DISSERTATION

IS JUDGMENT REACTIVITY REALLY ABOUT THE JUDGMENT?

Submitted by

Sarah J. Myers

Department of Psychology

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Summer 2023

Doctoral Committee:

Advisor: Matthew Rhodes

Anne Cleary
Gwen Fisher
James Folkestad

ABSTRACT

IS JUDGMENT REACTIVITY REALLY ABOUT THE JUDGMENT?

A common research tool used to measure one's understanding of their own learning is to collect judgments of learning (JOLs), whereby participants indicate how likely they are to remember information on a later test. Importantly, recent work has demonstrated that soliciting JOLs can impact true learning and memory, referred to as *JOL reactivity.* However, the underlying cognitive processes that are impacted when learners make JOLs and that lead to later reactivity effects are not yet well-understood. To better elucidate the mechanisms that drive JOL reactivity, I examined how changing the method of soliciting JOLs impacts reactivity. In Experiment 1, I manipulated how long participants had to make their JOLs; in Experiment 2, I compared JOLs made on a percentage scale versus a binary (yes/no) scale; and in Experiment 3 participants were required to explain why they made some of their JOLs. Judgments that require or allow for more in-depth processing (i.e., longer time in Experiment 1, percentage scales in Experiment 2, explaining in Experiment 3) should require more effort from participants to make their judgments. If these more effortful judgments lead to larger reactivity effects, it would suggest that reactivity is driven by processes that occur when making JOLs. However, findings from the experiments did not support this account. Although some differences in reactivity effects were seen after making binary and explaining JOLs compared to percentage JOLs, the hypothesis that more cognitive effort would result in stronger reactivity was not supported. Therefore, results suggest that the mere presence of JOLs during study may cause a general shift in participants' learning approach, resulting in later JOL reactivity.

ACKNOWLEDGEMENTS

I would first like to thank my advisor, Dr. Matthew Rhodes, for his assistance with my experiments and phenomenal mentorship throughout my PhD program. I cannot express how grateful I am to have him as my advisor. I would also like to thank my committee members – Dr. Anne Cleary, Dr. Gwen Fisher, and Dr. James Folkestad – for their intriguing questions, commitment to completing my defense, and enthusiasm for the research project.

I would like to thank the undergraduate research assistants and lab managers that helped me carry out the project. They are truly an irreplaceable resource. Lastly, I would like to thank my partner Xavier as well as everyone else in my personal and work families that were such a positive light and support system as I pursued my PhD. Thank you all.

TABLE OF CONTENTS

INTRODUCTION

When a student sits down to study, their degree of learning is not only based on the amount of information the student reviews, but also reflects their understanding of their own learning and the decisions they make based on that understanding. For example, a student may feel they have learned Unit 1 better than Unit 2. Subsequently, the student may choose to predominantly spend their limited time reviewing Unit 2 instead of Unit 1. Understanding how people assess their own learning is crucial for understanding their true learning and decisions. Research into *metamemory* (i.e., one's judgments and decisions regarding their own learning and memory) sets out to accomplish this goal (see Dunlosky & Tauber, 2016). A common metamemory tool used to measure how learners assess their own learning is to ask participants to make judgments of learning (JOLs) while studying materials (for a review, see Rhodes, 2016). A JOL asks participants to indicate the likelihood of remembering study material on a later test. These judgments of one's learning are then often compared to actual learning (i.e., test performance) to determine how accurately learners can assess their own learning (e.g., Kruger & Dunning, 1999).

However, recent research indicates that it may be inappropriate to compare JOLs and true learning because the act of making JOLs can change learning. Indeed, soliciting JOLs can change a participant's later test performance compared to not making JOLs (e.g., Mitchum et al., 2016; Witherby & Tauber, 2017), a finding termed *JOL reactivity* (see Ericsson & Simon, 1993, for a general discussion of reactivity)[1]. Although a handful of experiments using JOLs suggested

[1]JOL reactivity effects do not include changes in learning based on learning decisions (e.g., JOLs lead to a student realizing they know some material less than others, which causes them to focus future studying on less well-learned information). Instead, JOL reactivity effects refer to cases where, even when study time and decisions are controlled by an experimenter, the presence of JOLs can impact learning.

the possibility of reactivity (e.g., Arbuckle & Cuddy, 1969; King et al., 1980), JOL reactivity effects did not appear as the primary focus of a paper until 2015 (Soderstrom et al., 2015). Since Soderstrom et al.'s (2015) and Mitchum et al.'s (2016) seminal work examining JOL reactivity, research into reactivity has continued apace (see Figure 1). However, findings across studies are mixed and indicate that JOL reactivity effects vary by the particular learning situation. That is, JOL reactivity effects are specific to certain materials and test measures (e.g., Janes et al., 2018; Mitchum et al., 2016; Myers et al., 2020), and there is no conclusive evidence regarding the locus of JOL reactivity effects.

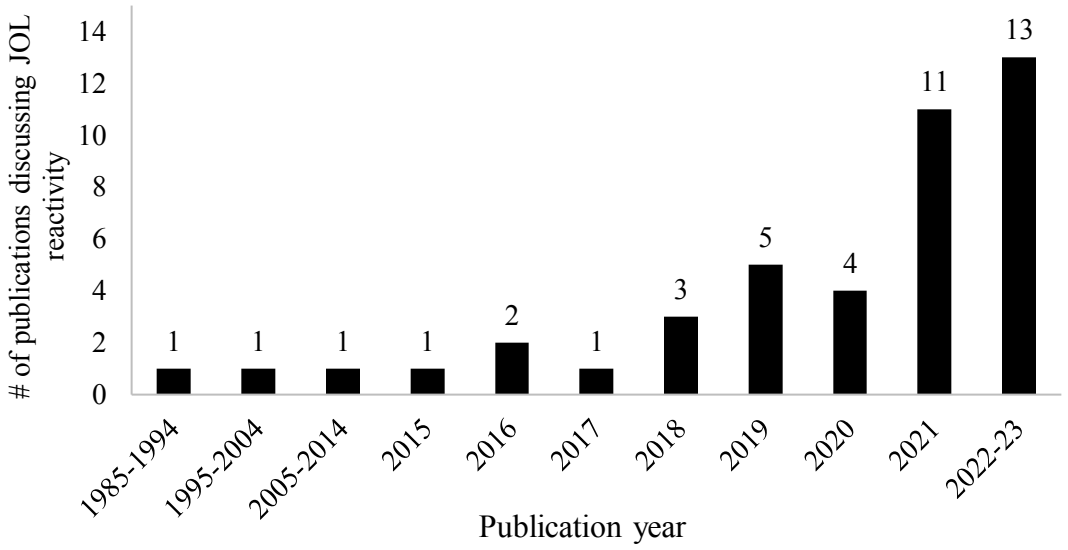

Figure 1. Number of publications discussing JOL reactivity from 1985-2021. The figure is based on a literature search via Google Scholar for publications that included the term "JOL reactivity". It is possible that some studies which discussed JOL reactivity were missed in this literature search.

One question yet to be thoroughly explored is when JOLs impact learning during the learning process. It is possible that JOLs impact learning via specific processes occurring while making that judgment. For example, JOLs might prompt participants to consider details about the study materials that inform the JOLs they make (Koriat, 1997); these details may not have been considered if JOLs were not solicited. Alternatively, the mere presence of JOLs in the study

phase may cause participants to make a global shift in their learning approach or strategies (Mitchum et al., 2016). For example, soliciting JOLs may lead participants to attend more to the study materials rather than passively viewing or reading each study item (but see Rivers et al., 2021). If the presence of JOLs leads to a general shift in learning, then how a JOL is asked should not impact reactivity effects. However, if reactivity partially reflects the cognitive processes that occur when deciding on and making a JOL, then manipulating the nature of the prompt should influence the magnitude and nature of reactivity effects. In the current study, I examined whether JOL reactivity is impacted by different JOL prompts or whether soliciting JOLs in any form causes similar reactivity effects.

JOLs in Metamemory Research

The JOL measure was first developed by Arbuckle and Cuddy (1969). In their seminal work, Arbuckle and Cuddy (1969) asked participants to study lists of word pairs (e.g., *Table – King*) and to indicate for each pair whether they would remember the pair on a later cued recall test (e.g., *Table – ?*). The researchers found that people were able to predict their later test performance above chance levels, indicating that learners have some insight into their state of learning. This then led to a consideration about how these internal judgments of one's state of learning (termed *monitoring*) led to changes in learners' decisions and behaviors (termed *control*; Nelson & Narens, 1994; see Son & Kornell, 2008; Rhodes, 2019, for reviews). In the example provided earlier, a student's realization that they understood Unit 1 better than Unit 2 (monitoring) would then inform their decision to focus study time on Unit 2 (control). Nelson and colleagues (1994) provided evidence of a relationship between monitoring and control by showing that, when given an opportunity to restudy a limited amount of learned vocabulary, participants chose to restudy items accorded low JOLs. This indicates that participants' in-the-

moment judgments inform their study choices and behaviors (see also Ariel et al., 2009; Metcalfe & Kornell, 2005).

Further research has documented flaws that can occur in people's assessments of their own learning, leading to sub-optimal study decisions (e.g., Rhodes & Castel, 2009). First, learners often show overconfidence whereby they believe they will remember more studied material than they do. For example, Fischhoff and colleagues (1977) asked participants trivia questions and had them indicate their confidence that their answers were correct. In their Experiment 1, of the questions for which participants stated they were 100% confident that their answer was correct, they were only truly correct 20-30% of the time (see also Hacker et al., 2000; Kruger & Dunning, 1999, for overconfidence in classroom settings). In addition, participants' assessments are sometimes based on factors that do not reflect true learning. For instance, Carpenter and colleagues (2013) had participants listen to one of two lecture videos. In one video, the instructor presented information fluently (e.g., spoke clearly, knew topic well) and in another video, the instructor spoke disfluently (e.g., used halting speech, looked at notes, little eye contact). Although participants predicted that they learned much less from the disfluent instructor than the fluent instructor, performance on a test of the lecture material found that participants' true learning was not significantly different between the two lecture styles.

From findings demonstrating that JOLs do not always match true learning (e.g., Fischhoff et al., 1977; Glenberg & Epstein, 1995), Koriat (1997) proposed the *cue-utilization framework*, which states that learners infer how well they learned something based on a set of cues available to them while studying. These cues can include information such as the difficulty of the learning material (e.g., Tauber & Rhodes, 2012a) or how fluent the information is to process (e.g., Carpenter et al., 2013; see also Begg et al., 1989), and learners often consider multiple cues to

inform their judgments (Undorf & Bröder, 2020). Some of these cues are related to actual memory strength, such as difficulty of the material (Tauber & Rhodes, 2012a), but other cues are not diagnostic of true memory strength, such as the appearance or nature of study materials (Rhodes & Castel, 2008; 2009).

Research has continued to examine how predictions compare to observed levels of learning. However, this comparison may not be warranted. Prior studies have almost exclusively regarded JOLs as neutral measurements of memory monitoring (see Nelson, 1990). That is, JOLs are expected to reflect an individual's assessment of their learning without exerting any influence on true learning. However, there was little empirical evidence to support this assumption. To verify that JOLs are truly a neutral measure, an experiment would need to compare learning between a group that made JOLs and a group that did not make JOLs. As an example, all participants might study a list of words in preparation for a later test; half of the participants may simply study the words at a fixed rate, whereas the other half may make a JOL for each word they study. Memory for the studied words can be compared between those who made JOLs during study and those who did not (i.e., a no JOL group). If differences in memory are not detected, then adding JOLs could be considered a neutral measure. Historically, few studies have included this necessary comparison.

The few early experiments that included comparisons of no JOL and JOL groups (although not the focus of the study) found mixed results regarding JOL reactivity. Some did not detect a difference in learning between those who made versus did not make JOLs, suggesting JOLs may be a neutral measurement (e.g., Begg et al., 1992; Keleman & Weaver, 1997). However, other studies did demonstrate evidence of JOL reactivity: soliciting JOLs changed participants' learning and memory compared to not requiring JOLs (e.g., Arbuckle & Cuddy,

1969; King et al., 1980). Thus, a focus on understanding reactivity was needed, but this need was not met until less than ten years ago.

JOL Reactivity

Empirical findings of JOL reactivity

Although a few papers had theorized the possibility of JOL reactivity (e.g., Spellman & Bjork, 1992; Rhodes, 2016), the first experiments exclusively focused on JOL reactivity were not published until 2015 (Soderstrom et al., 2015) with nearly 40 empirical papers focused on the topic appearing since (see Figure 1). JOLs can be solicited immediately after studying each item or after a delay from initial studying. I will focus my discussion on reactivity with immediate JOLs (see Kubik et al., 2022; Rhodes & Tauber, 2011; Tauber et al., 2015; Tekin & Roediger, 2021, for discussions of reactivity with delayed JOLs). In Soderstrom and colleagues' (2015) seminal study, they had participants study mixed lists of related (*Buzz – Bee*) and unrelated (*Table – King*) word pairs. While studying, some participants made JOLs for each word pair and others did not. On a later test for which participants were given the first word and had to provide the second (*Buzz – ?*), those who made JOLs correctly recalled more of the studied related word pairs than participants who did not make JOLs (although JOLs did not impact recall of unrelated word pairs). Among further studies examining reactivity with immediate JOLs, findings have remained mixed, with some work observing differences in test performance between JOL and no JOL conditions (e.g., Senkova & Otani, 2021) and some observing no differences (e.g., Ariel et al., 2021; Schäfer & Undorf, 2023).

One factor that appears to drive these divergent findings is differences in the materials being studied. For example, Senkova and Otani (2021) found that JOLs impacted memory of single-word lists (but see Tauber & Rhodes, 2012b) while Ariel and colleagues (2021) found that

JOLs did not impact performance on tests of reading passages. Even within single studies, reactivity differences appear based on type of studied materials. As described previously, Soderstrom and colleagues (2015) found that JOLs improved recall of related word pairs (mean percent correctly recalled on test[2] – JOL condition: 74%, no JOL condition: 55%) but did not influence recall of unrelated pairs (JOL condition: 20%, no JOL condition: 19%). Indeed, in a meta-analysis of eight independent studies, Double and colleagues (2018) determined that JOL reactivity occurred for related word pairs (g = 0.32) and single-item word lists (g = 0.38) but not for unrelated word pairs (g = -0.01)[3].

JOL reactivity effects also diverge based on how memory is measured on the final test (Myers et al., 2020). For example, JOLs improve memory of related pairs when these pairs are tested using cued recall, but not if tested using free recall (i.e., recall all the targets that were studied). JOL reactivity effects even seem sensitive to experimental design. Rivers et al. (2021) demonstrated that JOL reactivity occurs when JOL/no JOL conditions are manipulated between-lists (i.e., JOLs are made for every item in one list and no JOLs are made for another list) but not within-lists (i.e., JOL prompts appear for some words but not for others in each study list). Additionally, Janes et al. (2018) found that reactivity effects were reduced when JOLs were made for pure lists (i.e., only related pairs or only unrelated pairs; but see Maxwell & Huff, 2022). Overall, it is evident that JOL reactivity does not appear in every learning situation.

Accounts of JOL reactivity

Three main theories have been proposed so far that attempt to explain these differences in JOL reactivity based on materials, tests, and other manipulations. First, Soderstrom and colleagues (2015) accounted for material-specific reactivity (e.g., reactivity occurs for related but

[2]Means are estimated from Soderstrom et al.'s (2015) Figure 1b.
[3]Note that many other JOL reactivity studies have been added to the literature since this meta-analysis.

not unrelated pairs) by considering how the cue-utilization hypothesis (Koriat, 1997) may predict reactivity effects. Specifically, Soderstrom et al. (2015) suggested that making JOLs strengthens memory for the specific cues that learners use to inform their JOLs. If memory on a later criterion test is sensitive to the same cues, then making JOLs should improve performance on that test. For example, learners use relatedness between two words in a pair (e.g., *Buzz – Bee* vs. *Table – King*) as a cue when making JOLs, giving related word pairs higher JOLs than unrelated pairs (see Mueller et al., 2013, for a review). Soderstrom et al. (2015) proposed that when participants attend to relatedness to inform their JOLs they strengthen encoding of the relationship between the items in related pairs (e.g., for the pair *Buzz – Bee,* buzz is the sound a bee makes). However, this relational processing does little to enhance encoding for unrelated pairs, which have no semantic relationship (e.g., for the pair *Table – King,* there is no inherent relationship between these items). When tested later with a cued recall test (*Buzz – ?*), operations that strengthen cue-target relationships (such as making JOLs) should enhance performance. Soderstrom et al.'s (2015) findings supported this hypothesis by finding that JOLs elevated cued recall of related pairs but showed no influence on unrelated pairs. Myers et al. (2020) also provided support for this hypothesis, finding that JOLs did not increase related pair memory when measured via free recall (i.e., recall all studied words), a test that is much less dependent on cue-target relationships. Additionally, Halamish and Undorf (2022) found that the presence of JOLs increased participants' likelihood of selecting what type of pair an item was studied with (i.e., related, unrelated, or identical pair).

Mitchum and colleagues (2016) proposed a different mechanism to explain JOL reactivity. They argued that JOLs draw attention to the fact that some materials (e.g., related pairs) are easier to learn than others (e.g., unrelated pairs). Thus, JOLs cause participants to shift

from trying to learn all study materials to a more selective learning goal, whereby they focus on learning easy items at the expense of more difficult items. In support of this account, Mitchum and colleagues (2016) used a similar procedure to Soderstrom et al. (2015) and also found that JOLs bolstered cued recall performance for related pairs. However, contrary to Soderstrom et al.'s (2015) findings, they reported that JOLs impaired memory of unrelated pairs compared to not making JOLs, suggesting that making JOLs caused participants to spend more time and attention on learning related pairs and less attention on learning unrelated pairs.

Senkova and Otani (2021) have suggested an additional mechanism that drives JOL reactivity for single-item word lists (e.g., *Bee, King, Cat*) rather than word pairs (e.g., *Table – King*, *Buzz – Bee*). Senkova and Otani (2021) argue that making JOLs on an item-by-item basis increases item-specific processing. That is, learners focus on information based on each word individually, such as a personal connection to or mental image of the item, thus increasing its distinctiveness (Hunt, 2006; 2012). However, this increased focus on each individual item reduces relational processing, whereby connections are made between all the items seen within a list, such as recognizing that several items in a study list come from a common category (e.g., animals). Thus, Senkova and Otani (2021) propose that JOLs specifically increase item-specific processing but reduce relational processing. In support of this account, Senkova and Otani (2021) demonstrated that JOL reactivity was stronger when participants studied a categorized list of words (e.g., *tiger, horse, potato, squash*) than when participants studied an uncategorized list. They argued that this occurred because categorized lists already encourage relational processing (i.e., noticing that items across the list belong to the same category). Thus, when JOLs are introduced and increase item-specific processing, later test performance is boosted by both strong relational processing and item-specific processing during study. Uncategorized lists

already encourage item-specific processing (because there are no intuitive connections between different items), so making JOLs and increasing item-specific processing is redundant; hence, JOLs do not improve memory of uncategorized lists compared to not making JOLs. Moreover, Zhao et al. (2023) demonstrated that JOLs reduced performance on a serial order test, which asks participants to recall items they studied in the order they had studied them and relies heavily on relational processing.

Of note, theories of JOL reactivity are still emerging and include a number of assumptions that are still being examined (e.g., Janes et al., 2018; Myers et al., 2020; Rivers et al., 2021; Zhao et al., 2021), suggesting that the theories will most likely continue to evolve. Nevertheless, all three theories predict that the act of making JOLs changes the way learners process study materials, whether by focusing on specific details of the materials (Soderstrom et al., 2015), changing one's learning goals (Mitchum et al., 2016), or increasing focus on each item individually (Senkova & Otani, 2021). However, it is unclear whether these changes in processing occur specifically while making each JOL or whether soliciting JOLs causes a global shift of one's approach to learning during the study episode. For example, reactivity may be restricted to the time of making each JOL if participants consider specific details only while making the JOLs, such as processing the relationship between word pairs to inform their JOLs, as suggested by Soderstrom and colleagues (2015). Alternatively, JOLs may encourage a global shift in processing during the entire study episode, such as leading participants to focus on each study item separately and limiting focus on relationships between items, as suggested by Senkova and Otani (2021). To better understand when and how JOLs change learning and memory processes, more research is needed to determine whether the way JOLs are collected impacts their effect on later memory. If the type of JOL prompt impacts reactivity effects, then

theoretical approaches must incorporate this information to account for a more specific locus of the effect versus a global change in how material is encoded when JOLs are present.

Types of JOLs

The premise of a JOL is to ask participants to predict how likely it is that they will remember certain information on a later test. However, researchers have used many different methods to ask this question. One difference is the scale participants use for their JOLs: Some studies use Likert scales, such as selecting a number between 1 and 5 from "very unlikely" to "very likely" to remember each item (e.g., Arbuckle & Cuddy, 1969); others ask participants for a probability of remembering from 0 to 100% (e.g., Koriat, 1997); still others use a binary judgment (*Will you remember this? yes or no*; e.g., Saenz & Smith, 2018). Researchers have also compared different framings of JOL prompts. For example, rather than asking for a probability of remembering, Tauber and Rhodes (2012b) asked participants how long they thought they would remember each item, termed judgments of retention. Rawson and colleagues (2002) asked participants to rate their understanding of materials rather than performance on a later test, and other studies have asked participants how likely they were to *forget* each item (i.e., judgments of forgetting; JOFs) rather than *remember* each item (Finn, 2008; Halamish et al., 2011; McCabe & Soderstrom, 2011; Serra & England, 2012; Soderstrom & Rhodes, 2014).

These different judgment prompts impact how people use the prompts. For example, Serra and England (2012) found that participants were more confident in their memory when the judgment was made in terms of forgetting compared to remembering (see also Koriat et al., 2004; Kornell & Bjork, 2009). One reason for these differences may be based on how learners use the JOL/JOF scales. For example, England and Serra (2012) suggested that participants start by setting an anchor on a scale (e.g., a base judgment of 40% for each item) and then adjust their

judgment up or down based on other information that is available to them about the item (e.g., an item feels easier than others). When judgments are framed in terms of forgetting, this appears to change the initial anchoring value learners based their judgments on (England & Serra, 2012; England et al., 2017).

Researchers have also shown that even changing the values on JOL scales can impact which cues inform participants' decisions, thus influencing judgment accuracy (Hanckzakowski et al., 2013; Higham et al., 2016; McGillivray & Castel, 2017; Zawadzka & Higham, 2015). For example, Zawadzka and Higham (2015) found that, in a learning situation of multiple study-test cycles, binary JOLs (e.g., 0% vs. 100%) made after each study episode reflected true memory more accurately than 0-100% JOLs. In addition, certain (more accurate) cues were used more often when making binary JOLs than percentage JOLs.

These findings all reveal that interpreting participants' responses on these subjective measures is complex. Although the same general trends appear in how participants use subjective scales (e.g., higher numbers often reflect higher accuracy; Tekin & Roediger, 2017), interpreting fine-grained responses proves more difficult because participants' use of the scales does not always reflect what the scale is intended to measure. For example, when estimating probabilities of events, participants might use 50% as an indicator of "I don't know" rather than an actual prediction of a 50% likelihood of the event occurring (De Bruin & Carman, 2018; De Bruin et al., 2000). In the context of learning predictions, Zawadzka and Higham (2015) propose that percentage JOLs reflect participants' confidence in their future memory rather than the true probability they will remember an item or not (cf. Serra & DeMarree, 2016, in the context of students' predictions of future exam performance). Discussions continue regarding how participants map their subjective experiences onto the scales used in research and how their

mapping changes based on how the question is asked (see e.g., Higham et al., 2016; Jersakova et al., 2017). Thus, JOL scales may change how learners predict their memory even though, in practice, all metamemory judgments should measure the same underlying construct regardless of how it is collected.

Based on this evidence, the way JOLs are collected may impact learners' consideration and use of those judgments. However, past studies that have demonstrated divergences based on JOL prompts (e.g., Serra & England, 2012; Zawadzka & Higham, 2015) have only considered the accuracy of learners' judgments (i.e., comparing their judgments to actual performance) and not how these different JOL prompts might impact later memory of the learned materials. Thus far, only one study has compared the effects of different JOL scales on reactivity effects in an experiment (Mitchum et al., 2016, Experiment 3).

In Mitchum et al.'s (2016) Experiment 3, participants studied related and unrelated word pairs then either made JOLs on a 0-100% scale, made binary (yes/no) JOLs, or did not make JOLs (no JOL condition). Importantly, study was self-paced, meaning participants could study each word pair as long as they wished to they provided their JOL and/or moved onto the next item. (To anticipate, the current experiments used experimenter-paced study times.) In their experiment, Mitchum et al. (2016) found that making either percent or binary JOLs led participants to shift their study goals to focus more time on related pairs and less on unrelated pairs. Subsequently, JOLs (percent and binary) increased the discrepancy between memory of related and unrelated pairs, suggesting that requiring JOLs caused participants to focus on learning related pairs at the expense of unrelated pairs. This discrepancy was more prominent for binary JOLs than percentage JOLs (mainly driven by binary JOLs having a stronger positive effect on related pair memory), which may suggest that the type of JOLs influences reactivity

effects. However, the authors did not draw strong conclusions based on their comparison of percentage and binary JOLs. To more closely examine the impact of how JOLs are solicited, in the present study, I used several different prompts to solicit JOLs across three experiments. This provides a better understanding on how those different prompts might change learners' consideration of the study materials and how that change might impact later memory.

The Current Study

The current study consisted of three experiments, each using a common experimental procedure for investigating JOL reactivity (e.g., Soderstrom et al., 2015; Janes et al., 2018). Participants studied several lists of word pairs comprised of intermixed related (*Buzz – Bee*) and unrelated (*Table – King*) pairs. After studying each list, participants completed a cued recall test whereby they saw the first word of each pair they had studied and were prompted to type in the second word (e.g., *Buzz – ?*). Participants made JOLs for each item in some of these lists (JOL conditions) but not for others (no JOL condition). If JOL reactivity occurs, participants' cued recall performance should be different between the JOL and no JOL conditions. My primary interest was comparing the size of JOL reactivity (i.e., the difference in test performance between JOL and no JOL conditions) for different types of JOLs. For example, I determined whether the reactivity size differs for memory of related versus unrelated pairs. Based on prior findings (e.g., Soderstrom et al., 2015; Myers et al., 2020), I expected that reactivity size would be larger for related pairs than unrelated pairs.

Critically, in addition to comparing reactivity differences between types of pairs, the present study also examined the size of reactivity based on different JOL prompts. In Experiment 1, I manipulated the time participants had to type in their JOLs, which should change the amount of time participants have to deliberate on their judgments. Thus, if reactivity is driven by the act

of making JOLs, reactivity should be stronger when participants are given more time to consider their judgment. However, if requiring JOLs causes a shift in one's approach to the entire study episode and is not isolated to the act of making each JOL, reactivity effect sizes should not differ based on how long participants are given to make that judgment.

In Experiments 2 and 3, I kept the amount of time to make JOLs consistent and changed the complexity of the JOL and potentially the cognitive processes or amount of deliberation used to inform each JOL. For Experiment 2, participants made JOLs for some lists using the traditional 0 to 100% scale but made JOLs for other lists on a binary scale (*i.e., I will/will not remember this item*) as a conceptual replication of Mitchum et al.'s (2016) experiment. In Experiment 3, all JOLs were made on the traditional 0 to 100% scale but, for some study lists, participants completed an additional step whereby they selected from a list of options regarding why they chose that specific number for each JOL (similar to methods used in Jersakova et al., 2017).

In all three experiments, the more elaborate JOL (i.e., 4s JOLs in Experiment 1, 0-100% JOLs in Experiment 2, explaining reasoning in Experiment 3) should require or allow for more deliberation and processing of the study materials, whereas shorter judgments (e.g., binary JOLs) should restrict the amount of thought participants put into each judgment. If JOLs impact learning based on what participants consider while making their JOLs, then more elaborate JOLs should lead to larger reactivity effects compared to simpler JOLs. However, if requiring JOLs causes a general shift in processing toward the entire study episode, then changes to the JOL prompt should not impact the size of reactivity effects.

EXPERIMENT 1

In Experiment 1, I explored whether the time to make a JOL impacts the size of reactivity on a later cued recall test. Participants provided JOLs for two study blocks and did not provide JOLs for the other two blocks. For one of the JOL lists, participants were given 2 seconds to enter each of their JOLs and for another list participants had 4 seconds to make their JOL. To control for overall study time, each JOL study condition was compared to a no JOL condition whereby participants were given an equivalent amount of total study time (see details in Procedure). Having less time to make JOLs (i.e., 2s compared to 4s) should reduce the amount of processing and deliberation participants can put into each JOL. If this reduced time (and processing) diminishes reactivity effects, it would suggest that reactivity at least partially reflects processes operating while participants make their judgment. If judgment time does not impact reactivity size, it would suggest that the presence of JOLs causes a general shift in participants' processing or approach to learning throughout the entire study episode.

Method

Participants

A power analysis indicated a sample size of 67 using an expected effect size of $d = 0.35$ for a two-tailed paired-samples *t*-test, power of .80, and an alpha of .05. Based on a small meta-analysis of experiments reported in Myers et al. (2020), which used similar materials and procedures to the current study, I expect an overall reactivity effect of $d = 0.52$ between JOL and no JOL conditions for cued recall of related pairs[4]. To be more conservative, I planned to power the studies for a smaller effect size of $d = 0.35$. The power analysis used a paired-samples *t*-test

[4]The experiments tested cued recall of both related and unrelated pairs. However, based on past research (e.g., Myers et al., 2020), I expected reactivity effects to be exclusive to related pairs.

because my main analysis of interest is the difference in reactivity size between the 2s-JOL and 4s-JOL conditions for each pair type[5].

Participants were recruited via the Colorado State University (CSU) subject pool. Of those recruited, 136 participants began the study. A majority of participants completed the study online (n = 114), but some participants also completed the experiment in-person (n = 22)[6]. The experiment took approximately 50-55 minutes and participants who completed the experiment received one research credit for their participation. Fifty-two participants were removed because they did not complete all blocks of the experiment (n = 30), technical difficulties (n = 2), they did not respond to at least 21 (70%) of the 30 JOLs (n = 19) in at least one of the blocks, or they did not respond to at least 15 (50%) of the test items in at least one block (n = 1). Thus, only 62% of the respondents (59% of online participants, 81% of in-person participants) provided useable data. The final sample comprised 84 individuals (17 in-person, 67 online). Of those participants, 60 identified as female, 23 as male, and 1 as non-binary. Participants were between 17 and 29 years old (M = 19.4, SD = 1.82).

Design

This study used a 2 (judgment: JOL, no JOL) x 2 (total study time: 8s, 10s) x 2 (type of pair: related, unrelated) within-subjects design. There were four study-test blocks for each participant. Type of pair (related, unrelated) was manipulated within each block while total study time and judgment were manipulated between blocks. Thus, participants had one study-test block

[5]In all experiments, the sample size was larger than the initially proposed sample (67) because data collection was terminated on a set date rather than the number of participants collected. In addition, extra participants were recruited to account for anticipated attrition.

[6]Analyses suggested that the main comparisons did not differ between online and in-person participants (see Appendix C for full analyses), so data were collapsed across the two collection groups. However, the size of each group was underpowered and uneven, so differences might not have been statistically detectable.

for each of the following combinations: 1) JOL-8s, 2) no JOL-8s, 3) JOL-10s, 4) no JOL-10s. The order in which learners received these conditions was randomized for each participant[7].

Materials

One hundred twenty related cue-target word pairs (forward strength 0.40 - 0.74, M = 0.50, SD = 0.09) selected from the University of South Florida Free Association Norms (USF-FAN; Nelson et al., 1998) were used in Experiment 1. Based on the MRC Psycholinguistic Database (Version 2.00), target characteristics were as follows – frequency: 6.40-13.55 (M = 9.94, SD = 1.30), concreteness: 250-670 (M = 531.7, SD = 102.16), length: 3-8 letters (M = 4.75, SD = 1.15). Pairs were then divided into eight lists of 15 related pairs that were closely matched in average forward association, as well as target frequency, concreteness, and length. An unrelated version of each of the lists was created by randomly pairing the targets from each list with unrelated cues from another list. Lastly, eight lists of 30 pairs each (15 related, 15 unrelated) were created and counterbalanced so that target words were paired with a related cue for half the participants and an unrelated cue for the other half. For example, the target *Bee* was paired with a related cue (*Buzz*) for some participants and an unrelated cue (*Elbow*) for others. The lists were also counterbalanced so that each target appeared with each of the different judgment and study time conditions. Sixteen other related word pairs were used as buffers. Half the buffer pairs were randomly re-paired to make unrelated buffers. The experiment was conducted in Qualtrics (Qualtrics, 2020).

[7]Because conditions were manipulated within-subjects, order effects were possible. Order effect analyses and analyses using only participants' first block condition (i.e., eliminating any effects from exposure to other conditions) are available in Appendix D. Analyses largely indicated that order effects did not change the main patterns of findings (although these between-subject comparisons were underpowered).

Procedure

Procedures for all experiments were approved by the Colorado State University IRB before data collection. After providing consent, participants were informed that they would learn several lists of word pairs (e.g., *Buzz – Bee)* for later tests. They were also instructed that on the test they would be given the first word of each pair (e.g., *Buzz – ?*) and asked to type in the second word (e.g., *Bee*). After these initial instructions, participants proceeded with four study-test blocks. For each block, participants studied word pairs, solved addition problems, and then took a cued recall test over the word pairs. During study, 30 word pairs (15 related and 15 unrelated) were presented in a unique random order for each participant, with a 250ms interstimulus interval (ISI) between each pair. In addition to the 30 key pairs, two buffer pairs were added to the beginning and two to the end of each list to account for primacy and recency effects, resulting in 34 pairs total. Buffer pairs were not included on the later tests or in any analyses. After studying the word pairs, participants solved addition problems for 3 minutes and then began a cued recall test. On the test, participants saw each of the 30 cues one-at-a-time and had 7 seconds to type each corresponding target word. The cues were presented in a unique random order for each participant that was different than the order in which pairs were studied; guessing was encouraged, but no feedback was provided on the tests. Participants then completed another three blocks of studying 34 new pairs, solving math for 3 minutes, and then receiving a cued recall test on the block's pairs, with each block using different manipulations of study time and judgment.

For two of the blocks, each pair was presented for 8s during study; for the other two blocks, each pair was presented for 10s (see Figure 2 for a depiction of the different conditions). For one of the 8s blocks, participants simply studied each pair for 8s (no JOL-8s). For the other

8s block, participants provided a JOL for each pair they studied (JOL-8s): they saw the pair presented alone for 6s and then the JOL prompt appeared for the last 2s while the pair remained on the screen. Thus, participants had 2s to make their JOL for each pair. The JOL prompt stated "From 0-100%, how likely is it that you will remember this pair on a later test?".

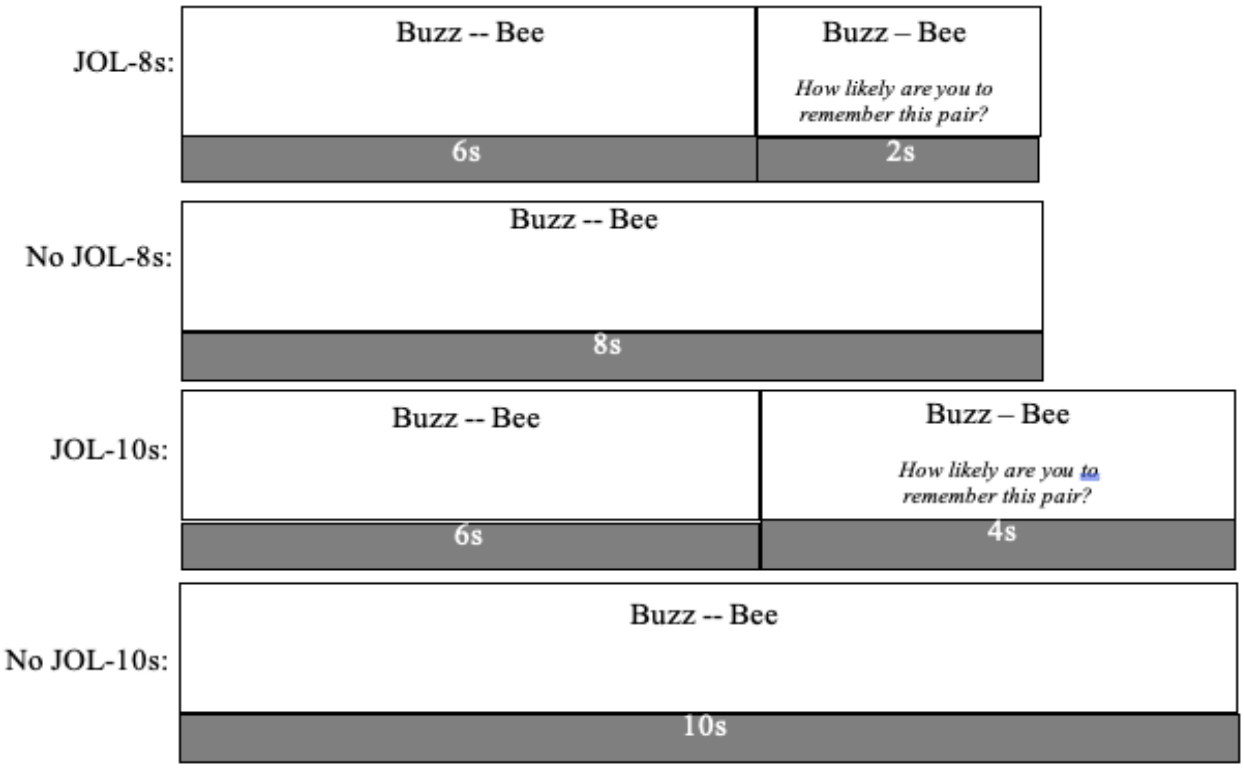

Figure 2. Schematic of the design for the four study conditions used in Experiment 1. For two blocks, participants see each study pair for 8s; for one of these two blocks participants also provide a JOL during the last 2s. For the other two blocks, participants see each pair for 10s, and in one of the blocks participants provide a JOL during the last 4s.

Similarly, one of the 10s blocks presented each pair for 10s without asking for any judgments (no JOL-10s) and the other block displayed only the pair for 6s then the pair and JOL prompt for 4s (JOL-10s). The 2s and 4s times were chosen for the JOL response deadlines because participants tended to make JOLs in 2.5-3s based on a prior JOL reactivity study (Myers

et al., 2020) that used a similar design. Therefore, 2s should introduce some time pressure on participants to make their JOL faster than they typically would, whereas 4s should result in less time pressure on participants. The order in which participants completed the four conditions was randomized.

Comparing these four conditions controls overall study time while manipulating only time given to make JOLs. To elaborate, in both JOL conditions, participants were given 6s to study the pair without being prompted to provide a JOL; then, one condition was given 2s with the prompt (8s total) and another condition was given 4s with the prompt (10s total). The JOL-10s condition included two more seconds overall than the JOL-8s condition. Thus, I compared participants' cued recall from each condition to a no JOL condition that controlled for that time difference. Reactivity size when participants were only given 2s to provide a JOL was calculated by subtracting average cued recall of the no JOL-8s list from the JOL-8s list. Similarly, reactivity size when given 4s to provide a JOL was calculated by subtracting cued recall of the no JOL-10s list from the JOL-10s list.

Scoring and analysis

Minor spelling mistakes were marked as correct provided the response was not a different word (e.g., "For" instead of "Fog" would be marked as incorrect). Plurals of target words were also marked as correct. Data were analyzed using JASP Version 0.16.2 (JASP, 2022).

For each analysis, I report the *p*-value, a standardized effect size measure (Cohen's *d* or η^2_p), and the Bayes factor (*BF*). Bayes factors give a ratio of the strength of the evidence in favor of the alternative hypothesis (i.e., differences in cued recall performance) relative to the null hypothesis (i.e., no difference; see Kruschke, 2013, for a discussion of Bayes factors). A Bayes factor of 1 means that the data are equally likely under the alternative and null hypotheses.

Unlike null hypothesis significance testing, Bayes factors can indicate that the null hypothesis is more probable than the alternative hypothesis (i.e., when $BF_{10} < 1$) and are reported as the reciprocal BF_{01}. Thus, BF_{10} indicates that the data are more probable if the alternative hypothesis were true, whereas BF_{01} indicates the data are more probable under the null hypothesis. Bayesian calculations require a prior, which gives a range of plausible effect sizes for the alternative hypothesis. I used the JZS prior when calculating Bayes factors because it requires the fewest assumptions about the range of the true effect size (Rouder et al., 2009).

Results

Judgments

Analyses of participants' provided JOLs are presented in Appendix A. For all experiments, participants gave higher JOLs to related pairs than unrelated pairs. Response times to provide a JOL were measured using participants' recorded reaction time of their first click (which should have corresponded to clicking on the JOL response box in most trials). Given that study trials induced more time pressure when JOLs had to be made within 2s rather than 4s, I expected participants' response times to be faster with the shorter 8s total study time (2s JOL) deadline. A paired-samples *t*-test confirmed that participants provided faster responses in the JOL-8s condition ($M = 1.02$s, $SE = .03$) than in the JOL-10s condition ($M = 1.73$s, $SE = .05$), $t(83) = 12.62$, $p < .001$, $d = 1.38$, $BF_{10} = 9.07 \times 10^{17}$. Thus, the manipulation was effective as participants provided their JOLs more rapidly when given a shorter response deadline.

JOL vs No JOL

A 2 (judgment: JOL, no JOL) x 2 (study time: 8s, 10s) x 2 (type of item: related, unrelated) repeated-measures analysis of variance (ANOVA) was performed on participants'

cued recall performance (see Figure 3). On average, participants recalled more related pairs (M = 82.92, SE = 2.06) than unrelated pairs (M = 33.51, SE = 2.06), $F(1,83) = 578.43$, $p < .001$, η^2_p = .88, $BF_{10} = 8.58\text{x}10^{35}$. Time did not significantly impact recall, $F(1,83) = 2.47$, $p = .12$, $\eta^2_p = .03$, $BF_{01} = 2.64$, nor did time interact with judgment or type of pair, p's $\geq .33$, BF_{01}'s ≥ 4.93. Although not significant and the Bayes factor favored the null, participants recalled more pairs for which they had made JOLs (M = 59.25, SE = 1.89) compared to not making JOLs (M = 57.18, SE = 1.89), $F(1,83) = 2.90$, $p = .09$, $\eta^2_p = .03$, $BF_{01} = 2.01$. The three-way interaction also was not significant, $F(1,83) = 0.002$, $p = .96$, $\eta^2_p < .01$, $BF_{01} = 7.24$.

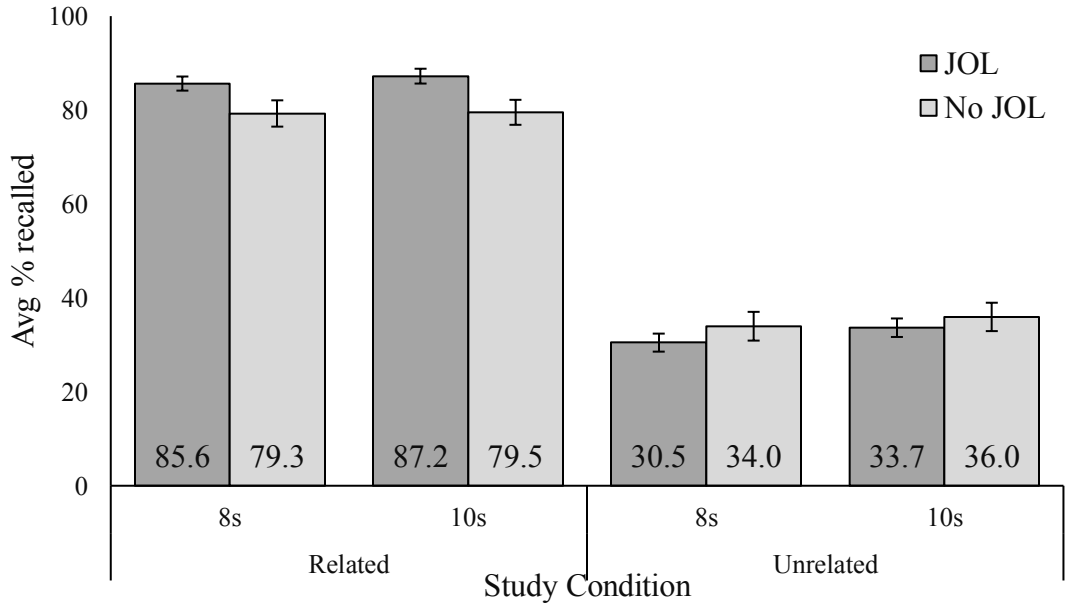

Figure 3. Average percent of related and unrelated word pairs correctly recalled after studied for 8s or 10s while either making JOLs or not making JOLs (No JOL). Errors bars represent 1 standard error of the mean.

However, there was a significant judgment x pair interaction, $F(1,83) = 30.19$, $p < .001$, $\eta^2_p = .27$, $BF_{10} = 2.53\text{x}10^4$. Collapsed across 8s and 10s study time blocks, JOLs significantly improved recall of related pairs, $t(83) = 5.59$, $p < .001$, $d = 0.61$, $BF_{10} = 4.72\text{x}10^4$. JOLs also reduced recall of unrelated pairs, although this was not significant and the Bayes factor favored

the null, consistent with the small effect size that characterized the difference, $t(83) = -1.68$, $p = .10$, $d = -0.18$, $BF_{01} = 2.17$.

8s vs 10s JOLs

Reactivity size (RS) was calculated for each participant by subtracting their performance for each no JOL condition from their performance in the corresponding JOL condition (e.g., related items from JOL-8s to related items from no JOL-8s; see Figure 4). A 2 (study time: 8s, 10s) x 2 (type of pair: related, unrelated) repeated-measures ANOVA was conducted to examine the effects of JOLs on memory. There was a main effect of pair, such that making JOLs (collapsed across time to make judgment) was beneficial for related pairs ($M = 7.02$, $SE = 1.51$) but harmful for unrelated pairs ($M = -2.90$, $SE = 1.51$), $F(1,83) = 30.19$, $p < .001$, $\eta^2_p = .27$, $BF_{10} = 1.91 \times 10^4$. The main effect of judgment was not significant, $F(1,83) = 0.42$, $p = .52$, $\eta^2_p = .01$, $BF_{01} = 5.51$, nor was the judgment x pair interaction, $F(1,83) = 0.002$, $p = .96$, $\eta^2_p < .01$, $BF_{01} = 7.08$. For completeness, follow-up paired-samples *t*-tests were still conducted. These verified that type of JOL did not impact reactivity size for related pairs, $t(83) = 0.57$, $p = .57$, $d = 0.06$, $BF_{01} = 4.59$, or unrelated pairs, $t(83) = 0.44$, $p = .66$, $d = 0.05$, $BF_{01} = 4.73$.

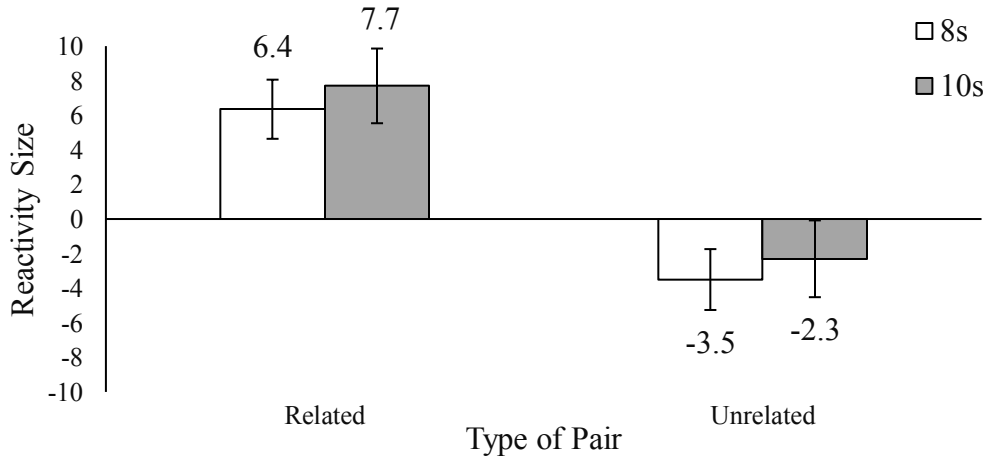

Figure 4. Average reactivity size (performance on recall test after not making JOLs during study subtracted from performance on recall test after making JOLs for the respective total study time) from Experiment 1 cued recall tests. Data are separated based on type of pair and the total study time given. 8s study time corresponds to 2s to make JOLs; 10s study time corresponds to 4s to make JOLs.

Discussion

Following prior findings (e.g., Soderstrom et al., 2015), there was an interaction between judgment and type of item, such that making JOLs increased cued recall of related pairs compared to not making JOLs, while there was only a minimal (non-significant) decrement to cued recall of unrelated pairs when participants made JOLs. However, the focus of the experiment was whether time to make JOLs would impact later recall. Analyses indicated that being given 2s vs. 4s to make JOLs did not impact JOL reactivity effects for either related or unrelated pairs, with Bayes factors supporting the null hypothesis (i.e., no differences in reactivity size) was 4-5 times more likely than the alternative (i.e., differences in reactivity size).

It was hypothesized that inducing time pressure to make JOLs (i.e., 2s instead of 4s) would lead participants to make faster JOL decisions, with potentially less overall deliberation for the JOL. If these differences in mental effort to make JOLs impacted the effect of JOLs on later memory, it would suggest that reactivity is driven by changes in cognitive processes that occurred at the time of making judgments. However, the response deadline for JOLs did not significantly impact later reactivity effects in Experiment 1. This might suggest that JOLs cause a general shift in participants' approach to learning, rather than reflect specific processes that occur while making JOLs.

EXPERIMENT 2

In Experiment 2, participants were given 4s to make all JOLs (i.e., study time was held constant). The key manipulation was how the JOL was asked. For one study list, participants did not make any judgments (i.e., no JOL condition) whereas participants made JOLs on the 0-100% scale used in Experiment 1 for another list. For a third study list, participants made binary JOLs, whereby they indicated *yes* or *no* regarding whether they would remember each pair for the later test. Binary JOLs are less elaborate judgments and thus may reduce participants' required effort to make each JOL. More specifically, participants must only decide between two options (they will remember the pair or not) rather than differentiating between whether they are 60%, 70%, or 80% likely to remember the pair. Because percent JOLs require more fine-grained distinctions than binary JOLs, I expected that participants would spend more effort to make these percent JOLs. Thus, JOL reactivity may be larger when making percent JOLs compared to binary JOLs if reactivity is driven by processing effort specific to making JOLs. In contrast, reactivity sizes should be similar if reactivity is not driven by the degree of processing that occurs specifically while making the judgments.

Experiment 2 replicated the comparison used in Mitchum et al. (2016) Experiment 3. Specifically, cued recall performance was compared after participants made percent JOLs, binary JOLs, or no JOLs during study. However, in the current experiment, study was experimenter-paced (rather than subject-paced) to isolate JOL effects on memory without allowing for JOL effects on study time.

Method

Participants

One hundred thirty-two (132) new participants were recruited via the CSU participant pool. All completed the study in-person and received one course credit. Of these, ten participants were removed because they did not complete all study-test blocks (n = 2), technical difficulties (n = 1), they did not provide at least 70% of JOLs (n = 4) or did not respond to at least 50% of items on the tests (n = 3) for at least one of the blocks. Thus, 122 participants (92%) provided useable data and were included in analyses. Of those 122 participants, age ranged from 17 to 49 (M = 19.7, SD = 3.5); 80 participants identified as female, 39 as male, and 3 as non-binary.

Design

This study used a 3 (judgment: binary JOL, percent JOL, no JOL) x 2 (type of pair: related, unrelated) within-subjects design. There were three study-test blocks for each participant. Type of pair (related, unrelated) was manipulated within each block while judgment type was manipulated between blocks.

Materials

Ninety word pairs were selected from those used in Experiment 1 (forward strength 0.40 to 0.54, M = 0.46, SD = 0.04) – Target frequency: 6.40-13.55 (M = 9.91, SD = 1.40), concreteness: 250-670 (M = 529.76, SD = 100.09), length: 3-8 letters (M = 4.92, SD = 1.18). Pairs were re-divided into six lists of 15 pairs that were closely matched in average forward association, as well as target frequency, concreteness, and length. An unrelated version of each of the lists was created by randomly pairing the targets with unrelated cues from another list. Thus, six lists of 30 pairs each (15 related, 15 unrelated) were created and counterbalanced so

that target words were paired with a related or unrelated cue and appear with each of the three judgment conditions. Twelve of the 16 buffer pairs from Experiment 1 were used.

Procedure

In Experiment 2, participants completed three study-test blocks. In each block, they studied 34 pairs (4 buffers, 30 main pairs) for 10s each with a 250ms ISI. Following the study phase, they solved addition problems for 3 minutes and then took a cued recall test (*Buzz – ?*) for each of the 30 main pairs.

For one of the study blocks, participants did not provide any judgments and instead viewed each word pair for 10s. For a second study block, participants were shown each pair one-at-a-time for 6s and then a percent JOL prompt appeared during the final 4s with the pair still on the screen. This 0-100% JOL used the same wording as the prompt from Experiment 1. For the third study block, participants were again shown each pair for 6s and had a JOL prompt appear with the pair for the last 4s. However, this JOL prompt used a binary scale and asked participants "Will you remember this pair on a later test?", and participants selected either "Yes" or "No". The order in which the three judgment conditions were completed was randomized for each participant. Scoring and analysis procedures were the same as Experiment 1.

Results

Judgments

To determine whether participants made binary judgments in less time than percent judgments, a paired-samples *t*-test was conducted to compare average response time for binary versus percent JOLs. The timing of participants' first click when entering their judgments was used as the response time measure. This test verified that participants took less time to provide binary JOLs (M = 1.24s, SE = .03) than percent JOLs (M = 1.87s, SE = .03), $t(121) = 15.73$, $p <$

.001, $d = 1.42$, $BF_{10} = 6.40 \times 10^{27}$. Thus, binary JOLs may require less cognitive effort to make a decision than percent JOLs.

JOL vs No JOL

To consider reactivity effects on recall, a 3 (judgment: binary JOL, percent JOL, no JOL) x 2 (type of pair: related, unrelated) repeated-measures ANOVA was used to compare cued recall performance (see Figure 5). Overall, participants recalled a higher percentage of related pairs correctly ($M = 84.59\%$, $SE = 1.15$) than unrelated pairs ($M = 43.12\%$, $SE = 2.09$), $F(1,121) = 792.18$, $p < .001$, $\eta^2_p = .87$, $BF_{10} = 4.67 \times 10^{51}$. The main effect of judgment was not significant, $F(2,242) = 2.22$, $p = .11$, $\eta^2_p = .02$, $BF_{01} = 5.23$, but there was a significant interaction, $F(2,242) = 13.78$, $p < .001$, $\eta^2_p = .10$, $BF_{10} = 4.29 \times 10^4$.

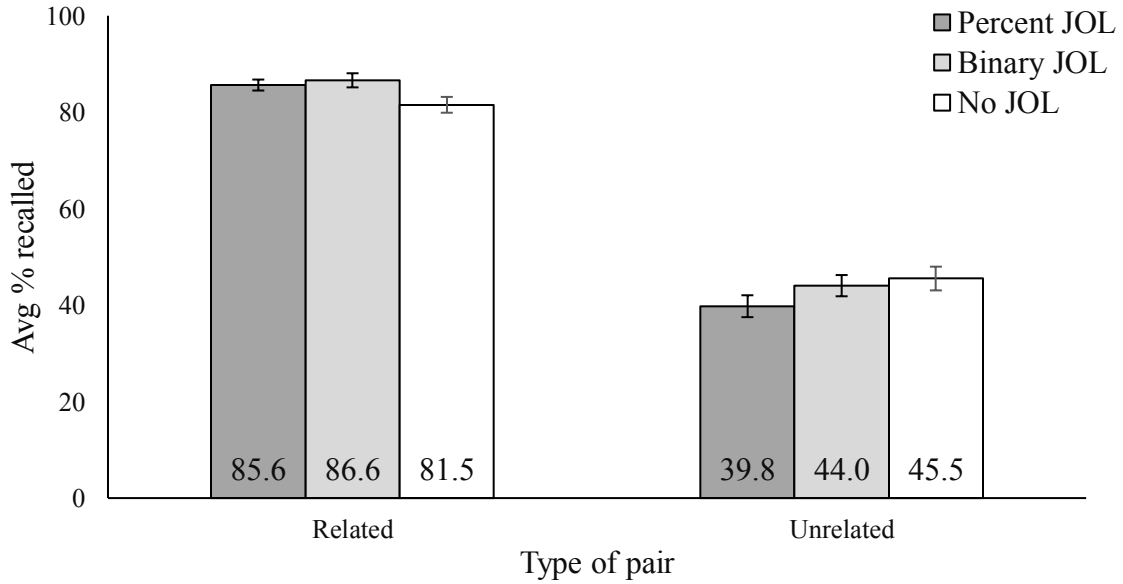

Figure 5. Average percent recalled correctly on the final cued recall tests based on type of pair and which judgment participants made during study. Error bars represent 1 standard error of the mean.

Based on prior findings (e.g., Myers et al., 2020; Soderstrom et al., 2015), I expected that binary and percent JOLs would lead to improved cued recall of related pairs but not impact recall of unrelated pairs compared to the no JOL condition. The present experiment supported these

predictions for related pairs, but not for unrelated pairs. Both percent and binary JOLs improved memory of related pairs compared to not making JOLs, percent JOLs: $t(121) = 2.64$, $p = .01$, $d = 0.24$, $BF_{10} = 2.78$, binary JOLs: $t(121) = 3.59$, $p < .001$, $d = 0.33$, $BF_{10} = 40.80$. In contrast to predictions, though, percent JOLs were associated with *reduced* memory for unrelated words compared to not making JOLs, $t(121) = 3.46$, $p < .001$, $d = 0.31$, $BF_{10} = 26.93$. Binary JOLs did not have an impact on recall of unrelated pairs, with the Bayes factor indicating that the null hypothesis (i.e., no difference) was 7 times more likely than the alternative, $t(121) = 0.82$, $p = .42$, $d = 0.07$, $BF_{01} = 7.19$.

Percent vs Binary JOLs

To explore further how type of judgment impacted reactivity effects, reactivity size was calculated for each participant by subtracting their performance for each no JOL condition from their performance on the JOL conditions (see Figure 6). A 2 (judgment: binary, percent) x 2 (type of pair: related, unrelated) repeated-measures ANOVA indicated a main effect of judgment, such that reactivity sizes were more positive when participants made binary JOLs ($M = 1.80$, $SE = 1.37$) compared to percent JOLs ($M = -0.82$, $SE = 1.26$), $F(1,121) = 4.88$, $p = .03$, $\eta^2_p = .04$, $BF_{10} = 1.13$. Of note, the Bayes factor indicated that the null and alternative hypothesis were equally likely under the data. The main effect of pair was also significant, $F(1,121) = 26.31$, $p < .001$, $\eta^2_p = .18$, $BF_{10} = 1.43 \times 10^4$. Collapsed across type of judgment, JOLs improved memory for related pairs ($M = 4.59$, $SE = 1.31$) but decreased memory for unrelated pairs ($M = -3.61$, $SE = 1.52$). The judgment x pair interaction was not significant, $F(1,121) = 2.77$, $p = .10$, $\eta^2_p = .02$, $BF_{01} = 1.84$, with the Bayes factor favoring the null hypothesis.

Follow-up paired-samples *t*-tests were still conducted to gain a better purchase on these data. Paired-samples *t*-tests indicated that type of JOL did not impact reactivity size for related

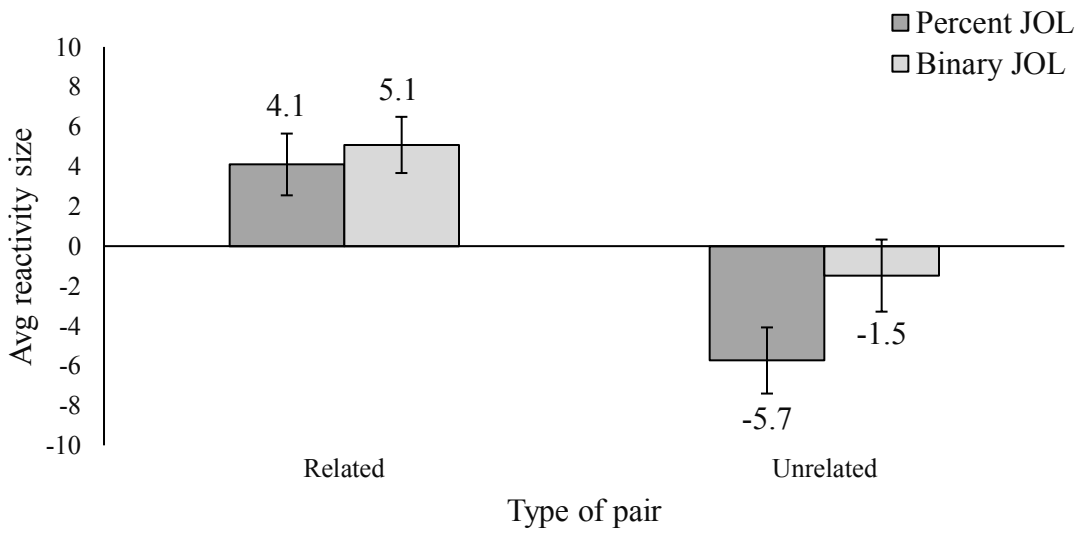

Figure 6. Average reactivity size (i.e., difference in percent correct between JOL and no JOL cued recall performance) based on type of pair and which judgment participants made during study. Error bars represent 1 standard error of the mean.

pairs, $t(121) = 0.70$, $p = .49$, $d = 0.06$, $BF_{01} = 7.84$, but reactivity size did differ moderately by type of JOL for unrelated pairs, $t(121) = 2.56$, $p = .01$, $d = 0.23$, $BF_{10} = 2.27$. Specifically, percent JOLs created a deficit in memory with a small effect size for unrelated pairs whereas binary JOLs did not.

Discussion

Experiment 2 found that both percentage and binary JOLs improved memory for related pairs. However, in contrast to some prior work (e.g., Myers et al., 2020; Soderstrom et al., 2015), percent JOLs harmed memory for unrelated pairs. This pattern of results aligns with some other reactivity experiments (Halamish & Undorf, 2022; Mitchum et al., 2016) that detected negative JOL reactivity on unrelated pairs when using percentage JOLs. I discuss this finding further in the General Discussion.

The main question of Experiment 2 was whether the reactivity effects on related and unrelated pairs changed with the type of judgment participants made – either a percentage or

binary (yes/no) judgment. Results indicated that binary JOLs had a more overall positive effect on memory than percent JOLs. However, this was driven by binary JOLs being significantly less detrimental to unrelated pairs compared to percent JOLs. This might suggest that more cognitive effort was required to make more nuanced judgments (i.e., percentage JOLs), leading to stronger impacts on later memory compared to less nuanced judgments (i.e., binary JOLs). I return to differences in reactivity sizes based on type of judgment in the General Discussion.

These results mirror findings from Mitchum et al.'s (2016) Experiment 3. Specifically, they found that, compared to percent JOLs, binary JOLs had stronger positive effects on memory for related pairs and weaker negative effects on memory for unrelated pairs. It is also important to note that results aligned between the present experiment, which used experimenter-paced study times, and Mitchum et al.'s (2016) experiment, which used self-paced study.

EXPERIMENT 3

In Experiment 3, the judgment prompt was again manipulated to determine whether the way JOLs are made changes their effects on later memory. Participants did not make judgments while studying one list of word pairs and made 0-100% JOLs while studying a different list. For a third study list, they also made a percent JOL, but they were also prompted to select a reason for making that JOL (e.g., relationship between the words, personal connection, not enough rehearsal). Requiring participants to explain why they made their JOLs should require more overt consideration than just choosing a number. Thus, if the specific act of making JOLs drives reactivity, then requiring a more in-depth process for making JOLs should lead to larger reactivity. However, if JOL reactivity is not specific to how participants make these JOLs, then requiring an explanation should not change the size of reactivity.

Method

Participants

One hundred fourteen (114) new participants were recruited via the CSU participant pool, completed the study in-person, and received one course credit. Of these, 13 participants were removed because they did not complete all blocks of the experiment (n = 5), did not provide at least 70% of JOLs (n = 5) during one of the lists, did not respond to at least 50% of items on one of the tests (n = 2), or due to a technical issue (n = 1), leaving a final sample size of 101 (89% of participants provided useable data). Age ranged from 17 to 24 (M = 19.0, SD = 1.2); 69 participants identified as female, 29 as male, 2 as non-binary, and 3 did not identify their gender. One participant did not provide their age.

Design

This study used a 3 (judgment: explain JOL, percent JOL, no JOL) x 2 (type of pair: related, unrelated) repeated-measures design. Type of pair was manipulated within each block while judgment was manipulated between blocks.

Materials and Procedure

The same ninety word pairs from Experiment 2 were used in Experiment 3. The procedure was nearly identical to that used in Experiment 2. Participants completed three study-test blocks whereby they studied 34 word pairs, solved addition problems for 3 minutes, and took a cued recall test over the 30 key word pairs.

Overall study time for each pair was increased to 14 seconds to provide sufficient time for participants to explain their JOLs. For the no JOL condition, participants simply studied each word pair for the full 14s. For the percent JOL condition, participants saw each word pair for 14s. The 0-100% JOL prompt used in the other experiments appeared after 6s and stayed on the screen for the remaining 8s. For explaining their JOLs, participants saw each word pair alone for 6s, then had 4s to type in their 0-100% JOL, and then had an additional 4s to select a reason for their JOL from a list of options.

A pilot experiment asked participants (n = 22) to study and make JOLs for word pairs, and then explain why they made that JOL. Participants' responses were most often based on one of the following: 1) relatedness between the two words in a pair, 2) a personal connection they had to one or both of the words, 3) a narrative (short story/phrase) they made to relate the two words together, or 4) guessing or simply indicating they would not remember it. Responses also seemed to differ based on whether participants provided a high or low JOL (cf., Jersakova et al., 2017). For example, if a participant thought they were likely to remember a pair, their reason

would be that the words were highly related or they had a strong connection to the words. If they thought they were not likely to remember a pair, they often indicated that there was no relation or they simply did not think they would remember the pair.

When asked to explain their reasoning, participants sometimes took close to 20 seconds to type in their responses. To equate study times between conditions in Experiment 3, I opted to show participants a list of options and asked them to select which best explained why they chose their JOL. They were allowed to select as many reasons for each JOL as they wished (see Appendix B for the options participants selected in Experiment 3). These were also separated based on whether they provided a high JOL (50% or higher) or low JOL (less than 50%) for each pair – when participants entered a JOL 50% or higher, they saw the list of options for high JOLs; when it was lower than 50%, they saw a different list of options. Figure 7 displays the options participants were provided.

If JOL 50% or higher		If JOL lower than 50%	
a)	Strongly related	a)	Only somewhat related
b)	Somewhat related	b)	Not related
c)	Rehearsed pair	c)	Not enough rehearsal
d)	Personal connection	d)	No personal connection
e)	Narrative	e)	No narrative
f)	Other	f)	Other

Figure 7. Options participants were given for why they made a JOL of 50% or higher or why they made a JOL of less than 50%.

There was a possibility that providing this list of reasons to participants would cause a shift in learning strategies. For example, seeing the option of "personal connection" may make participants realize they could make up a personal connection to themselves for each pair. To control for this possibility, these reasons were also displayed in the instructions for the no JOL

and percent JOL conditions. Specifically, participants were told "you could learn these pairs by thinking about the relationship between the two words, forming a personal connection, making up a narrative, repeating it to yourself, or using some other strategy".

Results

Judgments

Participants' recorded first click was again used as a measure of response time. Total time for the explain JOLs condition was calculated by adding the time for making the percent JOL and time for selecting a reason. This test verified that participants took more time overall to make and explain JOLs (M = 3.56s, SE = .05) than to simply make percent JOLs (M = 2.71s, SE = .07), although they were allotted 8s for both types of judgments, $t(100) = 11.86$, $p < .001$, d = 1.18, $BF_{10} = 1.93 \times 10^{16}$. Thus, being required to explain their JOLs may have required more cognitive effort than simply providing a percent JOL.

JOL vs No JOL

A 3 (judgment: explain JOL, percent JOL, no JOL) x 2 (type of pair: related, unrelated) repeated-measures ANOVA was used to compare cued recall performance (see Figure 8). Overall, participants correctly recalled a higher percentage of related pairs (M = 84.69%, SE = 1.69) than unrelated pairs (M = 41.34%, SE = 1.69), $F(1,100) = 788.16$, $p < .001$, $\eta^2_p = .89$, $BF_{10} = 9.83 \times 10^{45}$. The main effect of judgment was significant[8], although the Bayes factor indicated that the null and alternative hypotheses were equally likely, $F(2,200) = 3.88$, $p = .02$, $\eta^2_p = .04$, $BF_{10} = 1.10$. On average, the highest recall was observed after explaining JOLs (M = 65.31, SE = 1.73), followed by not making JOLs (M = 62.28, SE = 1.73) and making percent JOLs (M =

[8]Mauchly's test of sphericity was significant ($p < .05$), suggesting the assumption of sphericity was violated in this ANOVA. The Greenhouse-Geisser correction only slightly changed the p-value of this test, $F(1.7,168.7) = 3.88$, p = .03, $\eta^2_p = .04$.

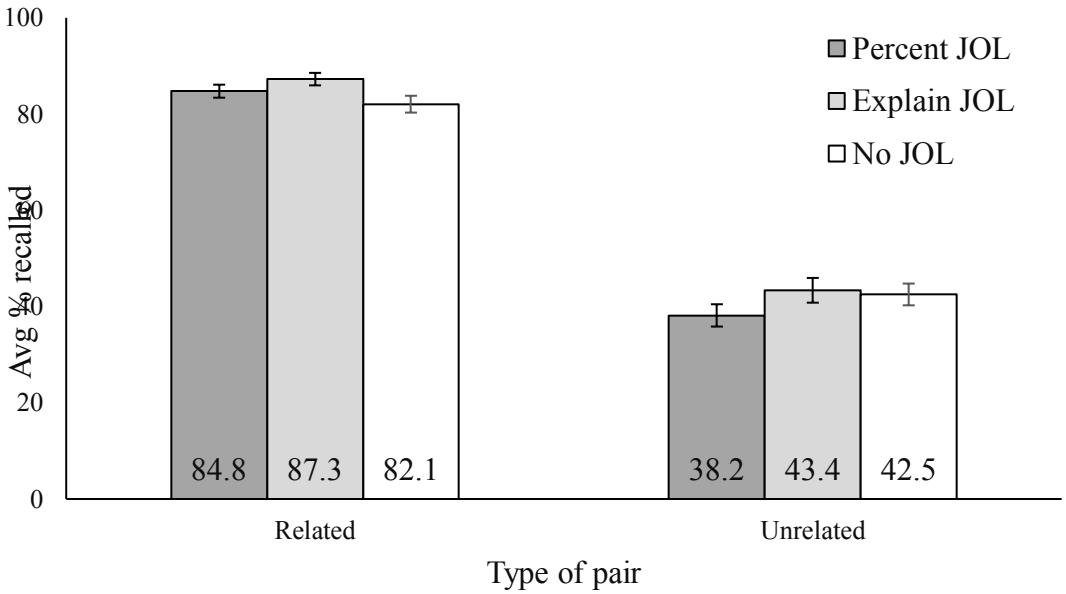

Figure 8. Average percent recalled correctly on the final cued recall tests based on type of pair and which judgment participants made during study. Error bars represent 1 standard error of the mean.

61.45, *SE* = 1.73). This was qualified by a significant judgment x type of pair interaction, $F(2,200) = 5.72$, $p = .004$, $\eta^2_p = .05$, $BF_{10} = 4.88$.

Explaining JOLs improved memory of related pairs compared to not making JOLs, $t(100) = 3.06$, $p = .003$, $d = 0.30$, $BF_{10} = 8.56$. However, providing percent JOLs did not significantly improve memory for related pairs (contrary to past experiments), with the Bayes factor indicating the null was about 2.5 times more likely than the alternative, $t(100) = 1.63$, $p = .11$, $d = 0.16$, $BF_{01} = 2.55$. Percent JOLs again harmed memory for unrelated pairs compared to not making JOLs, $t(100) = 2.14$, $p = .04$, $d = 0.21$, $BF_{10} = 8.56$, while explaining JOLs did not impact recall of unrelated pairs, $t(100) = 0.38$, $p = .70$, $d = 0.04$, $BF_{01} = 8.45$.

Percent vs Explain JOLs

Reactivity size was again calculated for each participant by subtracting their performance for each no JOL condition from their performance on the JOL condition (see Figure 9). A 2 (judgment: explain, percent) x 2 (type of pair: related, unrelated) repeated-measures ANOVA

indicated a main effect of judgment, such that reactivity sizes were more positive when participants explained their JOLs (M = 3.04, SE = 1.69) than when only providing percent JOLs (M = -0.83, SE = 1.69), $F(1,100) = 11.69$, $p < .001$, $\eta^2_p = .11$, $BF_{10} = 17.27$. The main effect of pair was also significant, $F(1,100) = 8.98$, $p = .003$, $\eta^2_p = .08$, $BF_{10} = 10.66$. JOLs improved memory for related pairs (M = 3.96, SE = 1.54) but decreased memory for unrelated pairs (M = -1.75, SE = 1.98). Although the judgment x pair interaction was not significant [$F(1,100) = 1.81$, $p = .18$, $\eta^2_p = .02$, $BF_{01} = 2.82$], follow-up paired-samples t-tests were still conducted to gain better purchase on these data. Type of JOL did not significantly impact reactivity size for related pairs, with the Bayes factor supporting the null, $t(100) = 1.80$, $p = .07$, $d = 0.18$, $BF_{01} = 1.90$. However, reactivity size did differ by type of JOL for unrelated pairs, $t(100) = 3.21$, $p = .002$, $d = 0.32$, $BF_{10} = 13.16$, whereby percent JOLs significantly harmed memory for unrelated pairs while explaining JOLs did not.

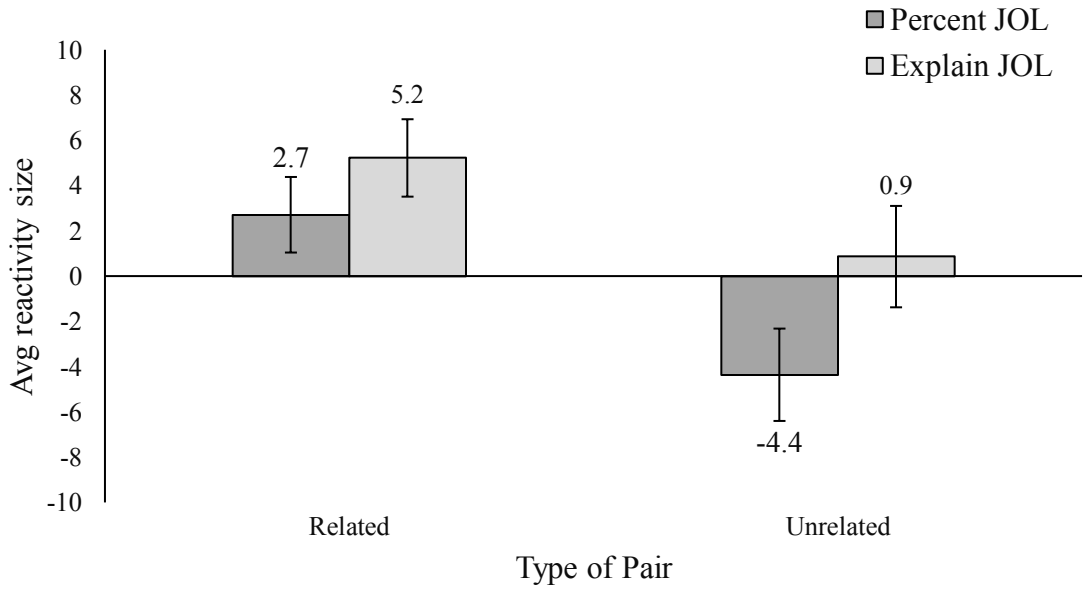

Figure 9. Average reactivity size (i.e., difference in percent correct between JOL and no JOL cued recall performance) based on type of pair and which judgment participants made during study. Error bars represent 1 standard error of the mean.

Discussion

Experiment 3 found that explaining JOLs improved memory for related pairs while percentage JOLs had minimal effects on related pairs. However, percent JOLs once again reduced memory of unrelated pairs (although this detriment was not detected when participants were required to explain their JOLs). Experiment 3, like Experiment 2, was characterized by changes in reactivity size based on type of judgment whereby selecting explanations to JOLs led to more positive effects on later memory than only making a percent JOL. However, this effect was again mainly driven by a difference of effect on unrelated pairs. These findings will be discussed further in the General Discussion.

GENERAL DISCUSSION

Main Findings

The purpose of the present experiments was to explore whether changing the method in which a JOL was solicited might change its reactive effects on later memory. Specifically, I proposed that JOLs that require or allow for more cognitive effort and deliberation while making the judgment might lead to stronger reactivity effects on a later cued recall test if reactivity effects reflected processes occurring during that judgment. Experiment 1 changed the time allotted to participants to make JOLs, either giving them 2s or 4s. Experiment 2 compared binary (yes/no) JOLs to 0-100% JOLs, and Experiment 3 required participants to sometimes choose why they made the JOL in addition to providing a percentage judgment. Across experiments, I expected that allowing 4s to make JOLs (Experiment 1), percentage JOLs, (rather than binary; Experiment 2), and requiring participants to select a reason for their JOL (Experiment 3) would lead to more cognitive processing while deciding on what JOL to make because each required or allowed for more fine-grained distinctions in JOLs. If reactivity effects were not sensitive to these changes in the JOL prompt, it might suggest that JOL reactivity is driven more by the requirement to make JOLs rather than the amount of cognitive processing that occurs while making the judgment.

In each experiment, I found that participants took longer to make the more cognitively demanding JOLs (i.e., 4s JOLs, percentage JOLs, explaining JOLs, respectively). Although not surprising, this would suggest that manipulating the JOL prompt caused some changes in participants' deliberation while making the JOLs. However, these differences in time to make judgments (and judgment complexity) did not lead to the predicted differences in reactivity

effects. In Experiment 1, the magnitude of JOL reactivity was similar regardless of whether participants were given 2s or 4s to make their JOLs. Findings in Experiment 2 and 3 suggested some differences in reactivity sizes based on JOL prompt, but not consistently in the predicted directions. In Experiment 2, percent JOLs led to stronger decrements in memory for unrelated pairs than did binary JOLs. This supports the proposed theory whereby the more effortful JOL (percent JOL) led to stronger reactivity effects. However, in Experiment 3, percent JOLs also more negatively impacted memory for unrelated pairs, whereas providing an explanation with the JOL did not impact memory of unrelated pairs. In other words, the more cognitive demanding judgment (explaining JOLs) in Experiment 3 led to weaker JOL reactivity effects than the less demanding judgment.

A main effect was found in Experiments 2 and 3 whereby, collapsed across pair type, binary JOLs and explaining JOLs more positively impacted memory than percentage JOLs. Mitchum et al. (2016) also found a similar pattern when they compared binary and percentage JOLs. This might suggest that different JOL prompts impact reactivity effects. Nevertheless, these results still do not support the hypothesis that JOL prompts that require more cognitive deliberation would lead to stronger reactivity effects. It is possible that the different JOL prompts lead participants to approach the study task differently. For example, perhaps having to select a reason for one's JOL increased attention toward the strategies the learner used in Experiment 3. However, this still would suggest a global shift in studying rather than reactivity effects being driven by cognitive processes during the JOL decision.

In all, the degree of reactivity sometimes differed based on the JOL prompt (more strongly for unrelated pairs), but patterns across experiments did not comport with the hypothesis whereby JOLs that prompted more cognitive deliberation would lead to stronger reactive effects.

Collectively, such data suggest that JOLs may introduce a global shift in processing that is insensitive to the specific format of the JOL. That is, based on the current study, the presence of the JOLs during study seems to drive reactivity effects more than specific cognitive processing that occurs at the time of making each judgment.

Integration with Past Findings

In the present experiments, I proposed that JOL reactivity would be isolated to benefitting memory for related pairs but have no impact on unrelated pairs, reflecting past studies (e.g., Myers et al., 2020; Soderstrom et al., 2015). However, in all three experiments, percentage JOLs (which have been most commonly used in prior studies) tended to harm memory of unrelated pairs in addition to strengthening memory of related pairs. Although not hypothesized, these findings are consistent with some other reactivity studies (e.g., Halamish & Undorf, 2022; Mitchum et al., 2016). The effects of JOLs on memory for unrelated word pairs continues to vary across ever-emerging reactivity studies. Some studies find negative reactivity (e.g., Mitchum et al., 2016), no reactivity (e.g., Soderstrom et al., 2015), different reactivity effects across experiments (e.g., Halamish & Undorf, 2022), or even positive reactivity (Myers et al., 2020; Rivers et al., 2021). The fluctuating impact of JOLs on memory for unrelated pairs could reflect a number of possibilities, including differences in association strength of word pairs, study conditions, difficulty of experimental task, and general variability associated with a typically small effect (effect sizes for unrelated pairs in the present experiments ranged from $d = 0.18$-0.31).

A common conclusion from JOL reactivity work is that reactivity effects greatly depend on the learning situation. These mixed findings might suggest that JOL reactivity effects cannot be explained by any one mechanism. Instead, different mechanisms may drive reactivity effects

in different learning situations. JOLs could impact approaches to learning (Mitchum et al., 2016), additional processing of certain cues (Koriat, 1997; Soderstrom et al., 2015), and attentional resources either positively by reducing mind-wandering and making items more distinctive (Murphy et al., 2023; Shi et al., 2022) or negatively by introducing a dual-attention task (Craik et al., 1996; cf. Mitchum et al., 2016). All of these possible effects of JOLs (as well as others) could interplay during the learning process, with some effects overshadowing others depending on the conditions. In the present experiment, perhaps binary JOLs more positively impacted memory than percentage JOLs because they required less cognitive effort (leaving more cognitive resources available for learning). In contrast, perhaps explaining JOLs led to more positive memorial benefits because it brought conscious attention to learning strategies. In other studies, reactivity effects for related and unrelated pairs after studying mixed lists might vary due to the difficulty of the task and test. For example, JOLs might beget more positive benefits for related but not unrelated pairs when the learning task is generally easier. In contrast, when a learning task or test is generally harder, JOLs might be too cognitively demanding to boost memory and thus lead to more negative impacts (cf., Mitchum et al., 2016). In all, studies must continue to be conducted so that the nuances of JOL effects on learning and memory can continue to be disentangled.

Another possibility for divergences in the effect of JOLs on memory across studies (and within the present study) could be due to individual differences between participants. Participants might differ in their overall cognitive effort toward the task or they might approach the learning situation differently. Moreover, other individual cognitive differences such as participants' general understanding of numbers (De Bruin et al., 2000), executive functioning (Komori, 2016), and their own metacognition (De Bruin et al., 2017; Kröner & Biermann, 2007;

Kruger & Dunning, 1999) might explain why differences continue to be found across studies that use very similar designs. A focus on individual differences in how learners use JOLs might shed more light into the different mechanisms driving JOL reactivity.

Theoretical Implications

The trends in the present experiments toward positive JOL reactivity for related pairs and negative JOL reactivity for unrelated pairs most closely align with the predictions of the changed goal account (Mitchum et al., 2016). To review, the changed goal account proposes that JOLs draw attention to some materials (e.g., related pairs) being easier to learn than other (e.g., unrelated pairs). Due to this increased focus on learning difficulty, JOLs cause learners to shift their efforts toward learning the easier material and spend less effort learning harder materials, resulting in positive reactivity for related pairs and negative reactivity for unrelated pairs. This pattern was also found in the present experiments. In addition, the changed goal account suggests an overall global shift in the approach to learning rather than JOLs impacting processing at the time of making each JOL. This is also corroborated by my findings, whereby more cognitively demanding JOLs did not result in stronger reactivity effects, suggesting that the mere presence of JOLs might cause a global shift in learning approach.

Regarding the other two theories, my findings provide less insight. The present results did not align with the cue-strengthening hypothesis (Soderstrom et al., 2015), which predicts that JOLs should not impact unrelated pairs. However, predictions about how different JOL prompts should impact reactivity effects are not directly addressed in the cue-strengthening hypothesis. Given that Soderstrom et al. (2015) proposed that JOLs increase processing of the cues used to make JOLs, this theory might suggest that reactivity effects should be impacted by how the question is asked if the prompt itself encourages additional processing of those cues. However, it

is also possible that JOLs encourage attention toward relatedness as a cue in general, leading to later reactivity effects, rather than participants' depth of processing of each relationship between the cue and target while making each judgment. Senkova and Otani (2021) have suggested that JOL reactivity occurs due to JOLs increasing item-specific processing but reducing relational processing. It is difficult to consider how the present findings fit with this theory, given that item-specific and relational processing were not isolated from one another in the present experiments.

To reiterate, theories of JOL reactivity are still being developed and tested (e.g., Maxwell & Huff, 2022; Rivers et al., 2021; Shi et al., 2022; Tekin & Roediger, 2020), so theories may continue to become more specified over time. The present experiments provide new information that can inform those theories by also examining how the type of JOL prompt impacts reactivity effects.

Limitations

One limitation of the present experiments was attrition, particularly in Experiment 1. Although all participants were recruited from the CSU subject pool, participants who completed the study online were more likely to not complete the experiment or not provide enough responses to the study questions than in-person participants, leading to an overall attrition rate of 38% in Experiment 1 (attrition rates in Experiments 2 and 3, with all in-person data collection, fell around 10%). Analyses suggest that findings did not differ between online and in-person participants (see Appendix C) in Experiment 1. Nevertheless, given marked disparities in the size of the samples (online: 67 participants, in-person: 17), a failure to detect differences may also reflect a lack of power. A future experiment should compare full online vs in-person samples to delve into whether differences in JOL reactivity effects occur.

Another concern with high attrition rates is the possibility of non-random attrition between conditions. Zhou and Fishbach (2016) note general caution concerning studies using online participants with high attrition rates, arguing that attrition can lead to non-random assignment if some experimental conditions are more difficult than others. As one example, Zhou and Fishbach (2016) compared attrition between participants assigned to writing 100 words without using the letters A and N and participants assigned to writing 100 words without using the letters X and Y. Those assigned to the more difficult task (writing without using A or N) were more likely to leave the experiment early and thus be removed from the data.

Differences in attrition among different conditions should not be a concern in the present study given the repeated-measures design: All of a participant's data were removed if a participant did not meet the requirements of any study block. However, there remains the possibility that the participants who did complete the experiment were not representative of typical college students. This could explain differences in reactivity effects across experiments (e.g., Experiments 2 and 3 found that JOLs reduced memory of unrelated pairs, but this effect was not significant in Experiment 1). Overall recall performance was consistent across all 3 experiments – around 85% of related pairs and 35% of unrelated pairs, on average, were recalled correctly in all three experiments. This suggests that overall attention or motivation to complete the task successfully was most likely similar across all three experimental samples, although there is still a possibility that other systematic differences occurred between samples. More attention to demographic information and individual difference measures would have been beneficial to explore in the present study.

Additionally, most JOL reactivity research has focused on a design of studying related and unrelated word pairs or individual word lists (but see Ariel et al., 2021; Schäfer & Undorf,

2023; Shi et al., 2017). A broader exploration of study materials and testing methods in future research could permit better purchase regarding the mechanisms underlying the effects of JOLs on cognition. Nevertheless, a focus on JOL reactivity experiments is a recent development, and initial findings and theories of JOL reactivity are still being refined. As the field continues to gain traction and understanding of the underlying cognitive mechanisms, a more generalized focus on JOL reactivity in different learning situations can be explored.

Practical Implications

JOLs have been used as a measure for over 50 years, beginning with Arbuckle and Cuddy (1969). With an increased focus on reactivity beginning recently (Soderstrom et al., 2015), the field has so far discovered that reactivity is not a unitary effect that occurs in every learning situation. This makes it difficult to predict how JOL reactivity might have impacted past metamemory findings or how JOLs will impact learning in real-world situations (see Ariel et al., 2021; Schäfer & Undorf, 2023). Thus, a full understanding of why and how JOL reactivity occurs is essential to both determine how reactivity has impacted past studies and to make future advances into JOLs as a learning strategy. For example, the cue-overlap theory predicts that JOLs strengthen memory for the cues that learners use to inform their JOLs. If learners can use those cues to access information on a later test, beneficial JOL reactivity will occur. To extrapolate this to real-world materials, we must gain a thorough understanding of what cues learners use when assessing their learning of complex materials and how those cues can impact later tests. If researchers can find cues that learners consistently consider for assessing complex materials and those cues are beneficial for a later test, this could be a potential application to how making JOLs can directly benefit student learning. Nevertheless, substantial work will be necessary to parse the information learners use when evaluating their learning in the classroom,

and how that information can be used in different evaluation methods. Additionally, given the lack of a unifying theoretical perspective, it is unlikely that only the effects predicted by any one theory can be isolated to extrapolate to effects on real-world learning materials.

Yet, monitoring one's own understanding while learning often leads to better study decisions and better subsequent performance when learning is self-regulated (Griffin et al., 2013; Nelson et al., 1994). Accordingly, telling students to assess their own metacognition while learning is not bad advice, even if the research cannot yet speak to the specific benefits of making JOLs on learning real-world materials.

Conclusions

Research focused on JOL reactivity has multiplied since Soderstrom and colleagues' (2015) seminal study. Subsequent studies have begun to explore reactivity effects with different study materials (e.g., Ariel et al., 2021; Double et al., 2018; Maxwell & Huff, 2022; Schäfer & Undorf, 2023; Shi et al., 2022), types of tests (Myers et al., 2020; Zhao et al., 2023), and experimental designs (e.g., Janes et al., 2018; Rivers et al., 2021). The present experiments examined the methods of JOL solicitation and indicated that the cognitive effort required to make JOLs is not consistently linked to stronger reactivity effects on later memory. Instead, JOL reactivity occurred to some degree across different formats of soliciting the judgment. These findings suggest that the presence of JOLs during study cause a more general shift in participants' approach to learning, subsequently leading to reactivity effects.

REFERENCES

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

Ariel, R., Dunlosky, J., & Bailey, H. (2009). Agenda-based regulation of study-time allocation: When agendas override item-based monitoring. *Journal of Experimental Psychology: General*, *138*(3), 432-447.

Ariel, R., Karpicke, J. D., Witherby, A. E., & Tauber, S. K. (2021). Do judgments of learning directly enhance learning of educational materials? *Educational Psychology Review*, *33*(2), 693-712.

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.

Begg, I. M., Martin, L. A., & Needham, D. R. (1992). Memory monitoring: How useful is self-knowledge about memory? *European Journal of Cognitive Psychology*, *4*(3), 195-218.

Carpenter, S. K., Wilford, M. M., Kornell, N., & Mullaney, K. M. (2013). Appearances can be deceiving: Instructor fluency increases perceptions of learning without increasing actual learning. *Psychonomic Bulletin & Review*, *20*(6), 1350-1356.

Craik, F. I., Govoni, R., Naveh-Benjamin, M., & Anderson, N. D. (1996). The effects of divided attention on encoding and retrieval processes in human memory. *Journal of Experimental Psychology: General*, *125*(2), 159-180.

De Bruin, W. B., & Carman, K. G. (2018). Measuring subjective probabilities: The effect of response mode on the use of focal responses, validity, and respondents' evaluations. *Risk Analysis*, *38*(10), 2128-2143.

De Bruin, W. B., Fischhoff, B., Millstein, S. G., & Halpern-Felsher, B. L. (2000). Verbal and numerical expressions of probability: "It's a fifty–fifty chance". *Organizational Behavior and Human Decision Processes*, *81*(1), 115-131.

De Bruin, A. B. H., Kok, E. M., Lobbestael, J., & de Grip, A. (2017). The impact of an online tool for monitoring and regulating learning at university: Overconfidence, learning strategy, and personality. *Metacognition and Learning*, *12*, 21-43.

Double, K. S., Birney, D. P., & Walker, S. A. (2018). A meta-analysis and systematic review of reactivity to judgements of learning. *Memory, 26*(6), 741-750.

Dunlosky, J., & Tauber, S. U. K. (Eds.). (2016). *The Oxford Handbook of Metamemory*. Oxford University Press.

England, B. D., Ortegren, F. R., & Serra, M. J. (2017). Framing affects scale usage for judgments of learning, not confidence in memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(12), 1898-1908.

England, B. D., & Serra, M. J. (2012). The contributions of anchoring and past-test performance to the underconfidence-with-practice effect. *Psychonomic Bulletin & Review*, *19*(4), 715-722.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data (Rev. ed.)*. The MIT Press.

Finn, B. (2008). Framing effects on metacognitive monitoring and control. *Memory & Cognition*, *36*(4), 813-821.

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1977). Knowing with certainty: The appropriateness of extreme confidence. *Journal of Experimental Psychology: Human Perception and Performance*, *3*(4), 552-564.

Glenberg, A. M., & Epstein, W. (1985). Calibration of comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*(4), 702-718.

Griffin, T. D., Wiley, J., & Salas, C. R. (2013). Supporting effective self-regulated learning: The critical role of monitoring. *International Handbook of Metacognition and Learning Technologies*, 19-34.

Hacker, D. J., Bol, L., Horgan, D. D., & Rakow, E. A. (2000). Test prediction and performance in a classroom context. *Journal of Educational Psychology, 92(1),* 160-170.

Halamish, V., McGillivray, S., & Castel, A. D. (2011). Monitoring one's own forgetting in younger and older adults. *Psychology and Aging, 26*(3), 631–635.

Halamish, V., & Undorf, M. (2022). Why do judgments of learning modify memory? Evidence from identical pairs and relatedness judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Hanczakowski, M., Zawadzka, K., Pasek, T., & Higham, P. A. (2013). Calibration of metacognitive judgments: Insights from the underconfidence-with-practice effect. *Journal of Memory and Language*, *69*(3), 429-444.

Higham, P. A., Zawadzka, K., & Hanczakowski, M. (2016). Internal mapping and its impact on measures of absolute and relative metacognitive accuracy. In J. Dunlosky & S. Tauber (Eds.), *The Oxford Handbook of Metamemory*, 39-61. Oxford University Press.

Hunt, R. R. (2006). The concept of distinctiveness in memory research. In R. Hunt & J. Worthen (Eds.), *Distinctiveness and Memory*, 3-25.

Hunt, R. R. (2012). Distinctive processing: The co-action of similarity and difference in memory. *Psychology of Learning and Motivation*, *56*, 1-46.

Janes, J. L., Rivers, M. L., & Dunlosky, J. (2018). The influence of making judgments of learning on memory performance: Positive, negative, or both?. *Psychonomic Bulletin & Review*, *25*(6), 2356-2364.

JASP Team (2022). JASP (Version 0.16.2) [Computer software].

Jersakova, R., Allen, R. J., Booth, J., Souchay, C., & O'Connor, A. R. (2017). Understanding metacognitive confidence: Insights from judgment-of-learning justifications. *Journal of Memory and Language*, *97*, 187-207.

Keleman, 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*(6), 1394-1409.

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

Komori, M. (2016). Effects of working memory capacity on metacognitive monitoring: A study of group differences using a listening span test. *Frontiers in Psychology*, *7*, 285.

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

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

Kornell, N., & Bjork, R. A. (2009). A stability bias in human memory: Overestimating remembering and underestimating learning. *Journal of Experimental Psychology: General, 138,* 449–468.

Kröner, S., & Biermann, A. (2007). The relationship between confidence and self-concept—Towards a model of response confidence. *Intelligence*, *35*(6), 580-590.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, *77*(6), 1121-1134.

Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, *142*(2), 573-603.

Kubik, V., Koslowski, K., Schubert, T., & Aslan, A. (2022). Metacognitive judgments can potentiate new learning: The role of covert retrieval. *Metacognition and Learning*, *17,* 1057-1077.

Maxwell, N. P., & Huff, M. J. (2022). Reactivity from judgments of learning is not only due to memory forecasting: Evidence from associative memory and frequency judgments. *Metacognition and Learning*, *17*(2), 589-625.

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

McGillivray, S., & Castel, A. D. (2017). Older and younger adults' strategic control of metacognitive monitoring: The role of consequences, task experience, and prior knowledge. *Experiment Aging Research, 43*(3), 233-256.

Metcalfe, J., & Kornell, N. (2005). A region of proximal learning model of study time allocation. *Journal of Memory and Language*, *52*(4), 463-477.

Mitchum, A. L., Kelley, C. M., & Fox, M. C. (2016). When asking the question changes the ultimate answer: Metamemory judgments change memory. *Journal of Experimental Psychology: General, 145*(2), 200-219.

MRC Psycholinguistic Database: Machine Usable Dictionary. Version 2.00

Mueller, M. L., Tauber, S. K., & Dunlosky, J. (2013). Contributions of beliefs and processing fluency to the effect of relatedness on judgments of learning. *Psychonomic Bulletin & Review, 20*(2), 378-384.

Murphy, D. H., Halamish, V., Rhodes, M. G., & Castel, A. D. (2023). How evaluating memorability can lead to unintended consequences. *Metacognition and Learning*, 1-29.

Myers, S. J., Rhodes, M. G., & Hausman, H. E. (2020). Judgments of learning (JOLs) selectively improve memory depending on the type of test. *Memory & Cognition*, *48*(5), 745-758.

Nelson, T. O. (1990). Metamemory: A theoretical framework and new findings. *Psychology of Learning and Motivation, 26,* 125-173.

Nelson, T. O., Dunlosky, J., Graf, A., & Narens, L. (1994). Utilization of metacognitive judgments in the allocation of study during multitrial learning. *Psychological Science*, *5*(4), 207-213.

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

Nelson, T. O., & Narens, L. (1994). Why investigate metacognition. In J. Metcalfe & A. Shimamura (Eds.), *Metacognition: Knowing About Knowing*, *13*, 1-25.

Qualtrics. (2020). *Qualtrics* (03-2022). Provo, UT. Available at: https://www.qualtrics.com

Rawson, K. A., Dunlosky, J., & McDonald, S. L. (2002). Influences of metamemory on performance predictions for text. *The Quarterly Journal of Experimental Psychology: Section A*, *55*(2), 505-524.

Rhodes, M. G. (2016). Judgments of learning: Methods, data, and theory. In Dunlosky, J. & Tauber, S. K. (Eds.), *The Oxford Handbook of Metamemory,* 65-80. Oxford University Press.

Rhodes, M. G. (2019). Metacognition. *Teaching of Psychology*, *46*(2), 168-175.

Rhodes, M. G., & Castel, A. D. (2008). Memory predictions are influenced by perceptual information: Evidence for metacognitive illusions. *Journal of Experimental Psychology: General*, *137*(4), 615-625.

Rhodes, M. G., & Castel, A. D. (2009). Metacognitive illusions for auditory information: Effects on monitoring and control. *Psychonomic Bulletin & Review, 16*(3), 550-554.

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

Rivers, M. L., Janes, J. L., & Dunlosky, J. (2021). Investigating memory reactivity with a within-participant manipulation of judgments of learning: Support for the cue-strengthening hypothesis. *Memory*, *29*(10), 1342-1353.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2), 225–237.

Saenz, G. D., & Smith, S. M. (2018). Testing judgments of learning in new contexts to reduce confidence. *Journal of Applied Research in Memory and Cognition*, *7*(4), 540-551.

Schäfer, F., & Undorf, M. (2023). On the educational relevance of immediate judgment of learning reactivity: No effects of predicting one's memory for general knowledge facts. *Journal of Applied Research in Memory and Cognition*.

Senkova, O., & Otani, H. (2021). Making judgments of learning enhances memory by inducing item-specific processing. *Memory & Cognition*, *49*(5), 955-967.

Serra, M. J., & DeMarree, K. G. (2016). Unskilled and unaware in the classroom: College students' desired grades predict their biased grade predictions. *Memory & Cognition*, *44*(7), 1127-1137.

Serra, M. J., & England, B. D. (2012). Magnitude and accuracy differences between judgements of remembering and forgetting. *The Quarterly Journal of Experimental Psychology*, *65*(11), 2231-2257.

Shi, A., Xu, C., Zhao, W., Shanks, D. R., Hu, X., Luo, L., & Yang, C. (2022). Judgments of learning reactively facilitate visual memory by enhancing learning engagement. *Psychonomic Bulletin & Review*, 1-12.

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

Soderstrom, N. C., & Rhodes, M. G. (2014). Metacognitive illusions can be reduced by monitoring recollection during study. *Journal of Cognitive Psychology, 26*(1), 118-126.

Son, L. K., & Kornell, N. (2008). Research on the allocation of study time: Key studies from 1890 to the present (and beyond). In J. Dunlosky & R. Bjork (Eds.), *A Handbook of Memory and Metamemory*, 333-351.

Spellman, B. A., & Bjork, R. A. (1992). When predictions create reality: Judgments of learning may alter what they are intended to assess. *Psychological Science*, *3*(5), 315-317.

Tauber, S. K. U., Dunlosky, J., & Rawson, K. A. (2015). The influence of retrieval practice versus delayed judgments of learning on memory: Resolving a memory-metamemory paradox. *Experimental Psychology*, *62*(4), 254-263.

Tauber, S. K., & Rhodes, M. G. (2012a). Multiple bases for young and older adults' judgments of learning in multitrial learning. *Psychology and Aging*, *27*(2), 474-483.

Tauber, S. K., & Rhodes, M. G. (2012b). Measuring memory monitoring with judgments of retention (JORs). *The Quarterly Journal of Experimental Psychology, 65*(7), 1376-1396.

Tekin, E., & Roediger, H. L. (2017). The range of confidence scales does not affect the relationship between confidence and accuracy in recognition memory. *Cognitive Research: Principles and Implications*, *2*, 1-13.

Tekin, E., & Roediger III, H. L. (2020). Reactivity of judgments of learning in a levels-of-processing paradigm. *Zeitschrift für Psychologie, 228*(4), 278-290.

Tekin, E., & Roediger, H. L. (2021). The effect of delayed judgments of learning on retention. *Metacognition and Learning, 16,* 407-429.

Undorf, M., & Bröder, A. (2020). Cue integration in metamemory judgements is strategic. *Quarterly Journal of Experimental Psychology*, *73*(4), 629-642.

Witherby, A. E., & Tauber, S. K. (2017). The influence of judgments of learning on long-term learning and short-term performance. *Journal of Applied Research in Memory and Cognition, 6*(4), 496-503.

Zawadzka, K., & Higham, P. A. (2015). Judgments of learning index relative confidence, not subjective probability. *Memory & Cognition*, *43*(8), 1168-1179.

Zhao, W., Li, J., Shanks, D. R., Li, B., Hu, X., Yang, C., & Luo, L. (2023). Metamemory judgments have dissociable reactivity effects on item and interitem relational memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Zhao, W., Li, B., Shanks, D. R., Zhao, W., Zheng, J., Hu, X., ... & Yang, C. (2021). When judging what you know changes what you really know: Soliciting metamemory judgments reactively enhances children's learning. *Child Development, 93,* 405-417.

Zhou, H., & Fishbach, A. (2016). The pitfall of experimenting on the web: How unattended selective attrition leads to surprising (yet false) research conclusions. *Journal of Personality and Social Psychology*, *111*(4), 493.

APPENDICES

Appendix A. JOL magnitudes

Experiment 1

A 2 (type of pair: related, unrelated) x 2 (time: 2s, 4s) repeated-measures ANOVA was conducted to compare differences in the magnitude of JOLs participants provided during study (see Figure A1). Note that the 2s corresponds to 8s of total study and the 4s time corresponds to 10s of total study. Overall, participants gave higher JOLs to related pairs (M = 70.07, SE = 1.82) compared to unrelated pairs (M = 34.47, SE = 1.82), $F(1,83) = 355.88$, $p < .001$, $\eta^2_p = .81$, $BF_{10} = 3.18 \times 10^{28}$. Participants' JOL magnitudes were similar when they were only given 2s to make their JOLs (M = 51.53, SE = 1.65) and when they were given 4s (M = 53.01, SE = 1.65), $F(1,83) = 1.68$, $p = .19$, $\eta^2_p = .02$, $BF_{01} = 2.56$. The pair x time interaction was not significant, $F(1,83) = 0.07$, $p = .79$, $\eta^2_p < .001$, $BF_{01} = 6.77$.

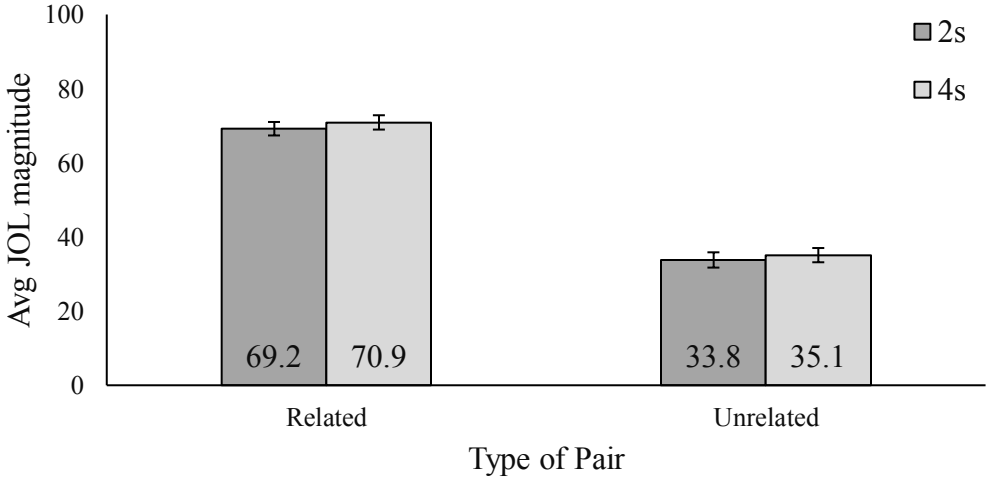

Figure A1. Average JOL magnitudes given during study for related and unrelated pairs based on whether participants were given 2s or 4s to make each JOL. Error bars represent 1 standard error of the mean.

Experiment 2

A 2 (type of pair: related, unrelated) x 2 (judgment: percent, binary) repeated-measures ANOVA was conducted on the average magnitude of participants' JOLs (see Figure A2). Average JOL magnitude for binary judgments was calculated based on the percent of pairs that participants responded "Yes, I will remember this" to. Percent JOLs and binary JOLs are still measured on different scales, but this method has been commonly used in past studies to evaluate binary JOL magnitudes (e.g., Hanczakowski et al., 2013; Jersakova et al., 2017; Zawadzka & Higham, 2015).

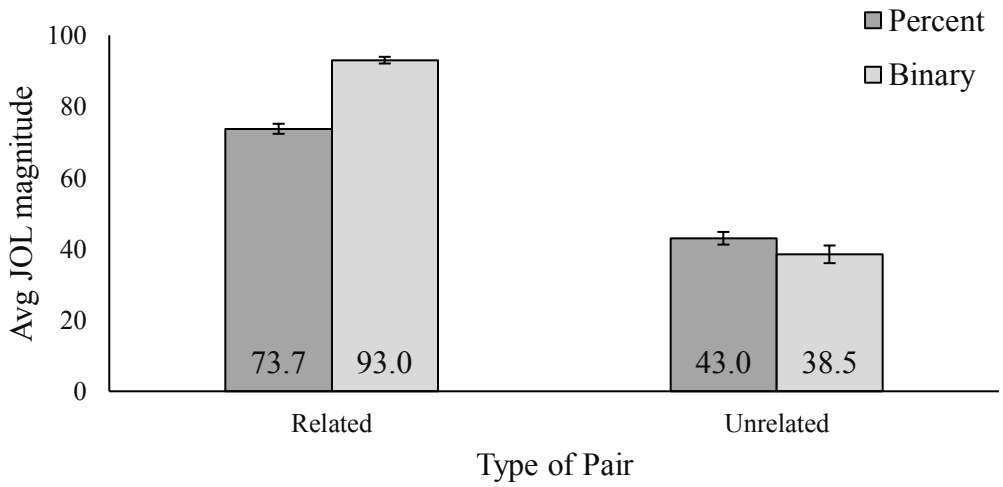

Figure A2. Average percent JOL magnitudes and percent of pairs participants responded "Yes" to (binary judgments) for related and unrelated pairs. Error bars represent 1 standard error of the mean.

Overall, participants gave higher JOLs to related pairs (M = 83.27, SE = 1.53) compared to unrelated pairs (M = 40.72, SE = 1.53), $F(1,121) = 669.24$, $p < .001$, $\eta^2_p = .85$, $BF_{10} = 4.37\text{x}10^{47}$. The main effect of participants' JOLs based on type of judgment was significant, $F(1,121) = 30.58$, $p < .001$, $\eta^2_p = .20$, $BF_{10} = 5.67\text{x}10^{3}$, as was the pair x judgment interaction, $F(1,121) = 138.21$, $p < .001$, $\eta^2_p = .53$, $BF_{10} = 3.98\text{x}10^{20}$.

Follow-up t-tests indicated that participants' average JOLs for related pairs were higher for binary judgments than percent judgments, $t(121) = 14.33$, $p < .001$, $d = 1.30$, $BF_{10} =$

$4.52x10^{24}$. To clarify, on average, participants believed they would remember about 92% of the related pairs for which they provided binary judgments. For percent JOLs, participants indicated they had, on average, a 73.7% likelihood of remembering the related pairs. For unrelated pairs, the pattern was reversed – the average percentage JOLs were significantly higher than average binary JOLs, although the Bayes factor indicated the alternative hypothesis was only slightly more probable than the null, $t(121) = 2.28$, $p = .02$, $d = 0.14$, $BF_{10} = 1.22$.

Experiment 3

A 2 (type of pair: related, unrelated) x 2 (judgment: percent, explain) repeated-measures ANOVA was conducted to compare the magnitude of JOLs participants provided during study (see Figure A3). The percent and explain judgments were both made on a 0-100% scale. The only difference was that the explain judgment condition required participants to select why they made their JOL.

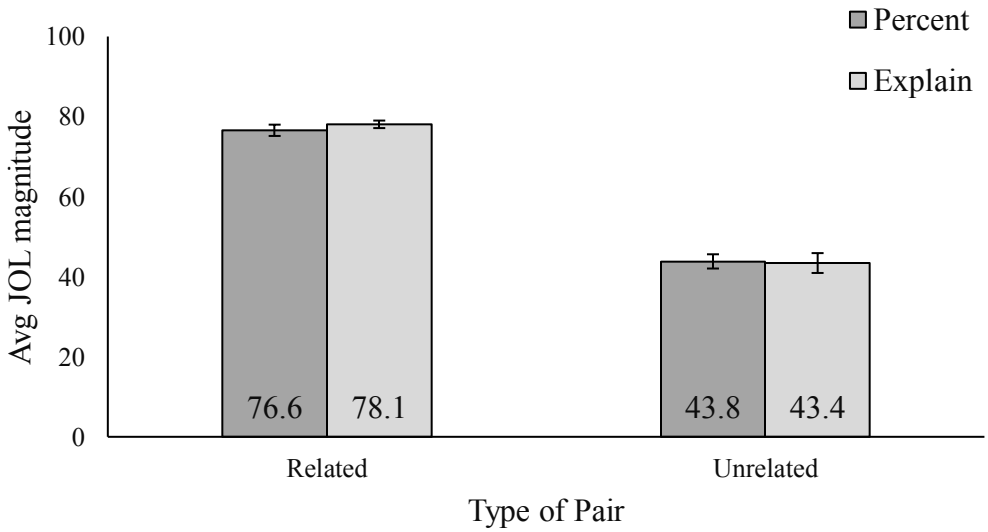

Figure A3. Average percent JOL magnitudes when participants completed the percentage JOL and explain JOL conditions, separated based on whether the pairs were related or unrelated. Error bars represent 1 standard error of the mean.

Overall, participants' JOLs were higher for related pairs (M = 77.32, SE = 1.49) compared to unrelated pairs (M = 43.61, SE = 1.49), $F(1,100) = 496.53$, $p < .001$, $\eta^2_p = .83$, $BF_{10} = 1.87\text{x}10^{37}$. Participants gave similar JOLs when they only provided percentage JOLs (M = 60.20, SE = 1.35) and when they provided a percentage JOL and an explanation, (M = 60.73, SE = 1.35), $F(1,100) = 0.40$, $p = .53$, $\eta^2_p = .004$, $BF_{01} = 3.52$, and the pair x judgment interaction was not significant, $F(1,100) = 2.18$, $p = .14$, $\eta^2_p = .02$, $BF_{01} = 2.33$.

Discussion

Across all three experiments, participants gave much higher JOLs to related than unrelated pairs, reflecting past research (see Mueller et al., 2013) and participants' true performance on the later recall tests (see the main manuscript). Participants gave similar magnitudes of JOLs regardless of whether they were given 2s or 4s to make the judgment (Experiment 1) and whether they had to explain their reasoning for the JOL or not (Experiment 3). Participants' binary and percentage JOLs differed from each other in Experiment 2, with participants showing a stronger differentiation between JOLs for related and unrelated pairs in their binary JOLs compared to percentage JOLs. However, it is important to note again that binary and percent JOLs are based on different scales and measure slightly different processes (Zawadzka & Higham, 2015).

Appendix B. Reasons Selected for Explaining JOLs (Exp 3)

In the explain JOL condition in Experiment 3, participants first provided a 0-100% JOL for each pair studied. If their JOL was less than 50%, they had to select from a list of reasons regarding why they made that JOL (only somewhat related, not related, not enough rehearsal, no personal connection, no narrative, other). If their JOL was 50% or greater, participants selected from a different list of options regarding why they made that JOL (strongly related, somewhat related, rehearsed pair, personal connection, narrative, other). In all, participants provided reasons for 976 JOLs (35%) that were under 50% and 1849 JOLs (65%) that were 50% or higher.

Participants were allowed to select as many reasons as they wished for each JOL. Table A1 provides frequencies of how many responses participants selected. Participants' selected reasons were then counted based on the numbers of each response given (e.g., selecting both "No personal connection" and "No narrative" for a JOL were counted as a "No personal connection response" and a "No narrative" response). Counts of the different responses are presented in the sections following.

Table A1. The number of reasons that participants selected (as many as they wished out of 6 possible) for JOLs they provided that were less than 50 and JOLs that were 50 or greater.

# of reasons selected	JOL less than 50	JOL 50 or greater
1 reason selected	708 (73%)	1485 (80%)
2 reasons selected	182 (19%)	293 (16%)
3 or more reasons selected	86 (9%)	71 (4%)

JOLs less than fifty

There were 1355 total reasons selected for the 976 JOLs participants provided that were less than 50%. The percentage of each reason being selected, out of the total number of reasons selected, is presented in Table A2. For JOLs that were 50 or higher, participants selected a total of 2307 reasons. The percentage of each reason selected is presented in Table A3.

Table A2. The percentage of selections for each of the six possible reasons participants could choose from when they provided a JOL less than 50%.

Reasons	% of selected responses
Only somewhat related	10.5%
Not related	42.3%
Not enough rehearsal	21.3%
No personal connection	15.8%
No narrative	7.6%
Other	2.6%

Table A3. The percentage of selections for each of the six possible reasons participants could choose from when they provided a JOL of fifty percent or more.

Reasons	% of selected responses
Strongly related	42.4%
Somewhat related	18.9%
Rehearsed pair	12.5%
Personal connection	14.0%
Narrative	10.3%
Other	2.0%

Discussion

Participants selected a range of possible reasons to explain why they made their JOLs. For some JOLs, they selected multiple reasons. When JOLs were less than 50%, participants most commonly selected that the words were not related (42%) or they did not use enough rehearsal (21%). For JOLs of 50% or higher, participants most often selected the words were strongly related (42%) or somewhat related (19%). Given that participants' responses varied (i.e., everyone did not select "Strongly Related" every time) and that the proportion of reasons differed (i.e., participants' responses were not evenly distributed across all 6 options), this would suggest that most participants considered a specific reason for each JOL and were not selecting a response at random.

Appendix C. Differences Between Online and In-Person Participants (Exp 1)

These analyses explore differences between online and in-person participants from Experiment 1. Attrition/data removal rates were higher for online participants (41%) than in-person participants (19%), which may suggest there were differences among those who completed the study online versus in-person. The analyses are based on the 84 participants who provided useable data (17 in-person, 67 online). However, note that the groups are not equal and are under-powered (particularly the in-person sample), so any results should be interpreted with caution.

Demographics

In-person participants were slightly younger (M = 18.82, SD = 0.95) than online participants (M = 19.52, SD = 1.96). Additionally, gender reports differed slightly between collection groups. Of the in-person participants, 65% identified as female, 29% as male, and 6% as non-binary. Of the online participants, 73% identified as female and 27% identified as male.

JOLs during study

A 2 (collection: in-person, online) x 2 (type of pair: related, unrelated) x 2 (deadline: 8s x 10s) mixed-design ANOVA was conducted on participants' average JOLs to determine whether participants' JOLs differed between online and in-person participants. Overall, collection method did not significantly impact JOL magnitude, and the Bayes factor indicated the null and alternative hypotheses were equally likely, $F(1,82) = 2.09$, $p = .15$, $\eta^2_p = .03$, $BF_{10} = 1.09$. In-person (M = 55.04, SE = 2.46) and online participants (M = 49.50, SE = 2.46) both provided an average JOL of approximately 50-55%. However, the 3-way interaction was significant (although the Bayes factor was inconclusive), $F(1,82) = 6.50$, $p = .01$, $\eta^2_p = .07$, $BF_{01} = 1.34$. Average JOL magnitudes given during study are plotted in Figure A4. Follow-up t-tests

indicated that in-person participants (M = 78.93, SE = 3.36) gave significantly higher JOLs than online participants (M = 68.87, SE = 2.21), for related pairs in the 10s study block (4s to make JOL), $t(82) = 2.13$, $p = .04$, $d = 0.29$, $BF_{10} = 1.77$. No other differences were significant between in-person and online participants' provided JOLs for the other three study conditions (all p's ≥ .29, all BF_{01}'s ≥ 2.27).

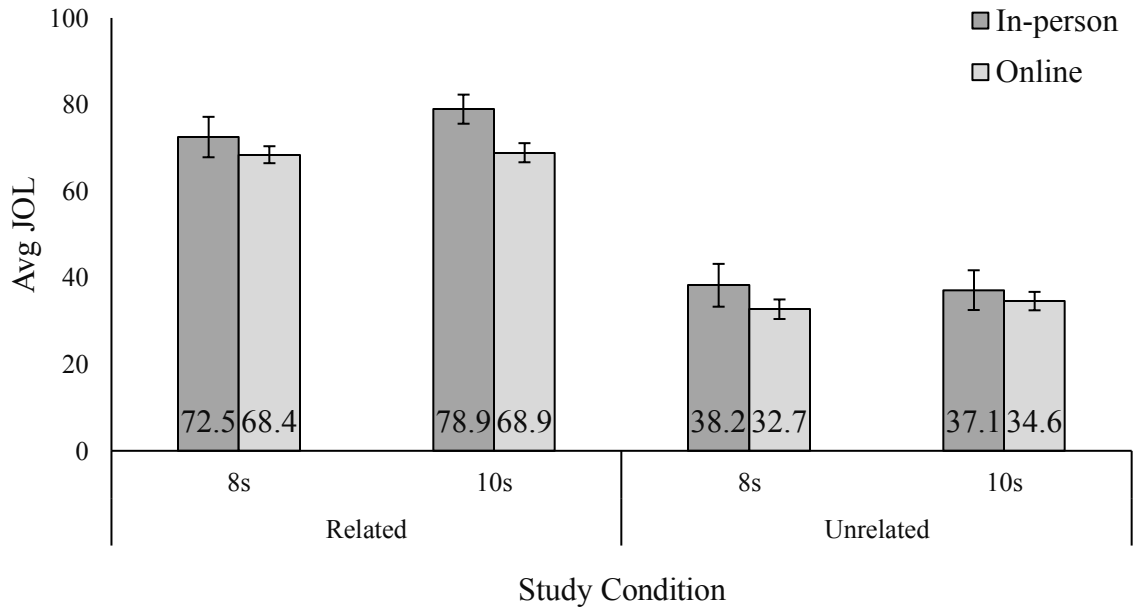

Figure A4. Average JOL magnitudes for each of the 4 study conditions. Data are split between in-person and online participants. Error bars represent 1 standard error of the mean.

Reactivity Size

A 2 (collection: in-person, online) x 2 (study time: 8s, 10s) x 2 (type of pair: related, unrelated) mixed-design ANOVA was also conducted to determine differences among the reactivity effects of JOLs on later memory performance (i.e., reactivity size; see Figure A5). Although online participants showed a trend toward more positive reactivity (M = 3.17, SE = 1.94) than in-person participants (M = 0.95, SE = 1.94), this difference was not significant and the Bayes factor favored the null, $F(1,82) = 0.54$, $p = .47$, $\eta^2_p = .01$, $BF_{01} = 2.42$. Additionally, collection method did not interact with any other variables (all p's ≥ .24, BF_{01}'s ≥ 1.50).

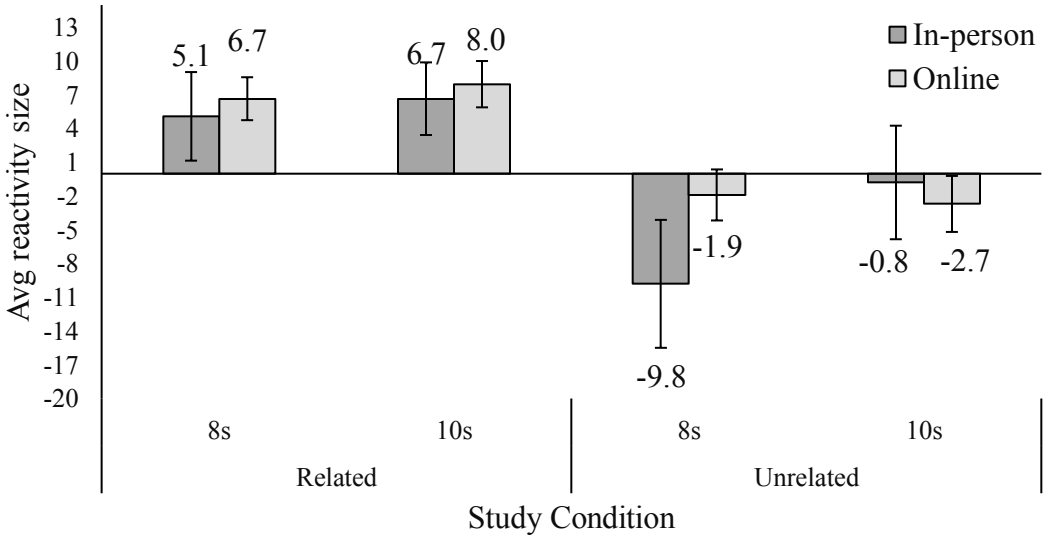

Figure A5. Average difference in percent recalled between trials where participants made JOLs vs did not (i.e., reactivity size) for each of the 4 study conditions. Data are split between in-person and online participants. Error bars represent 1 standard error of the mean.

Discussion

Barring differences in JOL magnitude for related pairs in the 10s study condition (4s to make JOL), no differences in data were observed between online and in-person participants. However, it is important to reiterate that the in-person group in particular was underpowered and the two groups were uneven. To fully understand differences between online and in-person participants, a full, well-powered comparison would be necessary.

Appendix D. Order Effects and First-Block Analyses

Because manipulations were within-subjects and all participants completed study blocks whereby they made JOLs and did not make JOLs, it is possible that being exposed to JOLs impacts participants' learning behaviors and performance in subsequent blocks (in addition to other possible carryover effects). Accordingly, in a series of analyses, I examined the influence of order.

Participants completed the conditions in a unique random order. With three or four conditions in the experiments, there were too many possible order combinations to separate all participants by their unique order. Because the main concern was that providing JOLs in an earlier block would impact how participants approached learning when not providing JOLs (or that experience with the study procedure would impact how participants used JOLs in subsequent blocks), I separated participants based on whether they completed a JOL study block first or a no JOL study block first.

Order Effects

Reactivity Size

Experiment 1. In Experiment 1, 35 participants (41.7%) made JOLs during the first study block while 49 (58.3%) did not make JOLs for the first study block. A 2 (order: JOL first, no JOL first) x 2 (type of pair: related, unrelated) x 2 (study time: 8s, 10s) mixed-design ANOVA was conducted on participants' reactivity size (i.e., the difference in recall performance after participants made JOLs compared to the respective block with the same study time for which participants did not make judgments). See Figure A6 for differences in reactivity size based on first study block. Overall, the effect of JOLs on later recall did not differ between those who made JOLs in the first study block (M = 2.99, SE = 1.73) and those who did not make JOLs

in the first block (M = 1.14, SE = 1.73), $F(1,82) = 0.57$, $p = .46$, $\eta^2_p = .01$, $BF_{01} = 4.03$. The order of study blocks also did not interact with type of pair ($p = .18$, $BF_{01} = 1.87$), nor was the 3-way interaction significant ($p = .65$, $BF_{01} = 4.22$). The order of study blocks (JOL or no JOL first) x study time (8s or 10s) interaction was not significant and the Bayes factor indicated the null was more probable, $F(1,82) = 2.82$, $p = .10$, $\eta^2_p = .03$, $BF_{01} = 1.60$. However, because the interaction approached significance, follow-up t-tests were still conducted.

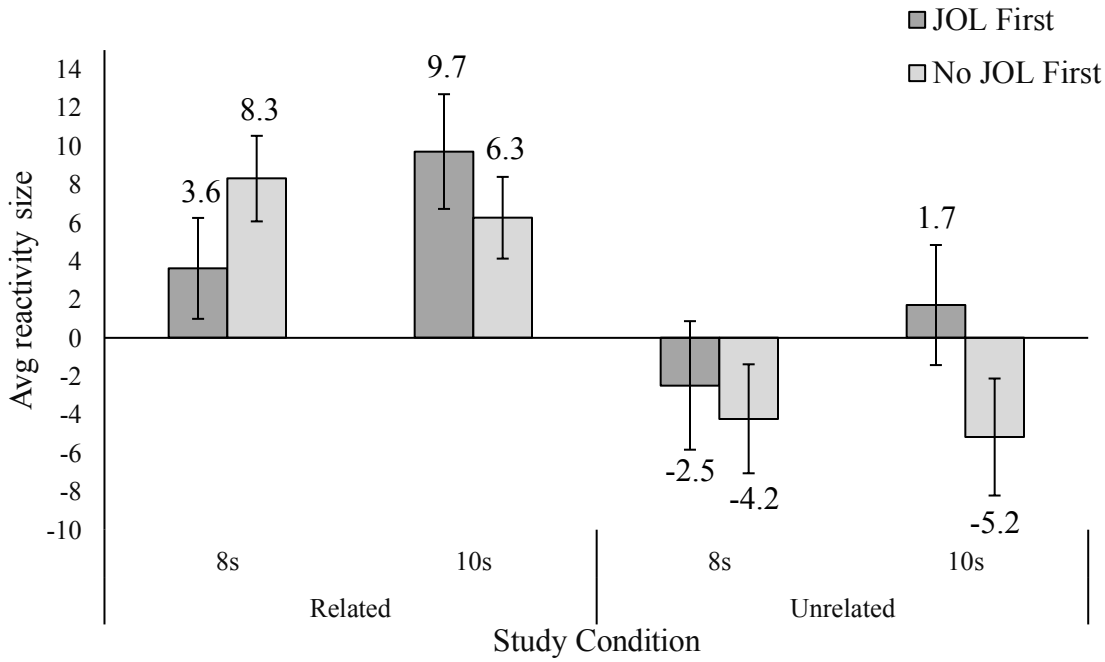

Figure A6. Average difference in percent recalled between trials where participants made JOLs vs did not (i.e., reactivity size) for each of the 4 study conditions. Data are split by whether participants made JOLs or not during their first study block. Error bars represent 1 standard error of the mean.

Those who made JOLs during the first block showed a more positive reactivity effect after they had a 10-second study time (M = 5.71, SE = 2.42) than after an 8-second study time (M = 0.57, SE = 2.42), although the difference was not significant, $t(34) = 1.57$, $p = .13$, $d = 0.27$, $BF_{01} = 1.82$. For those who did not make JOLs during the first block, the pattern was reversed - JOL reactivity was slightly more positive when study time was 8 seconds (M = 2.04, SE = 1.94)

than 10 seconds (M = 0.54, SE = 2.07), although the difference was again not significant, $t(48)$ = 0.63, p = .53, d = 0.09, BF_{01} = 5.34.

Experiment 2. Eighty-two (67.2%) participants made JOLs (either percent or binary) for the first study block, whereas 40 participants (32.8%) did not make JOLs. Figure A7 displays average reactivity sizes for those who made JOLs and did not make JOLs during the first block in Experiment 2. A 2 (order: JOL first, no JOL first) x 2 (judgment: percent JOL, binary JOL) x 2 (type of pair: related, unrelated) mixed-factor ANOVA was conducted on participants' reactivity size. Overall, the effect of JOLs on later recall did not differ between those who made JOLs (either percent or binary) in the first study block (M = 0.80, SE = 1.72) and those who did not make JOLs in the first block (M = 0.19, SE = 1.72), $F(1,120) = 0.51$, p = .81, $\eta^2_p < .001$, BF_{01} = 3.96. First block condition also did not interact with any other variables (all p's ≥ .21, BF_{01}'s ≥ 2.58).

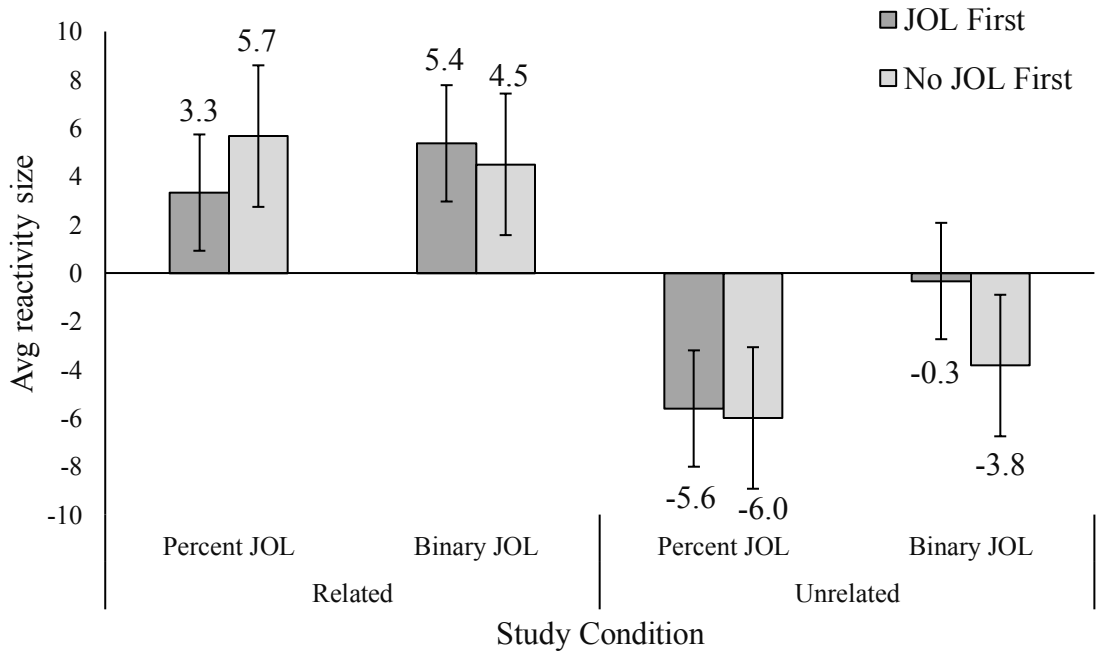

Figure A7. Average difference in percent recalled between trials where participants made JOLs vs did not (i.e., reactivity size) for each of the 4 study conditions (making percent or binary JOLs for related and unrelated pairs). Data are split by whether participants made JOLs or not during their first study block. Error bars represent 1 standard error of the mean.

Experiment 3. Sixty-seven (66.3%) participants were in the percent or explain JOL condition for the first study block, whereas 34 participants (33.7%) did not make JOLs. A 2 (order: JOL first, no JOL first) x 2 (judgment: percent JOL, explain JOL) x 2 (type of pair: related, unrelated) mixed-factor ANOVA was conducted on participants' reactivity size (see Figure A8), with judgment and type of pair manipulated within-subjects and order manipulated between-subjects. Overall, the effect of JOLs on later recall did not differ significantly between those who made JOLs (either percent or explaining) in the first study block (M = 2.42, SE = 2.18) and those who did not make JOLs in the first block (M = -0.21, SE = 2.18), $F(1,99) = 0.69$, $p = .41$, $\eta^2_p = .01$, $BF_{01} = 6.13$. First block condition did not significantly interact with type of judgment ($p = .98$, $BF_{01} = 16.24$), and the 3-way interaction was not significant ($p = .85$, $BF_{01} = 3.70$). However, the pair x first block interaction was significant, $F(1,99) = 5.54$, $p = .02$, $\eta^2_p = .05$, $BF_{10} = 1.48$.

Independent-samples t-tests indicated that, for related pairs, participants' average reactivity size did not differ regardless of whether they made JOLs or no JOLs for the first block, $t(99) = -0.62$, $p = .54$, $d = 0.21$, $BF_{01} = 3.84$. For unrelated pairs, however, making JOLs during the second study block was more detrimental than making JOLs during the first block, although the difference was not significant and the Bayes factor was inconclusive, $t(99) = 1.75$, $p = .08$, $d = 0.21$, $BF_{01} = 1.19$.

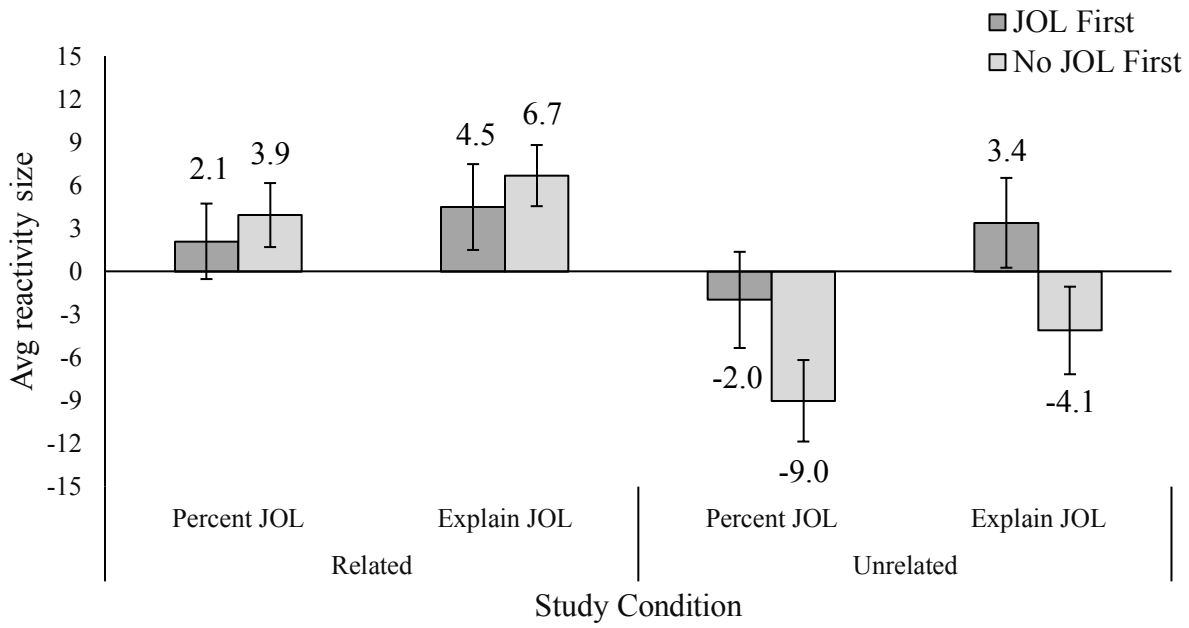

Figure A8. Average reactivity size for percent and explaining JOLs (compared to not making JOLs) for related and unrelated word pairs. Data are split between participants who made JOLs or not during the first study block. Error bars represent 1 standard error of the mean.

First-Block Data

Although which condition participants completed in the first block (JOL or no JOL) rarely interacted with variables across the experiments, order effects did not appear to systematically change the impacts of JOLs on later recall. However, the experiments were not sufficiently well-powered for detecting between-subject, order effect interactions. To obtain a purer measure of the effect of JOLs on memory without the possibility of carryover effects across the study blocks, I also analyzed data from only the participants' first completed block (e.g., for Experiment 1, making 2s JOLs in the first block vs 4s JOLs in the first block vs no JOLs in the first block). Because this resulted in a between-subject comparison rather than within-subjects, these analyses are also under-powered for the effect sizes anticipated. A sensitivity analysis indicated I could only reliably detect an effect size of d = 0.63-0.89 based on

our sample sizes for each experiment with a two-tailed *t*-test between two independent means, using power of 0.8 and alpha of 0.05.

Recall

Experiment 1. A 2 (type of pair: related, unrelated) x 4 (first block judgment: JOL-8s, JOL-10s, no JOL-8s, no JOL-10s) mixed-factor ANOVA was conducted, with type of pair manipulated within-subjects and first block judgment manipulated between-subjects. The analysis indicated that participants recalled more related pairs (M = 78.78, SE = 2.19) than unrelated pairs (M = 35.03, SE = 2.19), $F(1,80) = 385.00$, $p < .001$, $\eta^2_p = .83$, $BF_{10} = 8.60\text{x}10^{31}$. The main effect of judgment was not significant and the Bayes factor indicated that the null was almost six times more likely, $F(3,80) = 2.26$, $p = .09$, $\eta^2_p = .08$, $BF_{01} = 5.76$. More importantly, there was a significant pair x judgment interaction, $F(3,80) = 5.59$, $p = .002$, $\eta^2_p = .17$, $BF_{10} = 18.55$.

See Figure A9 for a display of participants' average recall based on their study condition in the first block. For related pairs, those who started with JOLs and had 8 seconds of study remembered fewer pairs on the later test than those who made JOLs and had 10 seconds of study ($t(33) = 4.83$, $p < .001$, $d = 1.63$, $BF_{10} = 599.85$), those who did not make JOLs and had 10 seconds of study ($t(44) = 3.28$, $p = .002$, $d = 1.00$, $BF_{10} = 17.21$), and those who did not make JOLs and had 8 seconds of study ($t(35) = 3.27$, $p = .002$, $d = 1.08$, $BF_{10} = 15.06$). Thus, the 8-second study time block may not have provided enough time to learn pairs while also making time-pressured JOLs. However, it is important to note that the detriments to memory for related pairs in the JOL-8s condition did not occur when comparing performance based on all four study blocks, rather than only participants' first study-test block. This might suggest that participants who had experience with the study block procedures before receiving the more demanding JOL-

8s condition were more capable of learning pairs efficiently within the restricted time. No other comparisons between the JOL-10s, no JOL-8s, and no JOL-10s were significant. In addition, the detriments to memory for the 8s-JOL block did not carry over to unrelated pairs, and no other conditions significantly differed in memory for unrelated pair (all p's $\geq .15, BF_{01}$'s ≥ 1.44).

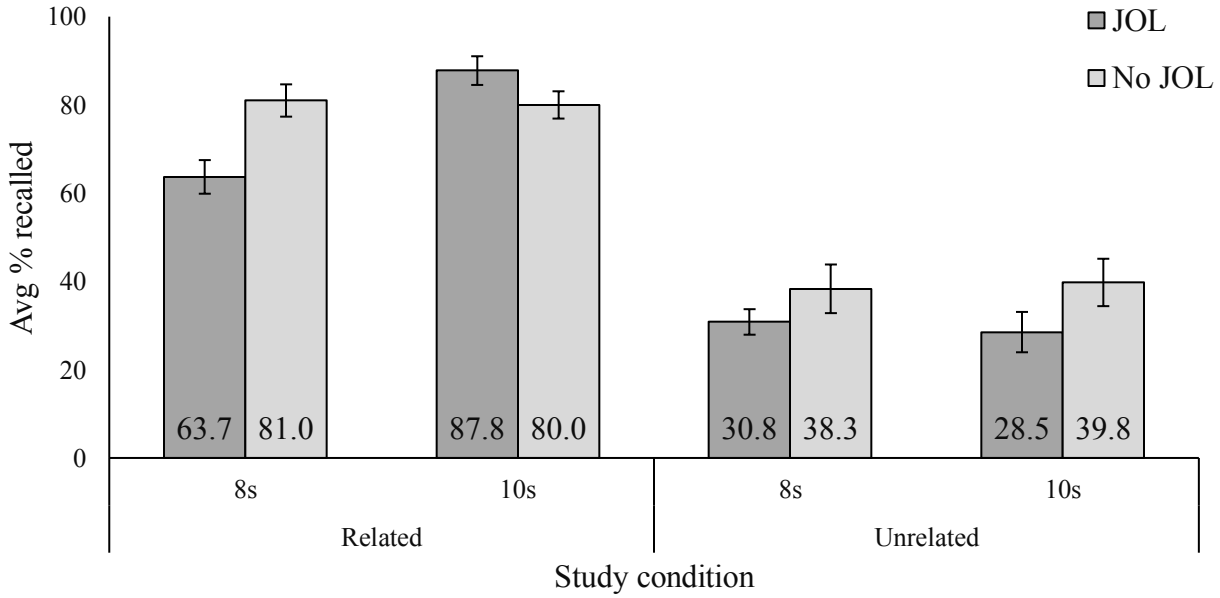

Figure A9. Average percent of related and unrelated word pairs correctly recalled after participants' first study block condition – JOL-8s, JOL-10s, no JOL-8s, no JOL-10s.

Except for JOLs harming memory for related pairs in the 8s study condition (this detriment was not found in the main manuscript when analyzing all four study blocks), results closely mirrored the patterns of findings when all 4 study blocks were analyzed (see p. 26 of the main manuscript). This may suggest that JOLs are more harmful when under time pressure (8s) and unfamiliar with the task, but this decrement is mitigated once participants gain more experience with the study procedures.

Experiment 2. A 2 (type of pair: related, unrelated) x 3 (first block judgment: percent JOL, binary JOL, no JOL) mixed-factor ANOVA was conducted, with type of pair manipulated

within-subjects and first block judgment manipulated between-subjects (see Figure A10 for average recall by condition). The analysis indicated that participants recalled more related pairs (M = 84.80, SE = 1.98) than unrelated pairs (M = 43.73, SE = 1.98), $F(1,119) = 521.74$, $p < .001$, $\eta^2_p = .82$, $BF_{10} = 1.17\text{x}10^{44}$. Although the main effect of judgment was not significant, $F(2,119) = 0.20$, $p = .82$, $\eta^2_p = .003$, $BF_{01} = 16.67$, there was a significant pair x judgment interaction, $F(2,119) = 4.62$, $p = .01$, $\eta^2_p = .07$, $BF_{10} = 3.30$. However, follow-up t-tests indicated that no comparisons between two groups were significant (all p's $\geq .16$, BF_{01}'s ≥ 1.77).

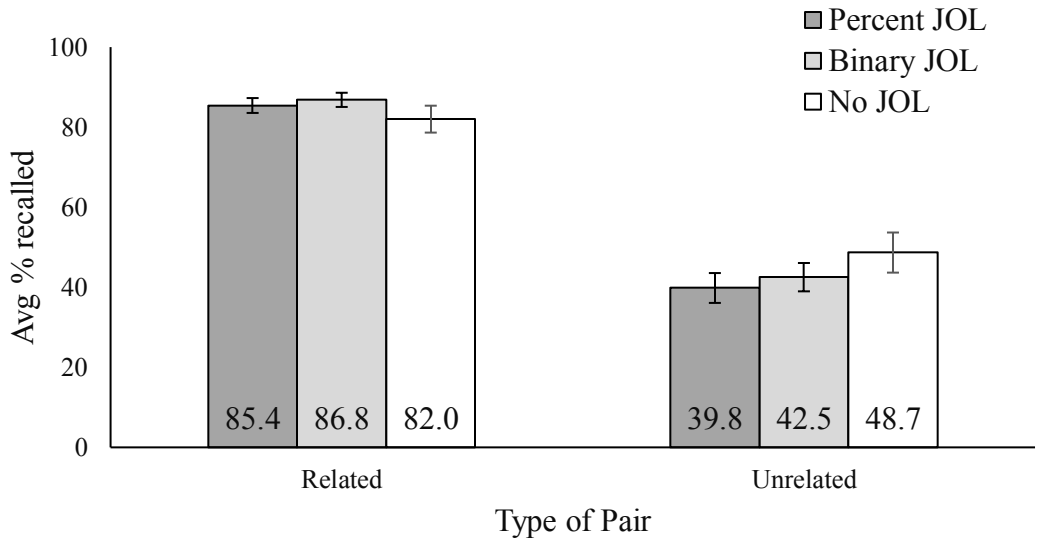

Figure A10. Average percent of related and unrelated word pairs correctly recalled after participants' first study block condition – percent JOLs, binary JOLs, or no JOLs.

Although results from only Block 1 did not significantly replicate the differences found when analyzing all three blocks, the pattern of results remained the same – JOLs slightly improved memory for related pairs and slightly harmed memory for unrelated pairs. Thus, it appears that potential carry-over effects across the three blocks did not impact the pattern of results in Experiment 2.

Experiment 3. A 2 (type of pair: related, unrelated) x 3 (first block judgment: percent JOL, explain JOL, no JOL) mixed-design ANOVA was conducted, with type of pair manipulated

within-subjects and first block judgment manipulated between-subjects (see Figure A11). Overall, participants recalled more related pairs (M = 84.90, SE = 1.97) than unrelated pairs (M = 42.69, SE = 1.97), $F(1,98) = 474.31$, $p < .001$, $\eta^2_p = .83$, $BF_{10} = 1.41\text{x}10^{39}$. Although the main effect of first block judgment was not significant, $F(2,98) = 1.68$, $p = .19$, $\eta^2_p = .03$, $BF_{01} = 6.79$, there was a significant interaction, $F(2,98) = 5.39$, $p = .01$, $\eta^2_p = .10$, $BF_{10} = 6.22$.

Follow-up t-tests indicated that making percentage JOLs or explaining JOLs improved memory of related pairs compared to not making JOLs, $t(67) = 2.83$, $p = .01$, $d = 0.68$, $BF_{10} = 6.75$ and $t(64) = 3.45$, $p < .001$, $d = 0.85$, $BF_{10} = 31.16$, respectively. Explaining and percentage JOLs did not significantly differ from one another in recall of related pairs, $t(65) = 1.01$, $p = .32$, $d = 0.25$, $BF_{01} = 2.59$. JOL condition did not significantly impact memory for unrelated pairs (all p's $\geq .23$, BF_{01}'s ≥ 2.12).

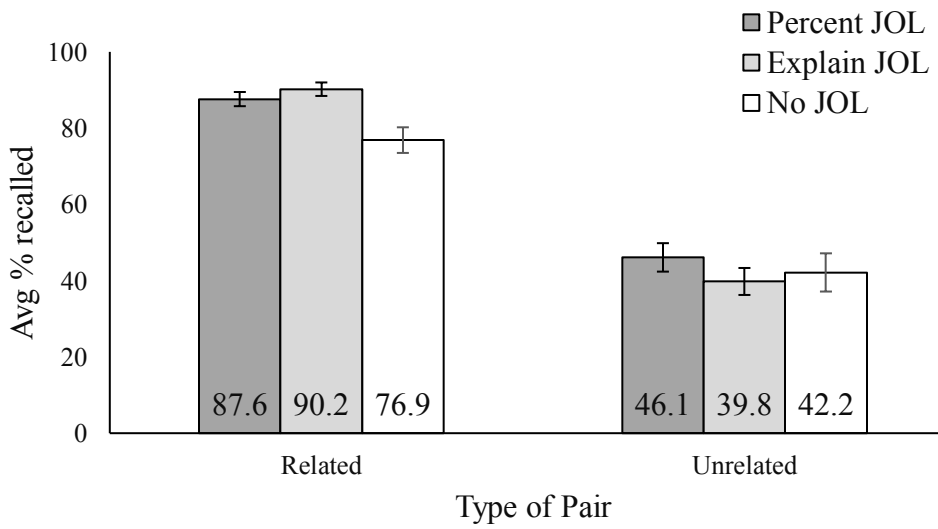

Figure A11. Average percent of related and unrelated word pairs correctly recalled after participants' first study block condition – percent JOLs, explain JOLs, or no JOLs. Error bars represent 1 standard error of the mean.

A few differences in findings occurred when only participants' performance on the first test was considered compared to their performance on all three tests in Experiment 3.

Specifically, although analyses of just the first block and all three blocks demonstrated that recall of related pairs was boosted by making JOLs, this benefit was only significant when considering only the first block. Moreover, JOLs significantly harmed memory for unrelated pairs when data from all three blocks were considered, but JOLs slightly helped (although not significantly) participants' memory of unrelated pairs. These divergences may suggest that JOLs impact participants' performance differently as they gain more experience with the task. However, these observed differences in Experiment 3 were not reflected in Experiment 2, which also compared reactivity effects for percent JOLs vs no JOLs. Therefore, speculation into why experience may change reactivity effects is premature.

Discussion

Overall, the patterns of results remained similar between experiments regardless of whether data were analyzed using only participants' first study-test block or all their study-test blocks. Some differences occurred whereby an effect would be significant with one set of data but not the other (e.g., JOLs significantly improved recall of related pairs in Experiment 2 when considering all study blocks but not when considering only the first block). However, the between-subjects analyses of only the first block were underpowered (because the experiments were powered for the within-subject comparisons using all study blocks), so some slight variations were expected.

Appendix References

Hanczakowski, M., Zawadzka, K., Pasek, T., & Higham, P. A. (2013). Calibration of metacognitive judgments: Insights from the underconfidence-with-practice effect. *Journal of Memory and Language*, *69*(3), 429-444.

Jersakova, R., Allen, R. J., Booth, J., Souchay, C., & O'Connor, A. R. (2017). Understanding metacognitive confidence: Insights from judgment-of-learning justifications. *Journal of Memory and Language*, *97*, 187-207.

Mueller, M. L., Tauber, S. K., & Dunlosky, J. (2013). Contributions of beliefs and processing fluency to the effect of relatedness on judgments of learning. *Psychonomic Bulletin & Review, 20*(2), 378-384.

Zawadzka, K., & Higham, P. A. (2015). Judgments of learning index relative confidence, not subjective probability. *Memory & Cognition*, *43*, 1168-1179.