suggest some implications of these data in applied settings.

JORIs, JOLs, and Retention Interval (RI)

Overall, several notable trends in predictions were evident in the experiments reported. First, Pilot Study 1 (also replicated in Experiment 1) demonstrated that JORIs were not extreme estimates of memory ability. That is, on average, people believed they would remember some bit of information for approximately 15 min (Pilot Study 1) or 24 min (Experiment 1). In contrast, prior research with JOLs has demonstrated more exaggerated predictions, such as predictions of equivalent memory performance immediately, in 1 week (10,080 min), or even up to 1 year (483,840 min) (Carroll, et al., 1997; Koriat et al., 2004).

Second, JORIs in Experiment 1 demonstrated some sensitivity to RI. Specifically, predictions varied such that a greater percentage of JORIs exceeded a 30 sec RI, whereas the opposite pattern obtained when anticipating a 20 min RI (i.e., greater percentage of JORIs below a 20 min RI). In other words, when anticipating a short RI, participants made more predictions that exceeded the RI. In contrast, participants who were expecting

a long RI made more predictions that fell below the RI. If participants were entirely insensitive to RI then one might expect the distribution of responses above and below the RI to be equal regardless of the anticipated RI. Participants' JORIs thus demonstrated some sensitivity to an anticipated RI as predictions were sensitive to the expected interval between study and test. However, this interpretation must be treated cautiously as mean JORIs were not reliably shorter when anticipating a 20 min RI compared with participants who were expecting a 30 sec RI. Further, only 6 participants in the continuous 30 sec condition provided JORIs that fell below the 30 sec RI providing small cell sizes for that condition. Thus, it is challenging to ascertain from the current data whether participants are truly able to attend to RI when making JORIs.[4] As well, these intriguing patterns must also be tempered with the post-experiment survey results obtained in Experiment 2.

Experiment 2 found no differences in retrospective self-reported bases for judgments between the JOL and JORI conditions. In both conditions participants reported considering strategies and information about the individual words when predictions were made and few participants reported considering the RI (approximately 24%, or 14 out of 54 total participants). These results may not be surprising, as it is not always necessary for participants to have explicit knowledge of the factors that can influence metacognitive

[4] In order to more directly compare JORIs with JOLs, I extended Experiment 1 to a JOL condition. That is, I also collected data asking participants to make standard JOLs (0-100%) anticipating either a 30 sec or 20 min RI. Results demonstrated that JOLs did not reliably vary when participants were expecting a 30 sec interval ($M = 41.78\%$, $SD = 21.54\%$) or a 20 min interval ($M = 42.18\%$, $SD = 14.36\%$), $t < 1$. These data cannot be further categorized around the RI as JOLs are on a percent scale and not in minutes. Thus, while these data might indicate that participants are not sensitive to RI when making JOLs, these data cannot be more directly compared to the JORI data leaving this conclusion necessarily limited.

monitoring (cf. Nisbett & Wilson, 1977). Perhaps adjustments for certain manipulations, like RI, are implicit and unavailable to awareness. However, it should be noted that the self-reported bases for judgments were retrospective in nature allowing participants to reflect back on their predictions across the entire experiment. Perhaps when making item-by-item JORIs participants *do* have explicit knowledge of the influence of an RI, but this may be better assessed with procedures (e.g., concurrent verbal reports; cf. Ericsson & Simon, 1980) to measure explicit awareness while making predictions.

Finally, Experiment 3 revealed that JOLs did not vary based on the length of time between study and test even when the long interval was framed in terms of minutes (i.e., 2,880 min for a 2 day RI). These data support prior work demonstrating that participants' JOLs are insensitive to RI and extend this work to indicate that reframing the RI is not an effective manipulation to make the RI more salient.

Taken together, JORIs, in contrast with JOLs, provide a very different picture of participants' ability to incorporate time into their assessments of future memory performance. That is, JORIs, compared with JOLs, are more conservative estimates of memory performance over time, and potentially demonstrate some sensitivity to RI. There are two potential explanations of these data. First, JORIs, compared with JOLs, are perhaps a more effective measure to assess time-based predictions. That is, participants might be just as aware, or unaware, of the influence of time on memory regardless of monitoring performance with JOLs and JORIs, but perhaps JORIs might be a somewhat better measure to assess time-based awareness. (I will return to this idea in the Basis for Judgment section below.) This suggests that there may be a small benefit when using JORIs that could be attributed to the measure itself. Second, when making JORIs,

participants may consider different kinds of information. For example, participants may be more aware of time and thus make more reasonable predictions when using a minute scale. This explanation contradicts the retrospective self-report data. Specifically, there were no differences in the self-reported bases for judgment between the JOL and JORI conditions. Alternatively, having participants make JORIs may provide a context in which people are primed to consider different information. Similar to Finn (2008), the JORI scale might induce theories of forgetting bringing to mind instances when memory failed over longer time intervals. In this way, participants' estimates would be more conservative due to different information coming to mind but not because of an increased understanding of the negative influence a long RI can have on memory performance. This explanation is feasible even in light of the retrospective self-reports because people may not be explicitly considering the RI; rather they might be considering memory failure in a more general sense.

Framing and Study Choices

Experiment 2 provided a direct comparison between JOLs and JORIs and explored the relationship between each judgment and restudy choices. Results demonstrated several intriguing patterns. Of particular interest, participants in the JORI condition made more restudy selections than the JOL condition or the restudy only condition. The discrepancy-reduction model (e.g., Thiede & Dunlosky, 1999) posits an interaction between control and monitoring (i.e., JOL or JORI) processes such that efforts will be made to continue to learn material until the performance goal is likely to be reached (but see Ariel, Dunlosky, & Bailey, 2009; Metcalfe & Kornell, 2005, for alternatives). Thus, the JORI data could be interpreted as improved awareness (at least

implicitly) of the general discrepancy between learning and memory performance leading participants to more frequently seek to restudy items. This claim is challenging to support with the current data alone as it requires determining if changes in study choices did in fact produce differences in the perceived discrepancy between learning and performance (see the limitations and future directions section of the GD). It is also possible that making JORIs, as opposed to JOLs, produced a different goal for study. That is, one might imagine that if a different goal for study were evident between the JOL and JORI conditions, participants might vary the amount of effort needed to achieve the goal, which would produce different levels of recall performance. This suggestion would support prior work demonstrating that manipulations of a study goal (i.e., an easy goal versus a more challenging goal) modify study choices (e.g., Thiede & Dunlosky, 1999). Finally, it is also possible that JORIs may prime theories of forgetting (Finn, 2008). In this case, when making JORIs, people may be more inclined to select items for restudy simply because memory failure is primed to a greater extent than in the JOL condition.

Experiment 2 additionally demonstrated that the item-by-item relationship (i.e., resolution) between JORIs, recall, and restudy mirrored those of the JOL condition. These data indicate that making JORIs maintained the same relationship with other measures on an item-by-item basis, while putting participants in a context that produced additional restudy selections. This suggests that JORIs may be equally suitable to measure metacognitive monitoring as JOLs, while providing a context in which different choices are made. It should be noted that framing in general is not a "cure-all" that always improves the ability to incorporate time information into predictions. For example, in Experiment 3, the RI was framed in larger or more specific units of time (i.e.,

5 min, 2 day, or 2,880 min) and JOLs were used as the metacognitive measure. Contrary to expectations, RI framing had a negligible impact on JOLs while recall was directly related to the RI. These data indicate that reframing the RI does not necessarily influence JOLs.

In sum, when the metacognitive judgment was framed in terms of minutes, participants made different study choices than when the metacognitive judgment was framed in terms of percentages. These data suggest the critical role framing can have for metacognitive judgments and has clear empirical implications: Researchers should carefully consider *how* metacognition is being measured to assess metacognitive awareness.

Basis for Judgments

Koriat (1997) proposed a cue utilization framework in which metacognitive predictions are inferential in nature, meaning that predictions are based on inferences made from the information available to the rememberer. Specifically, Koriat distinguishes between three different types of cues: intrinsic, extrinsic and mnemonic. Intrinsic cues are relevant to the individual items to-be-remembered and refer to information that makes certain items seem more memorable than others. Item difficulty is an example of an intrinsic cue and participants typically use intrinsic information as a basis for JOLs (but see Tauber & Rhodes, in press). Extrinsic cues are specific to either the testing situation or to encoding techniques employed by participants. RI is an example of an extrinsic cue, and participants in general are poor at incorporating extrinsic uses into JOLs (e.g., Carroll, et al., 1997, Koriat, et al., 2004). Finally, mnemonic cues are

internal indices of memory performance which are typically gained through experience with the task or practice.

Further, with experience, people can switch their bases for judgment from an intrinsic to an extrinsic or mnemonic cue (e.g., Tauber & Rhodes, 2010; Tauber & Rhodes, in press). For example, Tauber and Rhodes (2010) had participants learn, make JOLs, and tested memory for either a list of 10 words or 100 words. Results demonstrated that participants were largely unaware of the influence of the amount of material to-be-remembered as JOLs did not vary between list length conditions, while memory performance was reliably lower in the 100-item condition. However, participants were able to adjust predictions to incorporate list length information when the cue switched from an extrinsic cue to a mnemonic cue. In particular, when list length was manipulated within-subjects, allowing participants two study-test trials with lists of varying length, predictions on the second study-test trial were more diagnostic of memory performance (Experiment 4). Thus, after experience with a prior list, participants were able to adjust predictions and attend to mnemonic information about list length.

The current data extend the cue utilization framework to JORIs. For example, Pilot Study 2 (also replicated in Experiment 2) revealed that participants were able to modify their JORIs to reflect manipulations that influence memory performance, in a similar manner previously demonstrated with JOLs (e.g., Koriat et al., 2004; Koriat & Ma'ayan, 2005). Participants appear to be equally able to attend to intrinsic cues with JORIs or JOLs. The extrinsic cue of RI was explored in Experiment 1 and Experiment 3. Experiment 1 revealed that participants who made JORIs were better able to attend to RI than participants who made JOLs. Further, Experiment 3 reframing the RI did not

improve awareness measured using JOLs. Thus, according to the cue-utilization framework, people may be better able to attend to the extrinsic cue of RI with JORIs than JOLs. Based on the current data alone it is unclear whether this difference is specific to RI or can be generalized to other extrinsic cues.

In contrast to the cue utilization framework, a direct access account (e.g., Arbuckle & Cuddy, 1969; Hart, 1965; Dunlosky & Nelson, 1994) suggests that metacognitive judgments are based on the strength of a memory trace such that higher JOLs are associated with strong memory traces while lower JOLs are associated with weaker memory traces. By this account, metacognitive errors could result when predictions need to be adjusted for an anticipated RI. Specifically, at the time of encoding, the memory trace for each item would be approximately equivalent (assuming item difficulty is not manipulated); however, these memory traces need to be qualified by the anticipated RI. Thus, if participants base their predictions on the strength of the memory trace alone, predictions might be prone to errors demonstrating insensitivity to an intervening RI. However, a direct access approach would not be an effective explanation for the framing effects obtained in the current data. In particular, if participants are accessing memory strength when predictions are made, then identical patterns should be observed for JORIs and JOLs such that predictions would not vary based on RI. That is, the framing of the judgment should have no impact on memory strength. Because the data reported in the current experiment indicated several major discrepancies between JOLs and JORIs (e.g., for restudy choice) I would suggest that they are best accommodated by an inferential framework such as the cue utilization approach.

Limitations & Future Directions

There are several limitations of the experiments reported, most of which could be addressed with additional research. For example, one limitation was already identified with the post-experiment survey in Experiment 2. In particular, it is unclear whether people do not have explicit knowledge during self-report of previously considering the RI, or if this knowledge could be better measured concurrently with the monitoring task. Thus, future research could employ a qualitative component during the study and prediction phase in order to remove the retrospective nature of the post-experiment survey employed in Experiment 2. It might also be prudent to employ think-aloud protocols during the monitoring portion of the task. This type of methodology is not commonly employed in metacognition, but may greatly inform the online bases for metacognitive predictions.

Additionally, in Experiment 2, it seemed that participants in the JORI condition made better control choices, as they volunteered more items for restudy compared with the JOL condition. However, because participants were not actually provided an opportunity for restudy, it cannot be yet determined whether these choices are indeed better. Thus, I am currently collecting additional data using the same design employed in Experiment 2, but allowing participants to restudy the items they selected for restudy. In this manner improvements in memory performance should accrue as a result of additional study. Further, if participants continue to select more items for restudy in the JORI condition, enhanced memory performance should result in this condition in comparison with the JOL condition. (Preliminary results are consistent with this prediction.)

An intriguing trend obtained in Experiment 2 was that it appears that participants are better able to attend to RI with JORIs than with JOLs. However, it is unclear whether this is a generalized quality of JORIs, such that JORIs are a more sensitive metacognitive measure of extrinsic cues, or if this benefit is specific to RI alone. Thus, additional research should compare JORIs and JOLs for other extrinsic cues such as presentation time or the number of times items are presented at study. If such data indicate larger adjustments in JORIs based on these cues, then JORIs may be a more sensitive measure of participants' metacognitive awareness of extrinsic information. In contrast, if no differences are found between JOLs and JORIs, this indicates that the benefits to using JORIs are specific to RI information.

A final limitation that could be addressed with additional research is evident from Experiment 3. Specifically, this experiment revealed that reframing the RI did not influence JOLs. However, it would be informative to investigate whether reframing the RI would have a contrasting influence on JORIs. Specifically, it might be that when the RI is reframed onto a minutes scale (i.e., 2,880 min rather than 2 days) participants making JORIs might demonstrate sensitivity to the reframed condition. It is also possible that participants would anchor their JORIs on the RI information, resulting in inappropriately long JORIs in the 2,880 min condition. These data would increase our understanding of the relationship between JORIs and JOLs as well as the contexts in which participants modify their predictions based on the metacognitive judgment.

A more general limitation of this work is that it is challenging to directly compare JORIs and JOLs because each measure, by definition, exists on a different scale (i.e., JOLs were measured with a 0-100% scale, whereas JORIs were measured with a 0-60

min scale). Thus, while resolution and control choices (e.g., restudy) can be compared between the two judgment conditions, calibration could not be assessed for JORIs. It is unclear how future research should be designed to address this limitation as this might be inherent with this type of measurement question. A similar practical limitation of the current experiments is that in each experiment JORIs were constrained to a 0-60 min scale. Thus, participants' JORIs might be conservative estimates of time simply because the scale was limited to a maximum prediction of 60 min. I have explored this limitation in more detail by collecting additional data asking participants make JORIs without providing any scale. In other words, participants (n = 34) made predictions of future memory performance in terms of time, but were free to provide any duration they wanted; all JORIs were then converted to a minute scale. These data indicated that JORIs (M = 6.77 min, SD = 20.65 min) were even shorter than when no scale was provided. Further the majority of participants (i.e., 85%, or 29 out of the 34 participants) chose to use second or minute scales with larger time intervals used rarely (i.e., hours, days).

Overall, there are several limitations with the current set of studies, many of which can be addressed with additional research. This work is an initial step towards a new measure of metacognition, JORIs, and additional work in this area is likely to be fruitful. Despite the limitations of these experiments there are several implications of this work to applied contexts.

Implications

These data have direct implications for any situation in which predictions of future performance or choices are based on time. For example, the planning fallacy is a widely documented error whereby important information is not incorporated into

predictions when planning for future events, leading to underestimates of the amount of time, effort, or money necessary to complete a project (e.g., Buehler, Griffin, Ross, 2002; Dunning, et al., 2004). For example, construction companies are constantly making projections of when a project will be completed and often fall prey to this error. If people think specifically in terms of time when these predictions are made, perhaps by thinking about how much time it will take to complete every step necessary to finish a task, then these predictions could be more realistic and reliable. These data also directly apply to students. Students often ask when to expect the next exam, and these data suggest that students may not be as insensitive to time as prior reports would suggest. It may be that students *are* adjusting their study habits and expectations for the exam based on time. Thus, it may prove especially valuable to ask poorly performing students to make performance predictions in terms of time in order to see improvements in study habits.

Summary & Conclusions

In sum, these experiments were intended to address two specific questions. First, does the metacognitive judgment impact assessments of monitoring and control (Pilot Study 1, Pilot Study 2, and Experiment 2)? The current data suggest that the metacognitive measure itself does need to be considered to accurately assess metacognition. For example, Experiment 2 demonstrated that the type of judgment made influenced later study choices, such that participants were more likely to choose to restudy items given JORIs compared with JOLs. Second, do people lack awareness of the influence of an intervening RI on memory performance (Experiment 1, Experiment 3)? It seems clear from the present experiments that people do not easily incorporate an intervening RI into their judgments, whether predicted with JOLs or JORIs. Further work

is needed to definitively ascertain whether participants indeed lack knowledge of RI, or if alternative methods are needed to better assess this knowledge.

References

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

Ariel, R., Dunlosky, J., & Bailey, H. (2009). Agenda-based regulation of study-time allocation: When agendas override item-based monitoring. *Journal of Experimental Psychology: General, 138,* 432-447.

Benjamin, A. S., & Diaz, M., (2008). Measurement of relative metamnemoic accuracy. In J. Dunlosky, & R. A. Bjork (Eds.), *Handbook of Metamemory and Memory* (pp. 73-94). New York, New York: Psychology Press.

Buehler, R., Griffin, D., & Ross, M. (2002). Inside the planning fallacy: The causes and consequences of optimistic time predictions. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and Biases* (pp. 250-270). New York, New York: Cambridge University Press.

Busey, T. A., Tunnicliff, J., Loftus, G. R., & Loftus, E. F. (2000). Accounts of the confidence accuracy relation in recognition memory. *Psychonomic Bulletin & Review, 7,* 26-48.

Carroll, M., Nelson, T. O., & Kirwan, A. (1997). Tradeoff of semantic relatedness and degree of overlearning: Differential effects on metamemory and on long-term retention. *Acta Psychologica, 95,* 239-253.

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

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

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

Dunning, D., Heath, C., & Suls, J. M. (2004). Flawed self-assessment implications for health, education, and the workplace. *Psychological science in the Public Interest, 5,* 69-106.

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review, 87,* 215-251.

Finn, B. (2008). Framing effects on metacognitive monitoring and control. *Memory & Cognition 36,* 813-821.

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

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

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

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

Koriat, A., Bjork, R. A., Sheffer, L., & Bar, S. K. (2004). Predicting one's own forgetting: The role of experience-based and theory-based processes. *Journal of Experimental Psychology: General, 133,* 643-656.

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

Koriat, A., Ma'ayan, H., & Nussinson, R. (2006). The intricate relationship between monitoring and control in metacognition: Lessons for the cause-and-effect relation between subjective experience and behavior. *Journal of Experimental Psychology: General, 135,* 36-69.

LeBoeuf, R. A., & Shafir, E. (2009). Anchoring on the "here" and "now" in time and distance judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35,* 81-93.

Masson, M. E., & Rotello, C. M. (2009). Sources of bias in the Goodman-Kruskal Gamma Coefficient measure of association: Implications for studies of metacognition processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35,* 509-527.

McNeil, B. J., Pauker, S. G., Saks, H. C., Jr., & Tversky, A. (1982). On the elicitation of preferences for alternative therapies. *New England Journal of Medicine, 306,* 1259-1262.

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

MRC Psycholinguistic Database. (1987). *Kucera-Francis Word Frequency.* Available from MRC Psycholinguistic Database: Machine Usable Dictionary. Version 2.0 from http://www.psy.uwa.edu.au/mrcdatabase/mrc2.html.

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., & Leonesio, R. J. (1988). Allocation of self-paced study time and the "labor-in vain" effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 676-686.

Nelson, T. O., & Narens, L. (1994). Why investigate metacognition? In J. Metcalfe, & A. P. Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we know: Verbal reports on mental processes. *Psychological Review, 84,* 231-259.

Paivio, A. (1966). Latency of verbal associates and imagery to noun stimuli as a function of abstractness and generality. *Canadian Journal of Psychology, 20,* 378-387.

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

Rhodes, M. G., & Castel, A. D. (2009). Metacognitive illusions for auditory information: Effects on monitoring and control. *Psychonomic Bulletin & Review.*

Schwarz, N. (1999). Self-report: How questions shape the answers. *American Psychologist, 54,* 93-105.

Shafir, E. (1993). Choosing versus rejecting: Why some options are both better and worse than others. *Memory & Cognition, 21,* 546-556.

Tauber, S. K., & Rhodes, M. G. (2010). Does the amount of material to-be-remembered influence JOLs? *Memory, 18,* 351-362.

Tauber, S. K., & Rhodes, M. G. (in press). Metacognitive errors contribute to the difficulty in proper name learning. *Memory.*

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.

Tversky, A., & Kahneman, D. (1974). Judgments under uncertainty: Heuristics and biases. *Science, 211,* 453-458.

Tversky, A., & Kahneman, D. (1981). Judgment under uncertainty: Heuristics and biases. *Science, 211,* 453-458.

Appendices

Appendix A

A 10 min and 20 min RI were originally proposed. However, the 10 min RI was replaced with a 30 sec RI because no differences in recall performance were evident between the 10 min and 20 min conditions.

Table 1A

Data from a 10 min RI condition collected for Experiment 1

	Continuous 10 min RI	Binary 10 min RI
Average Recall	23.13% (12.72%)	27.19% (13.26%)
Categorized Recall: More than the RI	28.97% (6.53%)	42.73% (4.54%)
Categorized Recall: Less than the RI	18.78% (3.18%)	16.47% (3.64%)
Average JORI	18.83min (10.98 min)	N/A
Frequency of Binary JORIs above RI	N/A	49.27% (21.26%)
Frequency of Binary JORIs below RI	N/A	50.72% (22.70%)
Categorized JORI: More than the RI	48.80% (27.04%)	50.72% (22.70%)
Categorized JORI: Less than the RI	51.20% (27.04%)	49.28% (22.70%)

*Note. SD*s in parentheses. $n = 16$ in each 10 min RI condition.

Appendix B

Experiment 2 Post-Experiment Survey.

Front of Survey:

(1) What information did you consider when you made your memory predictions? Please include everything that came to mind that influenced your predictions.

(2) Do you think your memory predictions were accurate? Yes or No. Why or Why not?

(3) If you did this task again, would you make different predictions? Yes or No. Why or Why not?

Back of the Survey:

Which of the following did you think about when you made your memory predictions?

__ if the words reminded you of something else, something relevant to your life

__ if you thought the words would be easy or difficult or remember

__ if you tried to repeat the words in order to remember them

__ the amount of time between studying and testing

__ the number of words you were learning overall

__ whether you would choose to restudy the words or not

__ how much you paid attention to the words

__ if you tried to related the words with other words you were learning

__ how long the words were presented for

__ the other memory predictions you had made

__ how you think other people would make memory predictions

__ the order of words (e.g., word 1, word 2, word 3, etc.)

__ other:

Appendix C

Sample open-ended responses from the post-experiment survey in response to the question:

What information did you consider when you made your memory predictions? Please include everything that came to mind that influenced your predictions.

JOL Condition

(1) "when I try to remember something I try to relate it to myself so that helped me to remember easier."

(2) "I didn't really consider any outside information except that I knew I wasn't going to be able to recall much more than half."

(3) "objects would be easier"

(4) "my general ability to remember things such as lists of words, etc."

(5) "I'm usually very good at remembering things"

(6) "If it was a word I used often or if I could put it into a sentence to remember later then that influenced my predictions"

JORI Condition

(1) "trying to remember the words that were closely related to one another made it easier."

(2) "I really just guessed, but predicted that I would be able to remember common simple words for longer"

(3) "I'm tired from a long day so I knew my memory may not be as sharp as normal"

(4) "I considered where the word was on the sheet (number that I put) and whether I thought I could remember without reviewing. Also associated them with other words to remember."

(5) "how long/complicated the word was. If it was something I could relate to. What number it was on the list.

(6) "If an image came to mind, I chose no on the study part and typically marked less time, if I could not relate the words in a chain I would mark to see them again"