DISSERTATION

AN INVESTIGATION OF THE BASIS OF JUDGMENTS OF REMEMBERING AND KNOWING (JORKS)

Submitted by

Nicholas C. Soderstrom

Department of Psychology

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Summer 2012

Doctoral Committee:

Advisor: Matthew G. Rhodes

Anne M. Cleary
Deana B. Davalos
Dawn Rickey

ABSTRACT

AN INVESTIGATION OF THE BASIS OF JUDGMENTS OF

REMEMBERING AND KNOWING (JORKS)

Previous research indicates that prospective metamemory accuracy can be improved if participants are asked to monitor whether contextual details will be remembered or not (i.e., judgments of remembering and knowing; JORKs), as opposed to monitoring confidence (i.e., judgments of learning; JOLs), an important finding given that accurate memory monitoring has been linked to effective learning. Three experiments investigated whether the advantage for JORK is due to these judgments being based more on retrieval processes than JOLs. Experiment 1 showed that JORKs resemble retrospective confidence judgments (RCJs)—judgments known to be based on retrieval processes—in some ways but not in others. Experiment 2 demonstrated that JORKs benefit less from a delay than JOLs when judgments are made under some circumstances but not others, and Experiment 3 showed that JORKs are less susceptible to a manipulation of encoding fluency than JOLs. Thus, overall, the results provide mixed support for the idea that JORKs are more reliant on retrieval processes than JOLs, reinforcing the need for future research on this topic.

TABLE OF CONTENTS

CHAPTER I: Introduction

Knowing what one knows, and what one can do, is critical in everyday life. A pilot, feeling unsure about landing in thunderstorm conditions, might choose to delay the landing until the weather clears; a surgeon, not feeling confident in performing triple bypass surgery, might request that a more experienced surgeon observe the procedure, assisting when needed; or a student, realizing that material for an upcoming exam is not mastered, might re-read the material or ask a classmate for help. In these cases, it is paramount that one's subjective assessment closely matches with objective performance in order to successfully perform a given task and to avoid potentially disastrous consequences (see Dunning, Heath, & Suls, 2004). Stated more succinctly, people act on their subjective experiences, a fact that underscores the importance of the validity of such subjective states. The current dissertation investigated the basis of a recently developed metacognitive judgment that has been shown to lead to very accurate memory predictions.

Research in *metacognition* is aimed at understanding both the monitoring (i.e., awareness) of one's own cognitive processes, and the cognitive and behavioral control related to this evaluation (for reviews, see Koriat, 2007; Metcalfe, 2000). Nelson and Narens (1990) highlighted the conceptual relationship between monitoring and control, stating that meta-level processes (i.e., those involved in self-reflection) oversee object-level processes (i.e., basic operations such as encoding), and, in turn, regulate object-level processes accordingly (but see Koriat, Ma'ayan, & Nussinson, 2006, for evidence that control processes can sometimes inform monitoring). It should be of no surprise that researchers in this area have argued that metacognitive research represents a fruitful area in which the elusive topic of consciousness may be studied empirically, bridging an abstract idea with concrete data (see Nelson, 1996).

Multiple subareas of metacognition have emerged, such as metacomprehension and skill assessment; however, most research has examined one's own knowledge about his or her memory processes, termed *metamemory*. Particular interests in metamemory research include the bases of subjective memory assessments, the validity of these assessments, and the degree to which they influence subsequent behavior. Revisiting the example of a student assessing whether material for an upcoming exam has been sufficiently studied, metamemory researchers might be concerned with the process(es) by which the student concludes that material has, or has not, been learned, and how well these predictions match with actual performance on the subsequent exam. They might also examine how awareness of learning informs subsequent study habits.

The accuracy of metamemory judgments can be measured in two ways. First, *calibration* (also termed *absolute accuracy*) refers to the correspondence between mean judgment value and mean performance value. Based on the discrepancy between these values, over- or under-confidence can be determined. For example, a student might anticipate a 90% on an exam but get a 70%, demonstrating over-confidence. *Resolution* (also termed *relative accuracy*) refers to the extent to which items given high predictions are associated with high performance, and items given low predictions are associated with low performance, and is most commonly measured by within-person gamma correlations (Nelson, 1984; but see Benjamin & Diaz, 2008; Masson & Rotello, 2009, for alternatives). For example, a student might believe that material from one chapter is learned better than material from another chapter; if the student is correct, resolution is high. The current dissertation focused on the efficacy of metamemory judgments as reflected by measures of resolution, as this measure provides the most purchase on the particular research questions at hand. However, calibration is also discussed when appropriate.

Predicting Memory Performance: Judgments of Learning

Although other metamemory judgments such as feeling of knowing (FOK; Hart, 1965) and retrospective confidence judgments (e.g., Koriat & Goldsmith, 1996; Kelley & Lindsay, 1993) have been explored, the most common method of investigating metamemory in recent years has been to elicit judgments of learning (JOLs). Participants making JOLs are asked during learning to assess the likelihood of remembering a particular item on a later test. That is, people are asked to *predict* their future memory performance, typically on a 0%-100% scale. After studying a word pair such as *DOG – SPOON*, for example, the participant might be asked, "On a scale from 0%-100%, what is the likelihood that you will later recall *SPOON* if presented with *DOG*?" JOLs can be made after each item in a study list or for a group of items (termed *aggregate JOLs*; e.g., "How many of the 30 word pairs do you think you will you remember on a later test").

Given that participants' JOLs are usually moderately predictive of future performance (both in terms of calibration and resolution), researchers have speculated on the bases for these judgments. Two dominant theoretical accounts have been put forward: the direct access view and the cue utilization view. The *direct-access view* argues that people directly access memory trace strength for each item during study and make their judgments accordingly (e.g., Cohen, Sandler, & Keglevich, 1991; Hart, 1967; Jang & Nelson, 2005). If the memory trace is perceived to be strong, a relatively high JOL will be given. Conversely, if the trace is weak, the JOL will be relatively low. As an alternative view, Koriat (1997) proposed the *cue-utilization approach*, which suggests that JOLs are based on a variety of cues or heuristics, some of which may be more valid in predicting memory performance than others—that is, judgments are inferential in nature.

One prediction made by the trace access view is that judgments should always parallel memory performance, because both JOLs and memory performance are based on trace strength. However, in many cases predictions are *not* diagnostic of performance (e.g., Benjamin, Bjork & Schwartz, 1998; Koriat & Bjork, 2005; Koriat & Goldsmith, 1996; Koriat, Bjork, Sheffer, & Bar, 2004; Mazzoni & Nelson, 1995; Rhodes & Castel, 2008; Soderstrom & McCabe, 2011). For example, Soderstrom and McCabe (2011) showed that people may take into account irrelevant information when predicting future memory performance. Participants studied items that were paired with numbers (ranging from 1 to 6), denoting the value of remembering each item on a later test, before making item-by-item JOLs. Interestingly, participants gave relatively higher JOLs for high value items, even in cases in which the value of the items was not known until *after* the item was studied. That is, although value had no impact on future recall performance—it came after the item, thereby preventing value-based encoding strategies—people believed it would have a substantial influence on future memory performance (see also Kassam, Gilbert, Swencionis, & Wilson, 2009). Likewise, Rhodes and Castel (2008) found that participants gave higher JOLs for words in a large font relative to those presented in a small font, whereas future memory performance did not differ as a function of font size. These findings pose serious problems for the direct access view, but can be easily accommodated by the cue utilization approach by suggesting that invalid cues (i.e., value and font size) informed participants' JOLs.

These examples highlight cases in which JOLs and performance do not match; however, there are instances when JOLs are highly predictive of later performance. Most notably, making JOLs after a delay significantly improves their relative accuracy, a finding termed the *delayed-JOL effect* (Koriat & Ma'ayan, 2005; Nelson & Dunlosky, 1991; for a review, see Rhodes & Tauber, 2011a). The leading candidate explanation for this effect is that delayed JOLs, unlike

immediate JOLs, are not contaminated by information from short-term memory (e.g., encoding fluency). Rather, delayed JOLs are based on information from long-term memory, such as the retrievability of the to-be-learned material. Because the later memory test is also based on information from long-term memory (i.e., retrievability), delayed JOLs tend to be quite accurate. This hypothesis has been corroborated by work showing that the delayed-JOL effect is limited to situations in which only the cue word in a pair (e.g., *DOG - ?* for the pair *DOG - SPOON*) is presented for the delayed JOL; delaying judgment is not helpful when the judgment is made in the presence of the cue and target (Connor, Dunlosky, & Hertzog, 1997; Dunlsoky & Nelson, 1992; 1997). Given the robustness of the delayed-JOL effect and the idea that these judgments are accurate because of their reliance on retrieval processes, one major goal of future research should be to find ways to encourage people to base *immediate* prospective judgments on retrieval cues. Indeed, Begg, Duft, Lalonde, Melnick, & Sanvito (1989) state that "Hindsight is the best foresight…" (p. 631) and Dougherty, Scheck, Nelson, and Narens, (2005) conclude that "JOLs should be made by assessing one's confidence in past retrieval…" (p. 1113).

Predicting Memory Performance: Judgments of Remembering and Knowing

Based on the idea that the predictive accuracy of metamemory judgments can be improved substantially when such judgments rely on retrieval processes, our lab has been exploring a novel method of making memory predictions based on the episodic memory experiences of *remembering* and *knowing* (Tulving, 1985). This section will describe the new approach in some detail, first summarizing the relevant background information regarding episodic memory, and then highlighting why this area represents an interesting and important avenue of future metamemory research.

For decades, the nature of episodic memory has been heavily debated, with some proposing that a single unidimensional strength continuum underlies memory (akin to the direct access view described previously for metamemory judgments), whereas others have advocated a dual-process view composed of recollection (i.e., remembering in the presence of contextual details) and familiarity (i.e., remembering in the absence of contextual details; see Yonelinas, 2002, for a review). For current purposes, the subjective experiences of remembering and knowing are important, regardless of whether these accurately map onto the processes of recollection and familiarity, respectively. *Remembering* is often defined as recollecting contextual details associated with an event (e.g., seeing someone in the supermarket and being able to recall the person's name and where you know them from), whereas *knowing* involves a sense that something is familiar in the absence of recollective details (e.g., being confident that you know the person in the supermarket, but without memory for the person's name or anything else about them).

A common method to assess these experiences in the laboratory is to employ the remember-know procedure (Tulving, 1985). In this procedure, participants are asked to retrospectively assess whether an item is *remembered* from an earlier study episode (i.e., contextual details associated with the item are remembered) or just *known* (i.e., the item is remembered in the absence of contextual details). In reviewing the literature on remembering and knowing, Gardiner (2002) presented evidence that remembering is readily dissociable from knowing. For example, levels-of-processing manipulations have large, positive effects on remembering, but little-to-no effect on knowing (Gardiner, Java, & Richardson-Klavehn, 1996). Conversely, study modality effects have been shown to affect knowing but not remembering (Gregg & Gardiner, 1994). In addition to providing evidence that various experimental

manipulations have differential effects on these subjective experiences, Gardiner also emphasizes that certain populations show selective impairments in remembering, such as older adults (see McCabe, Roediger, McDaniel, & Balota, 2009, for a review) and those with Asperger's syndrome (Bowler, Gardiner, & Grice, 2000). Thus, the subjective experiences associated with remembering and knowing are dissociable, presumably representing qualitatively different subjective states.

As noted previously, remember-know judgments are retrospective metamemory judgments, requiring people to assess whether recollective details accompany a particular memory or not. Our lab, however, has developed *prospective* remember-know judgments, called judgments of remembering and knowing (JORKs; McCabe & Sodertsrom, 2011). That is, unlike traditional remember-know judgments that are made retrospectively, asking participants to 'look back' to determine if an item is remembered or known, JORKs require people to 'look forward' to determine whether they think contextual details will be remembered or not on a later test. Compared to JOLs that ask people to assess the likelihood that a particular item will be remembered on a later test (usually on a 0%-100% scale), JORKs ask people if they will 'remember,' 'know,' or 'forget' an item later. Thus, JOLs, as a consequence of how they are worded (i.e., *Will you remember this in the future?*), seek to determine *what* information people think will be remembered later, whereas JORKs seek to determine *how* people think information will be remembered later.

According to our research, JORKs and JOLs seem to be qualitatively different prospective judgments, with JORKs leading to higher levels of resolution than JOLs in most cases (McCabe & Soderstrom, 2011). In our first experiment, participants studied single words, making memory predictions—either JORKs or JOLs—after each item. JORKs were made on a

3-point scale ("Will you Recollect, Know, or Forget this item?"), as were JOLs ("WILL Remember <1—2—3> WON'T Remember"). As predicted, JORKs showed greater relative accuracy than JOLs. Subsequent experiments showed that the greater metamemory accuracy of JORKs as compared to JOLs is robust, replicating across different instructions, materials, and outcome measures.

These data from McCabe and Soderstrom (2011) suggested that JORKs explicitly encouraged the use of distinct cues when making prospective metamemory judgments (cf, Koriat, 1997), and because these cues were also diagnostic of later retrieval, they were useful in predicting later memory performance. In other words, we argued that JORKs are, to a greater extent than JOLs, based on retrieval processes, stating that, "…JORKs can be considered judgments that focus participants' attention on information that is more closely associated with target retrievability than are immediate JOLs" (McCabe & Soderstrom, 2011). However, this conclusion was based on the finding that JORKs were more accurate than JOLs, not on evidence directly testing this retrieval-processes hypothesis. Thus, tests that manipulate and measure retrievability at the time JORKs are made are required to determine whether JORKs are, indeed, largely based on retrieval processes. Hence, the current dissertation focuses on directly testing the idea that JORKs rely, to a larger extent than JOLs, on retrieval processes and that this reliance on retrieval processes is responsible for their superior predictive accuracy compared to JOLs.

Overview of Experiments

Three experiments were conducted investigating the degree to which JORKs rely on retrieval processes. Experiment 1 assessed the extent to which JORKs resemble retrospective confidence judgments (RCJs) in predicting memory performance. Because RCJs are heavily

reliant on retrieval processes (i.e., they require the rememberer to reflect on the likelihood that retrieved information is accurate) they have been shown to have superior predictive accuracy compared to JOLs, which are often contaminated by encoding fluency (Busey, Tunnicliff, Loftus, & Loftus, 2000; Dougherty et al., 2005). Experiment 1 employed the pre-judgment recall and monitoring (PRAM) methodology (see Nelson, Narens, & Dunlosky, 2004), which solicits item recall immediately *before* metamemory judgments are made (the PRAM methodology is schematized in Figure 1). The PRAM procedure was beneficial for current purposes because it permits a direct measure of retrieval when the metamemory judgment is made. If JORKs are based largely on retrieval processes, then JORKs should more closely resemble RCJs than JOLs in this respect.

	Pair Study	*Pre-Judgment Recall*	*Metamemory Judgment*	*Final Recall*
Example:	TRAFFIC – SOAP	TRAFFIC - ?	TRAFFIC - ?	TRAFFIC - ?
Activity:	Study the item	Recall the target (i.e., recall SOAP)	Make JOL, JORK, or RCJ	Recall the target (i.e., recall SOAP)

Figure 1. PRAM (pre-judgment retrieval and monitoring) methodology (modeled after Nelson et al. 2004). JOL = judgment of learning; JORK = judgment of remembering and knowing; and RCJ = retrospective confidence judgment.

Experiment 2 investigated the basis of JORKs in a different way. Instead of using the PRAM methodology (Experiment 1), a delayed JOL procedure was employed. For half of the items, judgments (either JORKs or JOLs) were made immediately after the item was studied whereas, for the other half of the items, judgments were made after a delay. As previously mentioned, delayed JOLs are highly predictive of future performance because they are based more on retrieval processes compared to immediate JOLs (for a review, see Rhodes & Tauber, 2011a). Furthermore, the delayed-JOL effect is limited to situations in which only the cue (e.g., *DOG - ?* for the pair *DOG – SPOON*) is present at the time the JOL is solicited (Connor et al.,

1997; Dunlosky & Nelson, 1992; 1997), further substantiating the idea that engaging in retrieval is critical for JOLs to be highly accurate. That is, presenting only the cue at the time of JOL encourages participants to attempt retrieval of the target word before a JOL is made, whereas this is not the case when both the cue and target are presented. Thus, in addition to delaying judgments, Experiment 2 also manipulated whether judgments were solicited in the presence of the cue only (*DOG - ?*), or with the cue and target (*DOG – SPOON*). If immediate JORKs are largely based on retrieval processes, then JORKs should show a smaller delayed judgment effect (as measured by resolution) compared to JOLs. Furthermore, the type of cue used to solicit the judgment (cue-only vs. cue-target) should have less of an effect on resolution for JORKs than JOLs.

Finally, Experiment 3 assessed the retrieval-processes hypothesis in yet another way. If JORKs are largely reliant on retrieval processes, then JORKs should be relatively immune to manipulations of encoding fluency. Rhodes and Castel (2008) showed that the font size in which studied words were displayed influenced JOLs (relatively higher JOLs were given to large words) despite font size having no bearing on later performance. They suggested that this font size illusion reflected the perceived fluency of large relative to small items. JORKs should be less prone to this metacognitive illusion if these judgments are more reliant on retrieval processes (and consequently less reliant on encoding processes) compared to JOLs. In all, the current experiments directly examine the bases for JORKs and thus provide evidence on whether their accuracy reflects reliance on retrieval processes.

CHAPTER II: Experiment 1

Experiment 1 focused on whether JORKs more closely resemble RCJs than JOLs in regards to predicting memory performance. Busey et al. (2000) investigated the degree to which prospective and retrospective confidence ratings (JOLs and RCJs, respectively), and recognition performance are based on the same information. Using state trace analyses, they observed that JOLs and recognition judgments were based on different information, whereas RCJs and recognition judgments were based on the same information—namely, retrieval success. Dougherty et al. (2005) provided similar results using the PRAM procedure (introduced by Nelson et al., 2004). In the PRAM procedure, participants study items, engage in pre-judgment recall for each item, and then make their respective metamemory judgments (see Figure 1). The critical phase of the procedure for current purposes is pre-judgment recall, in which participants are required to make a recall attempt *before* a metamemory judgment is made—in this case, before either a JOL or RCJ. This allows an assessment of the degree to which both judgments are based on retrieval processes. In two experiments, Dougherty et al. found that RCJs were influenced by pre-judgment recall to a greater extent than JOLs, as measured by gamma correlations between pre-judgment recall and judgment magnitude. Moreover, as indexed by the gamma correlation between the metamemory judgment and final recall, RCJs were better predictors of future performance than JOLs. This is striking given that JOLs explicitly ask one to predict *future* performance, whereas RCJs ask one to assess *past* performance.

Both Busey et al. (2000) and Dougherty et al. (2005) provide evidence that RCJs rely, to a greater extent than JOLs, on retrieval processes, and that this reliance is responsible for RCJs' superior accuracy in predicting memory performance. Experiment 1 sought to replicate those findings—specifically, those reported by Dougherty et al. using the PRAM methodology. More

importantly, one group of participants also made JORKs. This manipulation permitted me to determine whether JORKs are also heavily reliant on retrieval processes. If this is true, then JORKs should resemble RCJs as revealed by a host of analyses—most notably, gamma correlations between pre-judgment recall and judgment magnitude, and between judgment magnitude and final recall.

Method

Participants

One-hundred sixty-eight undergraduates from Colorado State University took part in this study and received course credit for their participation. Mean age was 18.86 years and 64% of participants were Female. There were 56 participants in each of the three judgment conditions (JOL, JORK, and RCJ). Participants were tested individually.

Materials

Stimuli included 52 unrelated word pairs (e.g., *TRAFFIC - SOAP*; taken from Castel, McCabe, Roediger, & Heitman, 2007), four of which served as buffer items. Thus, responses for 48 of the word pairs were analyzed. The presentation of all stimuli and the recording of responses were done on Dell PC computers programmed with E-Prime software.

Design and Procedure

The PRAM methodology was employed (see Figure 1). Judgment type (JOL, JORK, RCJ) was manipulated between-subjects.

Pair Study. Participants were instructed to study word pairs (e.g., *TRAFFIC – SOAP*), such that the second word in each pair (i.e., the target; *SOAP*) could be recalled when prompted with the first word (i.e., the cue; *TRAFFIC*). Altogether, 52 word pairs were studied one at a time at a 5 s rate (500ms interstimulus interval) with the first two pairs and the last two pairs

serving as buffers. These buffers remained constant for each participant and were excluded from analyses. The remaining 48 pairs were presented in 8 blocks. Each block consisted of 6 word pairs chosen randomly.

Pre-Judgment Recall. Following each study block, participants engaged in a pre-judgment recall phase in which the cue word of each of the 6 pairs studied in the previous study block was presented. These cue words were presented randomly with the restriction that the first 3 cue words corresponded to the first 3 pairs studied in that block, and the last 3 cue words corresponded to the last 3 pairs studied in that block (this was done to ensure that all of the pairs had the same delay between study and pre-judgment recall). After each cue was presented, participants typed the target word corresponding to that cue into the computer. The pre-recall phase was self-paced and terminated when the participant pressed the ENTER key to submit their response.

Metamemory Judgments. After each pre-judgment retrieval attempt, participants were prompted for their metamemory judgment when given the cue word. Those in the JOL condition were asked to assess the likelihood, on a scale from 1 (*very likely*) to 3 (*not likely at all*), that the target word would be recalled on a later test when given the cue word.[1] Those in the JORK remembered. Like JOLs, JORKs were made on a 3-point scale; however these judgments were made based on the following categories: *1 = Recollect, 2 = Know,* and *3 = Forget*. (Note that participants in the JORK condition were read instructions on what constitutes these responses prior to the start of the study phase [see Appendix for these instructions].) Finally, those in the

---

[1]As stated in the Introduction, JOLs are typically made on a 0% - 100% scale; however, given that JORKs are made a 3-point scale (*Recollect, Know, or Forget?*), JOLs also needed to be measured on a 3-point scale for the two judgments to be directly comparable.

condition were asked to assess whether contextual details associated with the word pair would be RCJ condition assessed their confidence in their just-provided answers for the cue word on a scale from 1 (*definitely correct*) to 3 (*definitely incorrect*). All metamemory judgments were self-paced and were made by pressing the 1-3 keys on the keyboard. Furthermore, participants were encouraged to use the entire scale when making their judgments.

Final Recall. Immediately following the 8 blocks of studying, pre-judgment recall, and metamemory judgments, participants engaged in a brief filler task (approximately 2 min) in which a demographic questionnaire was completed (asking for age, sex, etc.). Immediately following the filler task, participants completed a cued recall test, for which two different random-fixed order paper-and-pencil test sheets were created for purposes of counterbalancing. Half (24) of the studied cue words were printed on the left side of these sheets; the other half of the cues were printed on the right side. To the right side of each cued word, a blank space was provided in which recall of its corresponding target word was attempted (e.g., *DOG* - _____). Participants were told that either guessing or leaving the space blank were acceptable responses if the target word could not be remembered. When finished, participants placed their writing utensils atop their recall sheets to indicate to the experimenter that they were finished, and were then given debriefing forms and excused from the experiment. The entire experiment took approximately 45 min.

Results

Given that measures of resolution are of primary interest Experiment 1, those analyses are reported first, followed by supplementary analyses. *F* values, mean squared errors (*MSEs*), and effect sizes are reported for statistical tests in which *F* values were greater than 1. The alpha level for statistical tests was .05 unless otherwise noted.

Resolution

To determine the extent to which JORKs rely on retrieval processes, and whether these judgments closely resemble RCJs in this respect, a number of gamma correlations were calculated. Gamma measures the extent to which items given certain values on a given dimension (e.g., memory predictions) are associated with those same values on another dimension (e.g., memory performance), and is denoted with a correlation coefficient ranging from -1.0 to +1.0. As shown in Figure 2, gammas between (1) judgment and pre-judgment recall; (2) judgment and pre-judgment recall latency; (3) judgment and final recall; and (4) pre-judgment recall and final recall were calculated for all judgment conditions (JOLs, JORKs, and RCJs). Then, one-way analyses of variance (ANOVAs) were conducted on each of the gammas for the three judgment types.

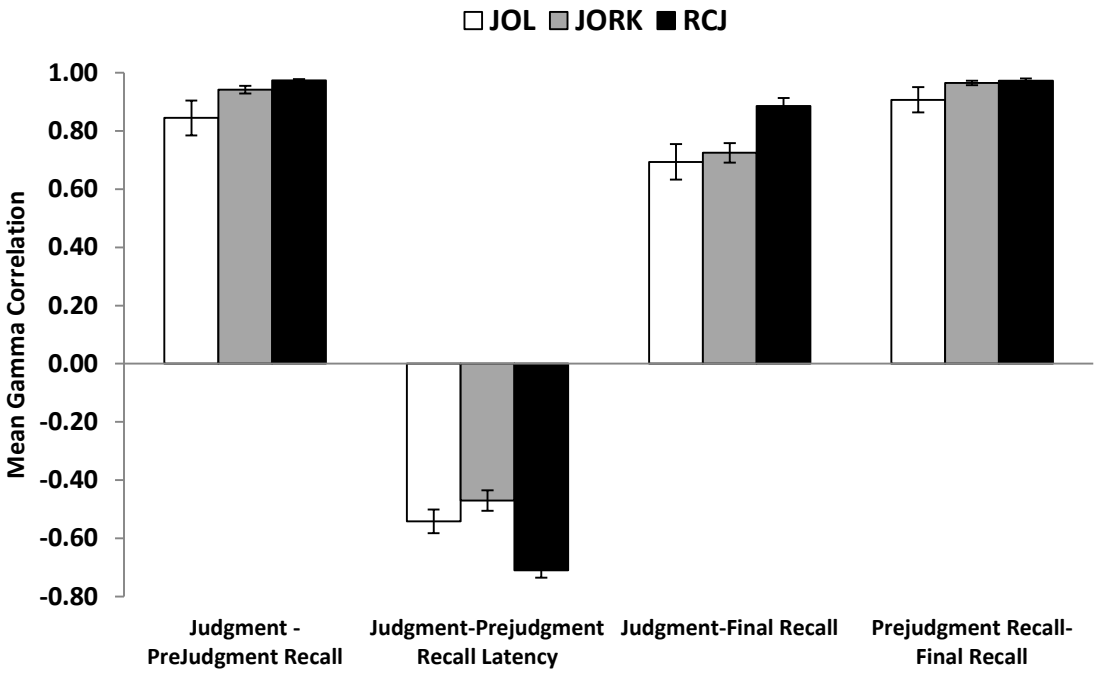


*Figure 2. M*ean gamma correlations for JOLs (judgments of learning), JORKs (judgments of remembering and knowing), and RCJs (retrospective confidence judgments) in Experiment 1. Error bars represent standard errors of the mean.

Turning first to gammas between metacognitive judgment and pre-judgment recall, a main effect was found, $F(2, 165) = 4.18$, $MSE = .06$, $\eta_p^2 = .05$. Post-hoc tests showed that RCJs were associated with the highest gammas ($G = .97$; $SE = .01$), exceeding both JOLs ($G = .84$; $SE = .06$) and JORKs ($G = .94$; $SE = .01$), $F(1, 110) = 5.37$, $MSE = .09$, $\eta_p^2 = .05$ and $F(1, 110) = 5.68$, $MSE = .01$, $\eta_p^2 = .05$, respectively. Furthermore, although not statistically reliable, gammas for JORKs were marginally higher than for JOLs, $F(1, 110) = 2.96$, $MSE = .09$, $\eta_p^2 = .03$, $p = .09$. These correlations suggest that pre-judgment retrieval informed RCJs the most, followed by JORKs, and then JOLs.

For gammas between metacognitive judgment and pre-judgment recall latency, a main effect was again found, $F(2, 165) = 15.28$, $MSE = .06$, $\eta_p^2 = .16$. Post-hoc tests revealed that this main effect was driven by RCJs showing lower gammas ($G = -.71$; $SE = .03$) than both JOLs ($G = -.54$; $SE = .04$) and JORKs ($G = -.47$; $SE = .04$), $F(1, 110) = 14.68$, $MSE = .06$, $\eta_p^2 = .12$ and $F(1, 110) = 36.37$, $MSE = .05$, $\eta_p^2 = .25$, respectively. No difference was found comparing JOLs to JORKs, $F(1, 110) = 2.09$, $MSE = .07$, $\eta_p^2 = .02$, $p = .15$. Thus, the time it took participants to complete pre-judgment recall attempts influenced RCJs the most, followed by JORKs and JOLs. All judgments, however, showed negative gamma correlations with pre-judgment recall latency, suggesting that judgment magnitude decreased as the time it took participants to recall the item increased.

Turning next to gammas between metacognitive judgment and final recall, a main effect was found, $F(2, 165) = 6.61$, $MSE = .09$, $\eta_p^2 = .07$, which was again driven by RCJs showing higher gammas ($G = .89$; $SE = .03$) than both JOLs ($G = .69$; $SE = .06$) and JORKs ($G = .72$; $SE = .03$), $F(1, 110) = 9.60$, $MSE = .11$, $\eta_p^2 = .08$ and $F(1, 110) = 16.02$, $MSE = .05$, $\eta_p^2 = .13$, respectively. No difference was found comparing JOLs to JORKs, $F < 1$. This indicates that

RCJs were the best at predicting future recall, followed by JORKs and JOLs, which did not differ in this regard. Finally, for gammas between pre-judgment recall and final recall, no main effect was found, $F(2, 165) = 2.27$, $MSE = .03$, $\eta_p^2 = .03$, $p = .11$. Thus, as expected, recalling an item at $Time_1$ (pre-judgment recall) predicted recall at $Time_2$ (final recall) equally as well for RCJs, JORKs, and JOLs.

Judgment Magnitude, Pre-Judgment Recall, and Final Recall as a Function of Judgment Type

My primary interest was in resolution; however, additional analyses were conducted to further examine the possibility that JORKs are based to a larger extent on retrieval processes than are JOLs. As shown in Table 1, judgment magnitude, pre-judgment recall, and final recall were examined as a function of judgment type. Turning first to mean judgment magnitude, a

Table 1
Judgment Magnitude, Pre-Judgment Recall, and Final Recall as a Function of Judgment Type in Experiment 1

	JOLs	JORKs	RCJs
Judgment Magnitude	1.78 (.38)	2.02 (.36)	1.68 (.39)
Pre-Judgment Recall	0.71 (.21)	0.68 (.20)	0.67 (.21)
Final Recall	0.56 (.24)	0.53 (.22)	0.50 (.23)

Note: The judgment scales for JOLs, JORKs, and RCJs were *1*=high confidence/remember, *2*=medium confidence/know, *3*=low confidence/forget. Standard deviations are reported in parentheses.

main effect was found, $F(2, 165) = 12.43$, $MSE = .14$, $\eta_p^2 = .13$. This was driven by JORKs showing a lower mean magnitude than both JOLs, $F(1, 110) = 11.89$, $MSE = .14$, $\eta_p^2 = .10$, and RCJs, $F(1, 110) = 23.81$, $MSE = .14$, $\eta_p^2 = .18$. (Note that the judgment scales were such that a "1" was the highest magnitude judgment.) No difference in judgment magnitude was found between JOLs and RCJs, $F(1, 110) = 2.10$, $MSE = .15$, $\eta_p^2 = .02$, $p = .15$. Finally, no differences were found in pre-judgment recall or final recall as a function of judgment type, F's < 1.

The between-condition differences in judgment magnitude were investigated further by examining response distributions across each judgment type (see Table 2). This was done by

Table 2
Distribution of Responses for JOLs, JORKs, and RCJs in Experiment 1

	JOLs	JORKs	RCJs
1/Remember	.50 (.25)	.35 (.19)	.57 (.23)
2/Know	.23 (.19)	.27 (.15)	.18 (.12)
3/Forget	.27 (.17)	.37 (.20)	.25 (.18)

Note: The judgment scales for JOLs, JORKs, and RCJs were *1*=high confidence/remember, *2*=medium confidence/know, *3*=low confidence/forget. Standard deviations are reported in parentheses.

conducting a 3 (Judgment Type: JOL, JORK, RCJ) x 3 (Response: 1/Remember, 2/Know, 3/Forget) mixed-model ANOVA. Note that because the proportion of study responses sum to 1 for each judgment type, the main effect of Judgment Type could not be calculated. Of greatest interest was the potential main effect of Response and the interaction. A main effect of Response was found, $F(2, 165) = 37.33$, $MSE = .07$, $\eta_p^2 = .19$; however, this was qualified by a Judgment Type by Response interaction, $F(2,165) = 12.62$, $MSE = .07$, $\eta_p^2 = .13$. Follow-up analyses were conducted to unpack this interaction. For the highest response of '1/Remember,' a main effect was found across judgment type, $F(2, 165) = 14.24$, $MSE = .05$, $\eta_p^2 = .15$. JORKs were associated with a lower proportion of items given the highest response compared to JOLs, $F(1, 110) = 11.86$, $MSE = .05$, $\eta_p^2 = .10$ and RCJs, $F(1, 110) = 31.25$, $MSE = .04$, $\eta_p^2 = .22$. No difference was found between JOLs and RCJs in the proportion of items given the highest response, $F(1, 110) = 2.99$, $MSE = .06$, $\eta_p^2 = .03$, $p = .09$. For items given the intermediate response of '2/Know,' a main effect was found, $F(2, 165) = 5.51$, $MSE = .02$, $\eta_p^2 = .06$. There were more intermediate JORKs than intermediate RCJs, $F(1, 110) = 14.44$, $MSE = .02$, $\eta_p^2 = .12$; however, JORKs did not differ from JOLs in this respect, $F(1, 110) = 1.58$, $MSE = .03$, $\eta_p^2 = .01$, $p = 2.21$. Also, JOLs and RCJs did not differ in terms of intermediate responses, although this was nearly the case, $F(1, 110) = 3.56$, $MSE = .03$, $\eta_p^2 = .03$, $p = .06$. Finally, for the lowest response of '3/Forget,' a main effect was found, $F(2, 165) = 7.56$, $MSE = .03$, $\eta_p^2 = .08$. JORKs

were associated with a higher proportion of items given the lowest response compared to both JOLs, $F(1, 110) = 9.07$, $MSE = .03$, $\eta_p^2 = .08$, and RCJs, $F(1, 110) = 12.19$, $MSE = .04$, $\eta_p^2 = .10$. JOLs did not differ from RCJs in this respect, $F < 1$.

Discussion

In Experiment 1, participants were tested using the PRAM methodology and made one of three metamemory judgments: JOLs, JORKs, or RCJs. One critical component of the PRAM methodology is that it elicits recall attempts *before* metamemory judgments, thus permitting a determination of the extent to which judgments rely on this pre-judgment recall. Consistent with the idea that JORKs are more reliant on retrieval processes than are JOLs, the gamma correlation between judgments and pre-judgment recall was higher for JORKs than JOLs (although not quite statistically reliable; see Figure 2). However, other gamma correlations—specifically, between judgments and prejudgment recall latency, and between judgments and final recall—were inconsistent with my predictions. I anticipated that these gammas would be similar for JORKs and RCJs, but these results indicated that JORKs more closely resembled JOLs. Finally, JORKs were associated with a lower mean judgment magnitude compared to both JOLs and RCJs (see Table 1), a difference driven by participants in the JORK condition giving relatively fewer high-magnitude responses and relatively more low-magnitude responses (see Table 2). Overall, then, the results of Experiment 1 provided mixed support for the idea that JORKs rely, to a larger extent than JOLs, on retrieval processes. Experiments 2 and 3 investigated this idea in different ways.

CHAPTER III: Experiment 2

Experiment 2 employed a standard delayed JOL methodology to further explore the possibility that JORKs are more reliant on retrieval processes than JOLs. Specifically, word pairs were studied and given either immediate or delayed judgments in the presence of either the cue only (e.g., *DOG-?*) or the cue and target (e.g., *DOG – SPOON*). As noted previously, the delayed-JOL effect is much larger for cue-only JOLs compared to cue-target JOLs, presumably because cue-only JOLs encourage participants to rely to greater extent on retrieval processes than cue-target JOLs (e.g., Dunlosky & Nelson, 1992; see Rhodes & Tauber, 2011a). Thus, I predicted that immediate JORKs should resemble delayed JOLs, or at least that the delayed judgment effect should be weaker for JORKs compared to JOLs. Furthermore, I predicted that the type of cue used at the time of the delayed judgment (cue-only vs. cue-target) should matter less for JORKs than JOLs. Again, both of these predictions are predicated on the idea that immediate JORKs are based, to a larger extent than immediate JOLs, on retrieval processes. Because delayed JOLs are also based on retrieval processes, their similarity to immediate JORKs should be evident.

Method

Participants

One-hundred forty-four undergraduates from Colorado State University participated in this experiment for course credit. Mean age was 19.22 years and 70% of participants were Female. There were 36 participants in each of the four between-subjects conditions (JOLs cue-only; JOLs cue-target; JORKs cue-only; and JORKs cue-target). Participants were tested in small groups of up to three individuals, each on their own computer.

Materials

The 48 unrelated word pairs were the same as those used in Experiment 1. The presentation of all stimuli and the recording of responses were done on Dell PC computers programmed with E-Prime software.

Design and Procedure

A standard delayed-JOL methodology was employed in which immediate and delayed judgments were made for both JOLs and JORKs. The type of cue (i.e., cue-only, cue-target) used when judgments were solicited was also manipulated. Thus, a 2 (Judgment Type: JOL vs. JORK) x 2 (Judgment Timing: immediate vs. delayed) x 2 (Cue Type: cue-only vs. cue-target) mixed-factor design was used. Both Judgment Type and Cue Type was manipulated between-subjects, whereas Judgment Timing was a within-subjects variable.

Participants studied word pairs one at a time for 4 s each with a 500ms interstimulus interval. JOLs or JORKs (depending on the condition) were made either immediately after studying the pair or following a delay. The study phase consisted of 2 blocks. In each block, 24 pairs were designated to receive either an immediate or delayed judgment (12 of each randomly presented). For those designated to receive an immediate judgment, the JOL or JORK was made immediately after the pair had been studied. For those pairs designated to receive delayed judgments, the JOL or JORK was not solicited until all of the immediate judgments for that block had been made. For example, once all the pairs in block 1 had been studied (those designated to receive immediate and delayed judgments) and after all of the immediate judgments had been given, delayed judgments were solicited corresponding to those pairs in block 1 that were not given immediate judgments. On average, approximately 2.5 min elapsed in a block before delayed judgments were made. The presentation order of delayed items was randomized. Furthermore, whether an item was designated an immediate or delayed judgment

was counterbalanced, such that those items given immediate judgments by a particular participant were given delayed judgments by the next participant, and vice versa.

In addition to the judgment-timing manipulation, the type of cue used at the time of the judgment was also manipulated; however, this manipulation was a between-subjects variable. Thus, for those participants in the cue-only condition (for either JOLs or JORKs), only the first word of each pair (e.g., *DOG - ?*) was presented at the time of the metamemory judgments, and this was true regardless of the timing of the judgment (i.e., immediate or delayed). For those in the cue-target conditions, however, both the cue and target of each pair (e.g., *DOG – SPOON*) was presented when the metamemory judgments were solicited. Instructions for JOLs and JORKs were identical to those used in Experiment 1. Participants were randomly assigned to one of the four between-subjects conditions: (1) JOLs with cue-only, (2) JOLs with cue-target, (3) JORKs with cue-only, and (4) JORKs with cue-target.

Immediately following the study phase, participants engaged in a brief filler task (approximately 2 min) in which they were required to fill out a demographic questionnaire (asking for age, sex, etc.). Finally, participants engaged in a paper-and-pencil final cued recall test in which each target word was to be recalled given its corresponding cue word (e.g., *DOG* - _____). Two random-fixed order test sheets were used for counterbalancing purposes in which half of the cue words were presented on the left side and the other half were presented on the right side. Participants were told that guessing or leaving a space blank were acceptable responses if the target word could not be remembered. When finished, participants placed their writing utensils on top of their recall sheets to indicate to the experimenter that they were finished. Once all participants completed the test, they were given debriefing forms and excused from the experiment.

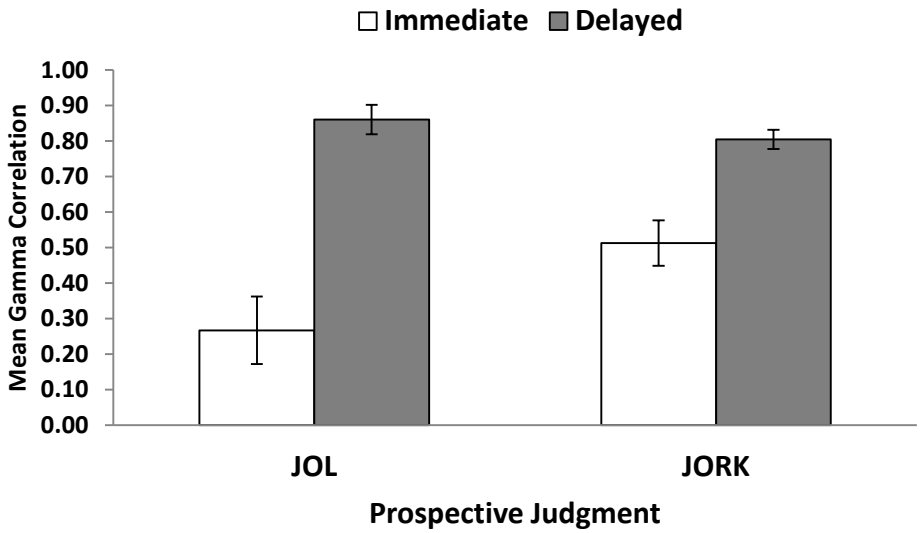
Results

Resolution

The primary focus of Experiment 2 (like Experiment 1) is on the relative accuracy of the metamemory judgments. Therefore, gamma correlations with final recall were calculated for each type of judgment (JOLs, JORKs) as a function of judgment timing (immediate, delayed) and cue type (cue-only, cue-target). Then, a 2 (Judgment Type: JOLs, JORKs) x 2 (Judgment Timing: immediate, delayed) x 2 (Cue Type: cue-only, cue-target) mixed-factor ANOVA was conducted on these gammas. These data are presented in Figure 3a and 3b. There was no main effect of Judgment Type ($F < 1$), but the main effect of Cue Type was reliable, $F(1, 125) = 10.67$, $MSE = 1.58$, $\eta_p^2 = .08$, indicating that cue-only gammas were greater than cue-target gammas. The Judgment Type by Cue Type interaction was not reliable, $F(1, 125) = 2.92$, $MSE = .43$, $\eta_p^2 = .02$, $p = .10$. Furthermore, the main effect of Judgment Timing was reliable, $F(1, 125) = 31.83$, $MSE = 3.39$, $\eta_p^2 = .20$, indicating that delayed gammas were greater than immediate gammas. Both the Judgment Type by Judgment Timing interaction, $F(1, 125) = 9.04$, $MSE = .96$, $\eta_p^2 = .07$, and the Cue Type by Judgment Timing interaction, $F(1, 125) = 24.69$, $MSE = 2.63$, $\eta_p^2 = .17$, were reliable. Thus, the timing manipulation (immediate vs. delayed) had a different impact on gammas for JOLs and JORKs, and also affected gammas differently depending on whether judgments were made in the presence or absence of the target word. The three-way interaction between Judgment Type, Judgment Timing, and Cue Type was not significant, $F(1, 125) = 1.26$, $MSE = .13$, $\eta_p^2 = .01$, $p = .26$.

To unpack the differential effects that Judgment Timing and Cue Type had on gamma correlations for JOLs and JORKs, separate ANOVAs were conducted for cue-only and cue-target items. Turning first to cue-only items (see Figure 3a), a 2 (Judgment Type: JOLs, JORKs)

(a) Cue-Only Items



(b) Cue-Target Items

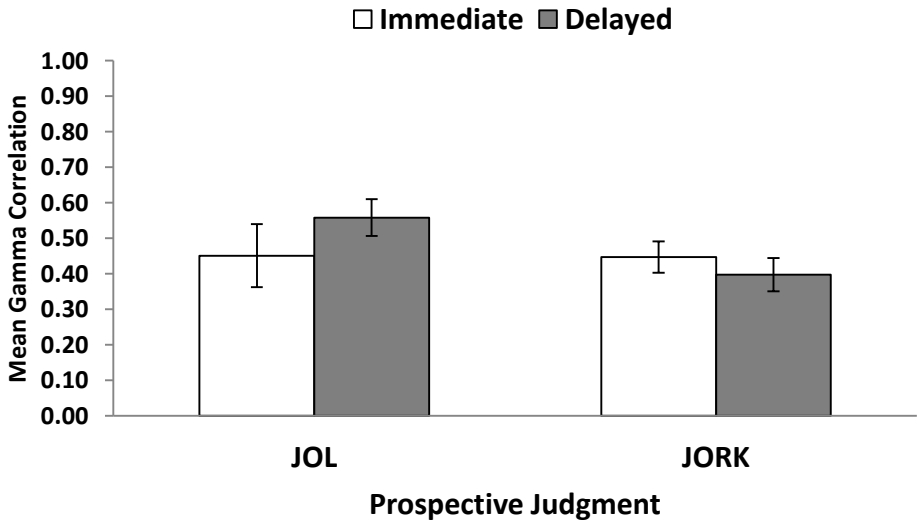


Figure 3. JOL-Recall and JORK-Recall gamma correlations as a function of judgment timing for (a) cue-only items and (b) cue-target items for Experiment 2. Error bars represent standard error.

x 2 (Judgment Timing: immediate, delayed) mixed-model ANOVA was conducted. There was no effect of Judgment Type, $F(1, 59) = 2.14$, $MSE = .31$, $\eta_p^2 = .04$, $p = .15$, but a main effect of Judgment Timing was found, $F(1, 59) = 52.86$, $MSE = 5.69$, $\eta_p^2 = .47$, indicating that delayed gammas were greater than immediate gammas for both JOLs and JORKs. However, this was

qualified by a Judgment Type by Judgment Timing interaction, $F(1, 59) = 8.00$, $MSE = .86$, $\eta_p^2 = .12$. Follow-up *t*-tests confirmed that delaying judgments improved resolution for both JOLs, $t(29) = 6.60$, $p < .001$, $d = 1.56$, and JORKs, $t(30) = 3.65$, $p < .01$, $d = 1.00$. Furthermore, gammas for immediate JORKs ($G = .51$, $SE = .06$) were greater than gammas for immediate JOLs ($G = .27$, $SE = .10$), $t(63) = 2.02$, $p < .05$, $d = 50$, replicating prior work (McCabe & Soderstrom, 2011). However, for delayed items, no difference was found between JOLs ($G = .86$, $SE = .04$) and JORKs ($G = .80$, $SE = .03$) ($p > .05$). Thus, although delaying judgments increased gamma correlations for both JOLs and JORKs for cue-only items, the delayed-judgment effect was attenuated for JORKs.

A separate 2 (Judgment Type: JOLs, JORKs) x 2 (Judgment Timing: immediate, delayed) mixed-model ANOVA was conducted on gammas for cue-target items (see Figure 3b). This analysis revealed no main effects of either Judgment Type or Judgment Timing (F's < 1), nor was the interaction reliable, $F(1, 66) = 1.90$, $MSE = .20$, $\eta_p^2 = .03$, $p = .17$. Given that no effects were found for cue-target items, no further analyses were performed. Thus, gamma correlations for JOLs and JORKs did not differ for cue-target items, and delaying judgments does not improve their resolution, a finding that is in contrast to cue-only items (cf. Dunlosky & Nelson, 1997).

Effects of Judgment Timing and Cue Type on Judgment Magnitude and Recall Performance

To assess differences between mean judgment magnitude of JOLs and JORKs, a 2 (Judgment Type: JOL, JORK) x 2 (Judgment Timing: immediate, delayed) x 2 (Cue Type: cue-only, cue-target) mixed-model ANOVA was conducted. These data are presented in the upper half of Table 3. A main effect of Judgment Timing was found, $F(1, 140) = 44.08$, $MSE = 2.58$,

η_p^2 = .24, but was qualified by a Judgment Timing by Cue Type interaction, $F(1, 140) = 39.73$, $MSE = 2.33$, η_p^2 = .22. Follow-up *t*-tests confirmed that delaying judgments lowered judgment

Table 3
Mean Judgment Magnitude and Proportion Recalled as Function of Judgment Timing and Cue Type in Experiment 2

		Cue-Only	Cue-Target
Judgment Magnitude			
JOLs			
	Immediate	1.90 (.42)	2.10 (.23)
	Delayed	2.28 (.37)	2.05 (.34)
JORKs			
	Immediate	1.99 (.28)	2.02 (.31)
	Delayed	2.36 (.34)	2.08 (.36)
Proportion Recalled			
JOLs			
	Immediate	.22 (.18)	.17 (.15)
	Delayed	.21 (.14)	.41 (.22)
JORKs			
	Immediate	.21 (.13)	.24 (.12)
	Delayed	.21 (.16)	.42 (.20)

Note: The judgment scales for JOLs and JORKs were *1*=high confidence/remember, *2*=medium confidence/know, *3*=low confidence/forget. Standard deviations are reported in parentheses.

magnitude for cue-only items, $t(71) = 7.96$, $p < .001$, $d = 1.04$, but not for cue-target items ($p > .05$). (Note that the judgment scales were such that a "1" was the highest magnitude judgment.) There was no main effect of Judgment Type ($F < 1$) or Cue Type, $F(1, 140) = 2.23$, $MSE = .37$, η_p^2 = .02, p = .14, nor was the Judgment Type by Cue Type interaction reliable, $F(1, 140) = 1.17$, $MSE = .19$, η_p^2 = .01, p = .28. The Judgment Type by Timing interaction was not reliable ($F < 1$), nor was the three-way interaction between Judgment Type, Judgment Timing, and Cue Type, $F(1, 140) = 1.17$, $MSE = .07$, η_p^2 = .01, p = .28. Thus, the manipulations of timing and cue type affected the magnitude of JOLs and JORKs in similar ways. Specifically, for both JOLs and JORKs, delaying judgments lowered judgment magnitude for cue-only items, but not for cue-target items. Presumably, this is because delayed cue-only items encourage target retrieval

before participants make their judgments. If participants fail to retrieve the target, their judgments are lowered compared to when the target is retrieved. This added influence of forgetting is not likely to influence cue-target items because the target is always presented at the time the judgment is solicited.

A 2 (Judgment Type: JOL, JORK) x 2 (Judgment Timing: immediate, delayed) x 2 (Cue Type: cue-only, cue-target) mixed-model ANOVA was also conducted on recall performance. These data are presented in the lower half of Table 3. There were main effects of Judgment Timing, $F(1, 140) = 82.58$, $MSE = .76$, $\eta_p^2 = .37$, and Cue Type, $F(1, 140) = 13.70$, $MSE = .64$, $\eta_p^2 = .09$, but these were qualified by a reliable Judgment Timing by Cue Type interaction, $F(1, 140) = 89.90$, $MSE = .82$, $\eta_p^2 = .39$. Follow-up *t*-tests showed that delaying judgments increased recall performance for cue-target items, $t(71) = 11.61$, $p < .001$, $d = 1.20$, but not for cue-only items ($p > .05$). The main effect of Judgment Type and the Judgment Type by Cue Type interaction was not reliable (F's < 1), nor was the Judgment Type by Judgment Timing interaction, $F(1, 140) = 1.41$, $MSE = .01$, $\eta_p^2 = .01$, $p = .24$. Finally, the three-way interaction between Judgment Type, Judgment Timing, and Cue Type interaction was not reliable, $F(1, 140) = 2.75$, $MSE = .03$, $\eta_p^2 = .02$, $p = .10$. Taken together, these results indicate that the manipulations of timing a cue type affected recall performance in similar ways for JOLs and JORKs. Specifically, delayed cue-target items were recalled more than any other item type.

Distribution of JOLs and JORKs as a Function of Judgment Timing and Cue Type

The impact of judgment timing and cue type on JOLs and JORKs was further investigated by comparing the distribution of predictions (1/Remember, 2/Know, 3/Forget) as a function of these variables (cf. Dunlosky & Nelson, 1994). These data are reported in Table 4. First, a 2 (Judgment Type: JOL, JORK) x 2 (Judgment Timing: immediate, delayed) x 2 (Cue

Table 4
Distribution of Responses for JOLs and JORKs as a Function of Judgment Timing and Cue Type for Experiment 2

	Cue-Only	Cue-Target
JOLs		
Immediate		
1 (high confidence)	.35 (.26)	.23 (.14)
2 (medium confidence)	.39 (.16)	.45 (.20)
3 (low confidence)	.26 (.19)	.32 (.16)
Delayed		
1 (high confidence)	.27 (.17)	.28 (.19)
2 (medium confidence)	.17 (.13)	.40 (.20)
3 (low confidence)	.55 (.22)	.33 (.20)
JORKs		
Immediate		
Remember	.29 (.16)	.31 (.17)
Know	.43 (.20)	.35 (.15)
Forget	.28 (.18)	.33 (.18)
Delayed		
Remember	.21 (.17)	.29 (.18)
Know	.22 (.13)	.33 (.16)
Forget	.57 (.20)	.38 (.22)

Note: Standard deviations are reported in parentheses.

Type: cue-only, cue-target) x 3 (Study Response: 1/Remember, 2/Know, 3/Forget) mixed-model ANOVAs was conducted. Note that because the proportion of study responses sum to 1 across both levels of Judgment Timing and Cue Type, the main effects of Judgment Type, Judgment Timing, and Cue Type could not be calculated. Thus, of greatest interest was the potential main effect of Study Response and the various interactions. The main effect of Study Response was reliable, $F(1, 140) = 15.88$, $MSE = 1.33$, $\eta_p^2 = .10$; however, this was qualified by a Judgment Timing by Study Response interaction, $F(1, 140) = 43.63$, $MSE = 1.28$, $\eta_p^2 = .24$, and a three-way interaction between Judgment Timing, Cue Type, and Study Response, $F(1, 140) = 39.80$, $MSE = 1.17$, $\eta_p^2 = .22$. No other interactions were reliable, F's ≤ 2.16, p's $\geq .14$

To unpack the three-way interaction between Study Response, Judgment Timing, and Cue Type, separate 2 (Judgment Timing: immediate, delayed) x 3 Study Response (1/Remember,

2/Know, 3/Forget) mixed-model ANOVAs were conducted for cue-only and cue-target items. Turning first to the cue-only items, no main effect of Judgment Timing was found ($F < 1$), but the main effect of Study Response was reliable, $F(1, 71) = 14.03$, $MSE = 1.25$, $\eta_p^2 = .17$. However, this was qualified by a Timing by Study Response interaction, $F(1, 71) = 63.11$, $MSE = 2.45$, $\eta_p^2 = .47$. Follow-up one-way ANOVAs were conducted to unpack this interaction. For immediate judgments, no effect of Study Response was found, $F(1, 71) = 1.57$, $MSE = .10$, $\eta_p^2 = .02$, $p = 22$; however, for delayed judgments, the effect of Study Response was reliable, $F(1, 71) = 56.55$, $MSE = 3.60$, $\eta_p^2 = .44$. *T*-tests revealed that 3/Forget judgments were given more than both 1/remember judgments, $t(71) = 7.52$, $p < .001$, $d = 1.68$, and 2/know judgments, $t(71) = 10.05$, $p < .001$, $d = 2.06$. No difference was found between judgments of 1/remember and 2/know judgments ($p > .05$). Turning to cue-target items, no main effects of Judgment Timing, $F(1, 71) = 3.70$, $MSE = 2.32E5$, $\eta_p^2 = .05$, $p = .07$, or Study Response, $F(1, 71) = 3.39$, $MSE = .27$, $\eta_p^2 = .05$, $p = 07$, were found. The interaction between Judgment Timing and Study Response was also unreliable ($F < 1$).

Discussion

In Experiment 2, both the timing of JOLs and JORKs and the cue used when judgments were solicited was manipulated. As anticipated, the predictive accuracy of JOLs and JORKs both benefited from a delay for cue-only items (see Figure 3a). However, given that immediate JORKs were more predictive of later recall than immediate JOLs, the delayed judgment effect was substantially attenuated for JORKs, providing evidence for the claim that immediate JORKs are based more on retrieval processes than immediate JOLs. Inconsistent with my predictions, however, was that the cue-type manipulation affected JOLs and JORKs in similar ways. The predictive accuracy of JOLs and JORKs as a function of timing was very similar for cue-target

items (see Figure 3b). It was predicted that this manipulation should matter less for JORKs than JOLs because JORKs are presumed to rely on retrieval processes, which the cue-type manipulation seeks to tease apart. That is, a manipulation intended to encourage the use of retrieval processes should not impact (or should have less impact on) judgments putatively based on retrieval processes. This prediction was not supported by the data, but as I argue in the General Discussion, this might be due to the large differences in recall produced by delaying judgments for cue-target items, a difference not shown for cue-only items (see Table 3). Finally, in terms of scale usage, the manipulations of timing and cue type affected JOLs and JORKs in similar ways (see Table 4).

CHAPTER IV: Experiment 3

Keeping with the common theme of JORKs and their reliance on retrieval processes, the final experiment sought to determine whether JORKs are immune to a manipulation of encoding fluency. Rhodes and Castel (2008; see also Kornell, Rhodes, Castel, & Tauber, 2011) showed that JOLs are sensitive to the font size in which the to-be-remembered items are presented, such that relatively higher JOLs were given to items in a larger font size. This was true despite recall not differing as a function of font size; thus, a metacogntive illusion was demonstrated as a result of the encoding fluency produced by the font size manipulation. If JORKs are based more on retrieval processes than JOLs, then JORKs should be less prone to this illusion.

Method

Participants

Eighty-eight undergraduates from Colorado State University participated in this experiment for course credit. The mean age was 19.13 years and 78% of participants were Female. There were 44 participants in each of the two between-subjects conditions (JOLs, JORKs). Participants were tested in small groups up to three, each on their own computer.

Materials

Following Rhodes and Castel (2008), studied items consisted of 40 single, concrete nouns that were taken from the Kucera and Francis (1976) norms. These were equated in terms of their frequency, number of letters, and number of syllables. Half of the items were presented in 18-pt Arial font, whereas the other half were presented in 48-pt Arial font. Each item was presented equally as often in both font sizes. Of the 40 total items, 4 served as buffer items (2 at the beginning of the list, 2 at the end of the list), which were presented in either small or large font (one of each on each end of the list). All items were presented randomly.

Design and Procedure

A 2 (Judgment Type: JOL, JORK) x 2 (Font Size: small, large) experimental design was used. Judgment Type was manipulated between-subjects, whereas Font Size was a within-subjects manipulation.

Participants studied words one at a time for 5 s each (with a 500ms interstimulus interval) and were instructed that they would predict their memory performance for each item. Immediately following the study of each item, JOLs or JORKs were solicited following the methods described in Experiment 1. After all words had been given a judgment, participants engaged in filler task for 5 min, which consisted of a psychology-term word search puzzle. After the filler task, participants were asked to recall, in any order, the words previously studied. This was done on a blank sheet of paper provided by the experimenter. Participants were given 4 min for recall, after which time debriefing forms were distributed and the participants were excused from the experiment.

Results

Resolution

Of primary interest in Experiment 3 was to determine whether JORKs are immune (or at least less susceptible) to the encoding fluency manipulation of font size. To do so, gamma correlations were calculated between judgment type and font size (see Figure 4; font size was coded such that *1 = small* and *2 = large*). Consistent with my predictions, the Judgment-Font Size gamma for JOLs (G = .62; SE = .04) was reliably greater than for JORKs (G = .47; SE = .05), $t(86) = 2.19$, $p < .05$, $d = .48$, suggesting that font size informed JOLs to a greater extent than it did JORKs. For completeness, additional gamma correlations between font size and recall, and between judgments and recall were calculated (see Figure 4). Turning first to Font

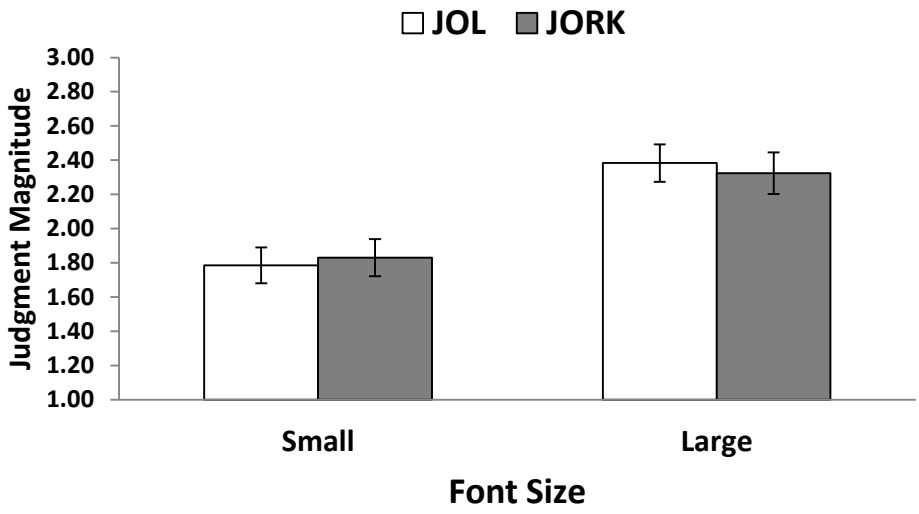
Size-Recall gammas, neither JOLs (G = -.03; SE = .07) nor JORKs (G = -.13; SE = .07) differed from zero (p's > .05). In addition, these gammas did not differ between each other (p > .05). Thus, for both JOLs and JORKs, recall performance did not differ as a function of font size. Finally turning to Judgment-Recall gammas, JOLs (G = .16; SE = .06) did not differ from JORKs (G = .22; SE = .07) (p > .05), indicating that both judgments types similarly predicted recall performance.

Figure 4. Judgment-Font Size, Font Size-Recall, and Judgment-Recall gamma correlations for Experiment 3. Error bars represent standard error.

Effects of Font Size on Overall Judgment Magnitude and Recall Performance

Figure 5a shows the impact of font size on overall judgment magnitude for both JOLs and JORKs. (Note that for purposes of the Figure, JOLs and JORKs were reversed scaled such that a higher value reflects a higher judgment, e.g., a '3' reflects a high confident JOL and a 'remember' JORK; the opposite was true when participants actually made their judgments during the experiment, e.g., a '1' reflected a high confident JOL and a 'remember' JORK). To determine the impact of font size on these judgments, a 2 (Judgment Type: JOLs, JORKs) x 2 (Font Size: small, large) mixed-model ANOVA was conducted. This revealed a main effect of

(a) Judgment Magnitude



(b) Recall Performance

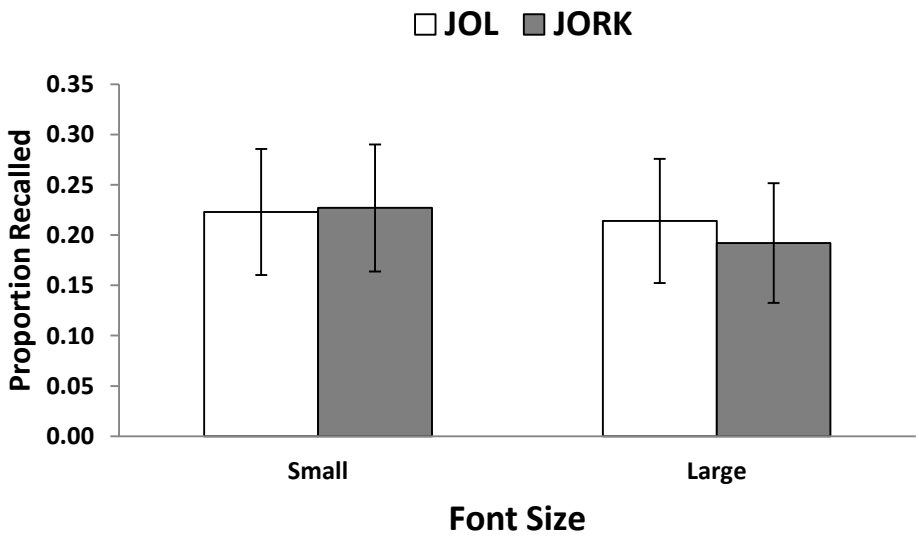


Figure 5. Panel (a) represents JOL and JORK magnitude as a function of font size and panel (b) represents proportion recall for JOLs and JORKs as a function of font size for Experiment 3. Note that for purposes of the Figure, JOLs and JORKs were reversed scaled such that a higher value reflects a higher judgment; the opposite was true when participants actually made their judgments during the experiment. Error bars represent standard error.

Font Size, $F(1, 86) = 160.79$, $MSE = 13.17$, $\eta_p^2 = .65$, indicating that both JOLs and JORKs increased in magnitude as a function of font size. For JOLs, mean judgment magnitude increased from 1.79 ($SD = .35$) to 2.38 ($SD = .37$) for small and large items, respectively, $t(43) =$

9.51, $p < .001$, $d = 1.64$. Similarly, mean JORK magnitude increased from 1.83 ($SD = .28$) to 2.33 ($SD = .29$), $t(43) = 8.40$, $p < .001$, $d = 1.79$. Neither the main effect of Judgment Type, $F < 1$, nor the Judgment Type by Font Size interaction, $F(1, 86) = 1.34$, $MSE = .11$, $\eta_p^2 = .02$, $p = .25$, was significant.

Figure 5b shows recall performance (indexed by proportion correct) for both JOLs and JORKs as a function of font size. To determine the impact of font size on recall performance, a 2 (Judgment Type: JOLs, JORKs) x 2 (Font Size: small, large) mixed-model ANOVA was conducted. This revealed no main effect of Font Size, $F(1, 86) = 2.80$, $MSE = .02$, $\eta_p^2 = .03$, $p = .10$, no main effect of Judgment Type, $F < 1$, and no reliable Judgment Type by Font Size interaction, $F(1, 86) = 1.15$, $MSE = .01$, $\eta_p^2 = .01$, $p = .29$. Thus, this analysis indicates that for both JOLs and JORKs, font size had no influence on recall performance, a finding that converges with the Font Size-Recall gamma correlations previously reported.

Distribution of JOLs and JORKs as a Function of Font Size

The impact of font size on memory predictions was further examined by comparing the distribution of responses for JOLs and JORKs as a function of font size (see Table 5). To do so, a 2 (Judgment Type: JOLs, JORKs) x 2 (Font Size: small, large) x 3 (Study Response: 1/Remember, 2/Know, 3/Forget) mixed-model ANOVA was conducted. Note that because the proportion of study responses sum to 1 for both JOLs and JORK, and across each font size, the main effects of Judgment Type and Font Size could not be calculated. Thus, the focus is on the main effect of Study Response and the various interactions. A main effect of Study Response was found, $F(1, 86) = 8.90$, $MSE = .57$, $\eta_p^2 = .09$; however, this was qualified by a Study Response by Font Size interaction, $F(1, 86) = 160.71$, $MSE = 6.53$, $\eta_p^2 = .65$, indicating that the distribution of study responses differed as a function of font size. Inspection of Table 5 shows

that for both JOLs and JORKs, there was general trend for small words to receive relatively more low judgments (i.e., JOLs of '3' and 'Forget' JORKs), whereas large words received relatively more high judgments (i.e., JOLs of '1' and 'Remember' JORKs). No other two-way interactions were reliable (F's < 1), nor was the three-way interaction between Judgment Type, Font Size, and Study Response, $F(1, 86) = 1.36$, $MSE = .06$, $\eta_p^2 = .02$, $p = .25$.

Although the previous analysis yielded no interactions with Judgment Type (JOLs vs. JORKs), a closer inspection of Table 5 shows that JOLs and JORKs may have, in fact, differed in terms of how font size affected study responses. Specifically, the difference between the

Table 5
Distribution of Responses for JOLs and JORKs as a Function of Font Size for Experiment 3

	Small	Large
JOLs		
1 (high confidence)	.18 (.14)	.50 (.28)
2 (medium confidence)	.43 (.17)	.38 (.22)
3 (low confidence)	.39 (.23)	.12 (.13)
JORKs		
Remember	.25 (.16)	.47 (.19)
Know	.33 (.18)	.38 (.15)
Forget	.42 (.17)	.15 (.13)

Note: Standard deviations are reported in parentheses.

proportions of each study response as a function of font size indicates that, for JOLs, large words received 32% more (.50 - .18) high confidence judgments than small words. For JORKs, however, this difference is markedly smaller, as large words received only 22% more (.47 - .25) 'Remember' responses than small words. Follow-up t-tests confirmed that the source of this difference comes from the fact that small words were given more 'Remember' JORKs (25%) than high confidence JOLs (18%), $t(86) = 2.30$, $p < .05$, $d = .50$, whereas no difference was found when making this comparison for large words ($p > .05$). Regarding the other study responses, the similarity between JOLs and JORKs was remarkable. For both JOLs and JORKs,

the difference between medium judgments (i.e., JOLs of '2' and 'know' JORKs) as a function font size was 5%, and the difference between low judgments (i.e., JOLs of '3' and 'Forget' JORKs) as a function of font size was 27% (p's > .05).

Discussion

In Experiment 3, font size was manipulated in order to determine whether JORKs are less susceptible to this encoding fluency manipulation compared to JOLs. Font size has been shown to have a robust effect on JOLs, but no effect on recall performance (Rhodes & Castel, 2008). Thus, to the degree that JORKs are based less on encoding fluency and more on retrieval processes than JOLs, JORKs should be less prone to the metacognitive illusion elicited by the font size manipulation. Indeed, while the current data replicate Rhodes and Castel's general pattern that memory predictions are impacted by font size whereas memory performance is not (see Figure 5), gamma correlations between judgment and font size indicated that JOLs were impacted more by font size than JORKs (see Figure 4). Furthermore, analyses of study response distributions revealed the source of this difference: 'Remember' JORKs were less susceptible to the font size manipulation than were high confidence JOLs (see Table 5). It should be emphasized that JORKs were not completely immune to the font size manipulation; rather, they were less susceptible to it than JOLs. One unexpected result from Experiment 3 was that the gamma correlation between judgment and recall did not differ between JOLs and JORKs—that is, JOLs predicted recall performance just as well as JORKs (see Figure 4). At first glance, this seems surprising given that JORKs have generally demonstrated better predictive accuracy than JOLs (McCabe & Soderstrom, 2011); however, as will be argued in the General Discussion, the JORK advantage hinges on the type of memory test employed, and tests of free recall may not be suited to generate this effect.

CHAPTER V: General Discussion

McCabe and Soderstrom (2011) showed that immediate judgments of remembering and knowing (JORKs) predict memory better than more traditional judgments of learning (JOLs). However, it is unclear why JORKs enhance predictive accuracy. One idea is that immediate JORKs are based, to a greater extent than immediate JOLs, on retrieval processes, a notion that was directly tested in the current experiments. If JORKs are largely based on retrieval processes, then JORKs should resemble RCJs (retrospective confidence judgments) in a number of important ways (Experiment 1); JORKs should resemble delayed JOLs in regards to their relative accuracy (Experiment 2); and JORKs should be relatively immune to manipulations of encoding fluency (Experiment 3). This section will first summarize the results of each experiment and how these results bear on the issue at hand. The broader implications—both theoretical and practical—of the current data will also be discussed, as well as potentially fruitful areas of future investigation.

Summary of Current Experimental Findings

Experiment 1 tested the hypothesis that JORKs, if more reliant on retrieval processes than JOLs, should resemble RCJs on a number of dependent measures. This hypothesis stems from previous research showing RCJs to be more reliant on retrieval processes than JOLs, resulting in RCJs showing greater predictive accuracy than JOLs (Busey et al., 2000; Dougherty et al., 2005). Following Dougherty et al., Experiment 1 utilized the pre-judgment recall and monitoring (PRAM) methodology, in which participants attempted to recall items immediately before making metamemory judgments for those items, thus allowing a direct measure of retrieval when the metamemory judgment is made (see Nelson et al., 2004). As a between-subjects manipulation, participants made JOLs, JORKs, or RCJs after the initial recall attempt.

These judgments were then related to a number of measures to assess the degree to which JORKs are based on retrieval processes.

Consistent with the idea that immediate JORKs are more reliant on retrieval processes than are immediate JOLs, JORKs resulted in higher gamma correlations with pre-judgment recall than JOLs. That is, JORKs were based more on whether an item was initially recalled or not, compared to JOLs. These gamma correlations were even higher for RCJs, replicating the finding that RCJs are based almost exclusively on retrieval processes (Dougherty et al., 2005), a notion consistent with the instruction for RCJs to assess past test performance (i.e., they are *retrospective*). JOLs and JORKs, on the other hand, are *prospective*, asking people to anticipate future memory experiences. Consequently, participants making JOLs and JORKs were likely influenced by additional information—perhaps the fluency in which the items were processed. This possibility is particularly likely for JOLs, which have been shown to rely on such processing fluency (e.g., Koriat & Bjork, 2005; Rhodes & Castel, 2008). Indeed, the finding that JORKs fell in between JOLs and RCJs in terms of their correlation with pre-judgment recall suggests that JORKs are based more on retrieval processes than JOLs but less than RCJs.

The correlations between metacognitive judgments and pre-judgment recall in Experiment 1 align with the idea that JORKs are reliant on retrieval processes, but other gamma correlations do not. First, the correlation between judgments and pre-judgment recall latency—an indirect measure of retrieval—indicated that JORKs resembled JOLs more than they resembled RCJs. That is, the time it took participants to complete pre-judgment recall attempts was related to JORKs and JOLs in similar ways. Although this was not expected, these correlations should be interpreted with caution because latency measures are proxies for more direct indices. The other pattern of gamma correlations that was unexpected was between

judgments and final recall. Dougherty et al. (2005) showed that, as a result of RCJs being based more on pre-judgment recall than JOLs, RCJs also showed relatively greater accuracy in predicting final recall, a pattern that was replicated in the current experiment. However, it is surprising that JORKs did not show higher correlations between judgments and final recall than JOLs. One possible explanation has to do with the fact that all judgments in Experiment 1 were technically delayed (i.e., at least a few seconds elapsed between initially studying items and making their corresponding metamemory judgments). Previous research indicates that delaying judgments substantially improves their relative accuracy (for a review, see Rhodes & Tauber, 2011a), and that only brief delays (on the order of seconds) are needed to produce this benefit (Kelemen & Weaver, 1997). Thus, it is possible that even though JORKs seem to be more reliant on retrieval processes than JOLs, this difference might not be enough to produce benefits in relative accuracy over and beyond that produced by delaying judgments.

Finally, Experiment 1 revealed that the three metamemory judgments—JOLs, JORKs, and RCJs—did not differ in terms of overall pre-judgment recall or final recall; all conditions produced similar levels of correctly recalled items before and after metamemory judgments were solicited. These findings are inconsistent with results obtained by Dougherty et al. (2005) who showed that JOLs were associated with higher final recall than RCJs, possibly because JOLs encourage participants to encode items differently than RCJs. Although this idea is not supported statistically by the current data, it should be noted that, numerically speaking, JOLs did show higher final recall than RCJs (56% vs. 50%) with JORKs falling in between (53%). Thus, JOLs (and perhaps other prospective metamemory judgments such as JORKs) might encourage people to think about to-be-remembered information differently than other judgments, but more research is clearly needed to examine this possibility.

Whereas Experiment 1 utilized the PRAM methodology to investigate the basis of JORKs, Experiment 2 employed a standard delayed-JOL methodology. This procedure has consistently shown that delaying JOLs improves their predictive accuracy—a finding termed the *delayed-JOL effect*—primarily because delaying JOLs encourages people to base their predictions on information from long-term memory (Dunlsoky & Nelson, 1992; Nelson & Dunlosky, 1991; see Rhodes & Tauber, 2011a). Following this procedure, in Experiment 2 studied items were given either immediate or delayed judgments (JOLs or JORKs) in the presence of either the cue only (e.g., *DOG - ?*) or the cue and target (e.g., *DOG – SPOON*). After this study phase, participants engaged in a final cued recall test, thus allowing for a determination as to the effects of delay and cue-type on the predictive accuracy of JOLs and JORKs.

There were two primary hypotheses in Experiment 2. First, because immediate JORKs have been shown to predict memory better than immediate JOLs (McCabe & Soderstrom, 2011), I predicted that the effect of delay would be substantially attenuated for JORKs relative to JOLs. That is, if immediate JORKs already encourage people to base their predictions on retrieval processes, then delaying these judgments should matter relatively less than delaying JOLs, which are thought to be particularly impacted by encoding fluency when made immediately after study. The second hypothesis was that the type of cue (i.e., cue-only vs. cue-target) presented at the time of the judgment should matter less for JORKs than JOLs. Similar to delaying JOLs, JOLs made in the presence of the cue only encourages people to rely to a greater extent on retrieval processes than when the cue and target are presented together at the time the judgment is made (Dunlsoky & Nelson, 1992; see Rhodes & Tauber, 2011a). Thus, it was predicted that for

JORKs—as a result of these judgments already being based more on retrieval processes than immediate JOLs—this cue-type manipulation should matter less than for JOLs.

Did the results of Experiment 2 support these predictions? The answer is both 'yes' and 'no.' For cue-only items, the effect of delay on predictive accuracy was, in fact, attenuated for JORKs relative to JOLs. That is, although there was an effect of delay for both JOLs and JORKs, the difference in predictive accuracy between immediate and delayed judgments was smaller for JORKs than JOLs. This effect was driven by immediate JORKs showing higher predictive accuracy than immediate JOLs; there was no difference between JOLs and JORKs for delayed items. Thus, for cue-only items, my expectations were supported. However, the cue-type manipulation affected JOLs and JORKs in similar ways, which runs counter to what I predicted. Specifically, both JOLs and JORKs showed a delayed effect for cue-only items (albeit JORKs showed a smaller effect), but for the cue-target items, the predictive accuracy of JOLs and JORKs looked very similar as a function of timing. This might reflect the general idea that cue-target items impede participants' ability to engage in retrieval processes prior to making their metamemory judgments (Dunlosky & Nelson, 1992). However, these comparisons are complicated somewhat by the fact that large recall differences were produced by delaying judgments for cue-target items—presumably due to spaced study (for a review on spacing effects, see Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006)—an effect not observed for cue-only items. For JOLs, 41% of the delayed items were recalled compared to 17% of immediate items; similarly, for JORKs, 42% of delayed items were recalled compared to 24% of immediate items. Thus, the lack of differences in predictive accuracy for cue-target items as a function of timing between JOLs and JORKs are difficult to interpret because such large recall differences were

observed. Ideally, one needs to equate memory performance to make strong conclusions regarding memory monitoring (see Hertzog, Dunlosky, & Sinclair, 2010).[2]

Finally, Experiment 3 used yet another methodology to investigate the idea that JORKs are base more on retrieval processes than are JOLs. Following Rhodes and Castel (2008), participants studied words in either large or small font. As a between-subjects manipulation, either JOLs or JORKs were made immediately after each item. Rhodes and Castel showed that JOLs were highly influenced by font size—specifically, words presented in large font were given higher JOLs than those presented in small font—despite font size having no impact on later recall. I predicted that if JORKs are more reliant on retrieval processes than JOLs, then JORKs should be immune (or at least less susceptible) to this metacognitive illusion that capitalizes on encoding fluency.

Experiment 3 replicated Rhodes and Castel's (2008) basic pattern of font size influencing metamemory judgments but not recall. To determine if JORKs were less susceptible than JOLs to the font size manipulation, gamma correlations between judgments and font size were computed, which measured the extent to which participants considered font size when making their judgments. Consistent with my prediction, this correlation was lower for JORKs than JOLs, suggesting that JORKs were based less on font size (i.e., encoding fluency) than JOLs. Furthermore, analyses of response distributions revealed the source of this difference: 'Remember' JORKs (i.e., the highest magnitude JORKs) were influenced less by font size than were the highest confidence JOLs. One peculiar finding from Experiment 3 was that JORKs did

---

[2]An experiment is currently underway attempting to equate memory performance between immediate and delayed cue-target items to better assess how this manipulation affects monitoring processes in JORKs compared to JOLs.

not predict overall recall better than JOLs as evidenced by equivalent gamma correlations between judgments and final recall. This is seemingly inconsistent with McCabe and Soderstrom (2011) who have shown that JORKs generally demonstrate better predictive accuracy than JOLs. However, the JORK advantage hinges on the type of test used to assess memory. Indeed, McCabe and Soderstrom state,

> …if making JORKs during study encourages participants to distinguish between items that will later include contextual details and those that will not, then accuracy (i.e., gamma correlations) would depend on having an outcome measure in which items that include contextual details are distinguished from those that do not. (p. 613)

Thus, it might be the case that tests of free recall are not suitable to generate the JORK advantage because these tests do not allow one to make judgments on the basis of the amount or type of information retrieved.[3]

On the whole, the current experiments provide mixed support for the idea that immediate JORKs rely more heavily on retrieval processes than do immediate JOLs. The major data in favor of this idea include the following: (1) JORKs were more reliant on pre-judgment recall than JOLs (Experiment 1), (2) delaying judgments for cue-only items had less of an effect on the predictive accuracy of JORKs than JOLs (Experiment 2), and (3) JORKs were less susceptible than JOLs to the encoding fluency manipulation of font size (Experiment 3). The major data in opposition to this idea include the following: (1) JORKs showed more similarity to JOLs than RCJs in terms of their ability to predict future recall (Experiment 1), (2) the cue type

---

[3]In their Experiment 3, McCabe and Soderstrom (2011) showed that the JORK advantage does not generalize to a yes-no recognition test, presumably for the same reason. That is, this type of test does not allow for one to distinguish between items that are accompanied by contextual details and those that are not.

manipulation affected JORKs and JOLs in similar ways (Experiment 2), and (3) JORKs were influenced by an encoding fluency manipulation (albeit less so than JOLs) and failed to better predict free recall performance than JOLs (Experiment 3). Thus, although it may be the case that JORKs are, in fact, based more on retrieval processes than JOLs, the current experiments did not provide definitive support for this idea.

Given that many of my predictions were not supported, alternative theoretical accounts of the JORK advantage reported by McCabe and Soderstrom (2011) must be considered. For example, rather than JORKs being based more on retrieval processes than JOLs, perhaps JORKs show better predictive accuracy than JOLs because JORKs provide a more defined scale for participants to use when making their memory predictions. Indeed, instructions for making JORKs are lengthy and detailed (see Appendix), whereas for JOLs participants are relatively free to determine the types of information or cues that are associated with a given point on the scale (for a similar argument regarding JOLs scales, see Benjamin & Diaz, 2008). Consequently, there may be substantial variation across participants making JOLs in how the scale is used, which may ultimately result in relatively lower predictive accuracy of these judgments. One prediction of this explanation, it seems, is that JORKs should always show better predictive accuracy than JOLs; however, this is clearly not this case. Experiment 3 of the current study showed that JORKs were equivalent to JOLs in predicting free recall, and McCabe and Soderstrom showed that the JORK advantage did not emerge when yes-no recognition was used as the memory test. Thus, the JORK advantage interacts with test type, a finding that is not easily accommodated by the explanation that JORKs simply supply participants with a more defined scale than JOLs. Nevertheless, alternative explanations of the JORK advantage—such as this one—should be considered.

It is also important to note that even if JORKs are more reliant on retrieval processes than JOLs, the current data clearly show that immediate JORKs are not exclusively based on retrieval processes, which makes sense given that these prospective judgments are made *immediately* after studying a bit of information. Consequently, encoding operations (e.g., fluency) also seems to affect JORKs, but less so than JOLs. For example, Experiment 2 demonstrated that, although JORKs showed a smaller effect of delay for cue-only items on their predictive accuracy, JORKs still showed a delayed effect, indicating that non-diagnostic information from the encoding experience was influencing immediate JORKs. Likewise, JORKs were not entirely immune to the font size manipulation in Experiment 3; rather they were less susceptible to it than JOLs. Thus, the current data suggest that JORKs are still influenced by encoding operations, but that these operations might inform JORKs less than JOLs.

Future Directions

Further research is clearly needed to bolster the idea that JORKs are based more on retrieval processes than JOLs. For example, if this is true, then JORKs' predictive accuracy should suffer if inaccurate information is retrieved at the time the judgment is made. This possibility could be investigated by modifying the PRAM methodology used in Experiment 1to include deceptive items (see Kelley & Sahakyan, 2003; Rhodes & Tauber, 2011b). For example, a participant might study *NURSE – DOLLAR*, which, when later given the cue word and three letters of the target (*NURSE – DO_ _ _ R*), a highly semantically-related target competitor might be recalled—in this case, *DOCTOR*—before the JORK is made. If JORKs are reliant on retrieval processes, then their predictive accuracy for deceptive items should suffer compared to control items.

In a related vein, arguing that JORKs are based relatively more on retrieval processes than JOLs and that JOLs are based relatively more on encoding processes than JORKs says very little in regards to what, precisely, these judgments are based on. Indeed, actual remembering and knowing were not measured in the current experiments, and thus one can only infer such information from how participants were instructed to make these judgments and the emergent data. Perhaps one way to get a better understanding of the basis of JOLs and JORKs would be to use a think aloud protocol in which participants are asked to continuously verbalize their thoughts during the study and test episodes of each item (see Fox, Ericcson, & Best, 2011; McCabe, Geraci, Bowman, Sensenig, & Rhodes, 2011). If JORKs are based more on retrieval processes than JOLs, then JORK participants might verbalize more details during study that are reinstated at test, whereas JOL participants might focus relatively more on fleeting information related to the fluency of each item (i.e., "that was an easy word" or "that word was in large font"). Such a methodology might reveal important differences (and similarities) regarding the information that is used by participants when making JOLs and JORKs.

In addition to establishing the basis of JORKs, future research might also investigate the generalizability of the JORK advantage reported by McCabe and Soderstrom (2011). Does the advantage, for example, generalize to situations in which participants read and learn from text—that is, does it improve *metacomprehension* (see Maki & McGuire, 2002; Thiede & Anderson, 2003)? Such research could provide real-world validity to the JORK advantage as metcomprehension tasks more closely resemble how students study for exams. Furthermore, it might be valuable to investigate JORKs in the context of aging. Regarding JOLs, younger and older adults often show equivalent predictive accuracy (measured via resolution; Hertzog & Dunlosky, 2011; for a review, see Hertzog & Dunlosky, 2004). Is the same true for the accuracy

of JORKs? Preliminary evidence suggests that, although relative accuracy using JORKs might also show age-equivalency, older adults over-predict recollective experiences relative to younger adults (Soderstrom, McCabe, & Rhodes, submitted). This is noteworthy given that aging is specifically related to deficits in recollection (McCabe et al., 2009), a finding that may have implications for predictions of future memory performance (see Toth, Daniels, & Solinger, 2011). Nevertheless, more research is needed to provide a comprehensive picture of how aging affects memory monitoring; the use of JORKs could be of benefit to this end.

Perhaps the most promising line of future research for JORKs relates to the consequences of their superior predictive accuracy. As stated in the Introduction, research in metacognition seeks to understand how people monitor their own cognitive processes *and* how such monitoring affects future behavior (i.e., control; for a review, see Koriat, 2007). Nelson and Narens (1990) proposed that monitoring directly influences behavior, and subsequent research using JOLs has bolstered such a relationship (Metcalfe & Finn, 2008; Rhodes & Castel, 2009). For example, Rhodes and Castel (2009) had participants listen to words in either a loud or quiet volume, making JOLs after each word. In addition to giving relatively lower JOLs to quite words, participants also chose to restudy quite words more frequently than loud words. Thus, participants' subjective experiences (as measured by their JOLs) were directly related to their behavioral control processes (measured by restudy choices). Furthermore, Kornell and Metcalfe (2006) showed that later learning was enhanced when restudy choices were honored as compared to a condition in which restudied items were randomly chosen, suggesting that metacognition benefits learning. Following these results, if JORKs lead to better monitoring than JOLs, then they should also lead to better study decisions. Such a possibility could be investigated by employing the honor/dishonor paradigm used by Kornell and Metcalfe. If JORKs do lead to

better study decisions than JOLs, it would provide further evidence that instructors should encourage their students to monitor *how* they will remember material during learning, rather than *if* they will remember this material.

Concluding Remarks

The current experiments provide preliminary evidence that immediate JORKs may be more reliant on retrieval processes than are immediate JOLs, providing evidence for the potential mechanism underlying McCabe and Soderstrom's (2011) JORK advantage—that JORKs show superior predictive accuracy relative to JOLs. Theoretically, this suggests that participants can be encouraged to focus *immediate* metamemory judgments on characteristics that are diagnostic of future retrieval, rendering these judgments more accurate than other prospective judgments—most notably JOLs—that are often biased by fleeting or irrelevant information (e.g., encoding fluency; e.g., Koriat et al., 2004; Rhodes & Castel, 2008). Additionally, the metamemory literature is currently dominated by theories generated by findings using JOLs. Thus, the introduction of JORKs represents an interesting wrinkle that will need to be accommodated by extant theories of metamemory. Finally, finding ways to improve metamemory monitoring—as JORKs seem to do—and discovering their corresponding mechanisms of operation, has potential implications in various domains, such as educational settings and memory rehabilitation.

REFERENCES

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

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

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

Bowler, D. M., Gardiner, J. M., & Grice, S. (2000). Episodic memory and remembering in adults with Asperger's syndrome. *Journal of Autism and Developmental Disorders, 30*, 305-316.

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

Castel, A.D., McCabe, D. P., Roediger, H. L. III, & Heitman, J. L. (2007). The dark side of expertise: Domain specific memory errors. *Psychological Science*, *18*, 3-5.

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin*, *132*, 354-380.

Cohen, R. L., Sandler, S. P., & Keglevich, L. (1991). The failure of memory monitoring in a free recall task. *Canadian Journal of Psychology*, *45*, 523-538.

Connor, L., Dunlosky, J., & Hertzog, C. (1997). Age-related differences in absolute but not relative metamemory accuracy. *Psychology and Aging, 12*, 50-71.

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

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

Dunlosky, J., & Nelson, T. O. (1997). Similarity between the cue for judgments of learning (JOL) and the cue for test is not the primary determinant of JOL accuracy. *Journal of Memory and Language*, *36*, 34-49.

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.

Fox, M. C., Ericsson, K. A., and Best, R. (2011). Do procedures for verbal reporting of thinking have to be reactive? A meta-analysis and recommendations for best reporting methods. *Psychological Bulletin, 137,* 316-344.

Gardiner, J. M. (2002). Episodic memory and autonoetic consciousness: A first-person approach. *Episodic memory: New directions in research* (pp. 11-30). New York: Oxford University Press.

Gardiner, J. M., Java, R. I., & Richardson-Klavehn, A. (1996). How level of processing really influences awareness in recognition memory. *Canadian Journal of Experimental Psychology, 50,* 114-122.

Gregg, V. H., & Gardiner, J. M. (1994). Recognition memory and awareness: A large effect of study-test modalities on 'know' responses following a highly perceptual orienting task. *European Journal of Cognitive Psychology, 6,* 137-147.

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

Hart, J. T. (1967). Memory and the memory-monitoring process. *Journal of verbal learning and verbal behavior, 6,* 685-691.

Hertzog, C., & Dunlosky, J. (2004). Aging, metacognition, and cognitive control. In *The psychology of learning and motivation: Advances in research and theory, Vol 45* (pp. 215-251). San Diego, CA US: Elsevier Academic Press.

Hertzog, C., & Dunlosky, J. (2011). Metacognition in later adulthood: Spared monitoring can benefit older adults' self-regulation. *Current Directions in Psychological Science, 20,* 167-173.

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

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

Kassam, K. S., Gilbert, D. T., Swencionis, J. K., & Wilson, T. D. (2009). Misconceptions of memory: The Scooter Libby effect. *Psychological Science, 20,* 551-552.

Kelemen, W. L., & Weaver, C. A. (1997). Enhanced metamemory at delays: Why do judgments of learning improve over time? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23,* 1394-1409.

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

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

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. Zelazo (Ed.), *The Cambridge handbook of consciousness* (pp. 289-325). 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., & Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review, 103,* 490-517.

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

Koriat, A., Ma'ayan, H., & Nussinson, R. (2006). The intricate relationships 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.

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

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

Kornell, N., Rhodes, M. G., Castel, A. D., Tauber, S. K. (2011). The ease of processing heuristic and the stability bias: Dissociating memory, memory beliefs, and memory judgments. *Psychological Science, 22,* 787 – 794.

Maki, R. H., & McGuire, M. J. (2002). Metacognition for text: Findings and implications for education. In T. Perfect and B. Schwartz (Eds.), *Applied metacogntion* (pp. 39-67). Cambridge: Cambridge UP.

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

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

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

McCabe, D. P**.**, Roediger, H. L., McDaniel, M. A., & Balota, D. A. (2009). Aging decreases veridical remembering but increases false remembering: Neuropsychological test correlates of remember/know judgments. *Neuropsychologia*, *41*, 2164-2173.

McCabe, D. P., Geraci, L., Bowman, J. K., Sensenig, A. E. (2011). On the validity of remember-know judgments: Evidence from think aloud protocols. *Consciousness and Cognition*, *20*, 1625-1633.

Metcalfe, J. (2000). Metamemory: Theory and data. In E. Tulving & F. I. M. Craik (Eds.), *The Oxford handbook of memory* (pp. 197-211). New York: Oxford UP.

Metcalfe, J., and Finn, B. (2008). Evidence that judgments of learning are causally related to study choice. *Psychonomic Bulletin and Review*, *15*, 174-179.

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

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

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

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

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

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

Rhodes, M. G., & Castel, A. D. (2009). Metacogntive illusions for auditory information: Effects on monitoring and control. *Psychonomic Bulletin & Review, 16*, 550-554.

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

Rhodes, M. G., & Tauber, S. K. (2011b). Eliminating the delayed JOL effect: The influence of the veracity of retrieved information on metacognitive accuracy. *Memory, 19*, 853-870.

Soderstrom, N. C., & McCabe, D. P. (2011). The interplay between value and relatedness as bases for metacognitive monitoring and control: Evidence for agenda-based monitoring. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37,* 1236-1242.

Thiede, K., & Anderson, M. (2003). Summarizing can improve metacomprehension accuracy. *Contemporary Educational Psychology, 28*, 129-160.

Toth, J. P., Daniels, K. A., & Solinger, L. A. (2011). What you know can hurt you: Effects of age and prior knowledge on the accuracy of judgments of learning. *Psychology and Aging, 26*, 919-931.

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

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language, 46*, 441-517.

APPENDIX

JORK Study Instructions. A little later in the experiment you are going to study a list of words, and you will take a test for those words. When people remember things, they can experience them in different ways. You are going to distinguish between two different types of memory experiences on the test that you'll be taking. These two types of memory are called Recollection and Knowing. I am going to explain the difference between these two types of memory in some detail now. Please listen carefully.

Recollection is a type of memory that is accompanied by the ability to recall details associated with a past event. For example, if I asked you to remember breakfast this morning you'd likely be able to *recollect* where you were, what you ate, who you ate with, what you talked about, what you were thinking about, and other details. Another way to explain recollection is that it involves *mentally traveling back to the moment that an event occurred.* In this experiment the "events" we're talking about remembering are going to be word pairs that you'll study. When you recollect a word pair you may be able to recall a specific thought that came to mind when you studied the pair, or a mental image that came to mind when you studied it. Or, you may remember a personal association you made, or your emotional reaction to the pair. *The important point is that recollection in this experiment involves bringing to mind some details of what happened, or what was experienced, at the time a word pair was originally studied.*

Knowing is a type of memory where you recognize something as a memory, but you can't remember any specific details about the experience. This is like when you see someone on campus and you know you've met them before but you have no idea where, and can't remember anything else about them. In this experiment, when you believe you studied a word pair but you

cannot consciously recollect any specific details from when you studied the pair earlier, that's the experience of Knowing. In other words, when you *know* a word pair you recognize it as having been studied, but you do not re-experience the exact details of what you were thinking or feeling when you studied it.

The way the study phase is going to work is as follows: you are going to see word pairs presented on the computer screen, one at a time. For each pair you see, I want you to just think of whatever pops into your head related to that word. Spend the full time that the word pair is on the screen thinking about whatever pops into mind about the word pair. Immediately after studying each pair, a screen will come up with the words, "(1) Recollect, (^{2}) Know, or (3) Forget?", just like the screen in front of you right now. For each word pair you will try to predict whether later, on the test you take, you will be able to *Recollect* the word pair, you will just *Know* the word pair was studied earlier, or whether you will *Forget* the word pair later. In other words, for each word pair, you'll be predicting what your future memory for that pair will be like. If you believe you'll be able to Recollect specific details from when you studied the word pair, like the specific thought that came to mind, your emotional reaction, a mental image, or some personal association you made for that pair, you should press the "1" key to predict that you will Recollect the pair. If you do not think you'll be able to recall these sorts of details, but you still believe you'll be able to recognize the word pair as one you studied, you should press "2" key to indicate that you'll Know the pair later. If you think that you won't be able to recognize the word pair as one you studied at all, press the "3" key to indicate you believe you'll forget the word pair. As soon as you make your response, the next word pair will appear, you'll study it, and then decide Recollect, Know, or Forget for that pair too, and so on. (These instructions were adopted from McCabe & Soderstrom, 2011).