UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

HOW JUDGMENTS OF LEARNING CAN CREATE ILLUSIONS OF EPISODIC MEMORY

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

DOCTOR OF PHILOSOPHY

By

WILLIAM JOSEPH MUNTEAN
Norman, Oklahoma
2014

HOW JUDGMENTS OF LEARNING CAN CREATE ILLUSIONS OF EPISODIC MEMORY

A DISSERTATION APPROVED FOR THE
DEPARTMENT OF PSYCHOLOGY

BY

_____
Dr. Daniel Kimball, Chair

_____
Dr. Maeghan Hennessey

_____
Dr. Scott Gronlund

_____
Dr. Robert Terry

_____
Dr. Rick Thomas

This dissertation is dedicated to Cristina Neagos Nelson.

Acknowledgements

I would like to thank Dr. Daniel Kimball for all this guidance throughout my graduate training. His enthusiastic approach to scientific research and experimentation influenced my desire to become scientist. I am grateful for Dan's countless contributions to my research projects. As a mentor, Dan not only encouraged me think critically about experimental design and theoretical development but also equipped me with the skills and resources to do so. From my first project as an undergraduate to my last project as a graduate student, I am thankful for each and every opportunity Dan has given me. I look forward to our future collaborations.

I would also like to thank my committee members and fellow lab members for their insight, training, and support throughout my graduate career at Oklahoma. I would like to thank Troy Smith for always pointing me in the right direction whenever I had a research, programming, experimental design, or theoretical question. Similarly, I would like to thank William Landon for always willing to engage me in research discussions; I am thankful for all our creative brainstorming sessions and all the long nights in the lab collaborating on research and class projects. Furthermore, Will was a great friend who supported me throughout all the challenges of graduate school.

Lastly, I want to thank my family for all their support in my journey of becoming an academic scholar. I want to especially thank my mom, Cristina Nelson, for everything she has done so that I could complete my education.

Table of Contents

Acknowledgements iv

List of Tables viii

List of Figures ix

Abstract xi

How Judgments of Learning Can Create Illusions of Episodic Memory 1

Metacognition 5

Direct-access hypothesis 5

Cue utilization 8

Encoding fluency 11

Retrieval fluency 16

Goals of the current research 19

Experiment 1AB 25

Method 30

Participants 30

Design 30

Materials 31

Apparatus 31

Procedure 32

Results 34

Method of analysis 34

Gamma coefficient 34

Generalized linear models 35

Generalized linear mixed models 40
Relationship between JOLs and recall 47
JOLs……47
Recall……...50
Calibration 51
Gamma correlation between JOLs and recall..... 51
GLMM: Regressing recall on transfer item JOLs 52
LMM: Regressing transfer item JOLs on reference item JOLs 55
LMMR Regressing transfer item JOLs on reference item recall 58
LMM: Regressing JOLs on serial position..... 59
Global predictions 60
Discussion..... 61
Experiment 2 63
Method..... 65
Participants 65
Design……..65
Materials 65
Apparatus..... 65
Procedure..... 66
Results. 67
JOLs……67
Recall……...68
Calibration 69

Gamma correlation between JOLs and recall 70
GLMM: Regressing recall on transfer item JOLs 71
LMM: Regressing transfer item JOLs on reference item JOLs 72
LMM: Regressing transfer item JOLs on reference item recall 74
Global predictions 75
Discussion 76
Experiment 3 78
Method 78
Participants 78
Design 78
Materials 79
Apparatus 80
Procedure 80
Results and Discussion 81
Aggregated data analysis 81
Item-level data analysis 84
General discussion 87
Summary 87
Utilization of episodic information 91
References 94

List of Tables

Table 1. LR test: Predicting recall in Experiment 1.....108

Table 2. JOLs predicting recall in Experiment 1.....109

Table 3. LR test: Predicting transfer item JOLs in Experiment 1.....110

Table 4. Reference item JOLs predicting transfer JOLs in Experiment 1.....111

Table 5. LR test: Reference item recall predicting transfer JOLs in Experiment 1.....112

Table 6. Reference item recall predicting transfer JOLs in Experiment 1 113

Table 7. LR test: Predicting recall in Experiment 2.....114

Table 8. JOLs predicting recall in Experiment 2.....115

Table 9. LR test: Predicting transfer item JOLs in Experiment 2.....116

Table 10. Reference item JOLs predicting transfer JOLs in Experiment 1.....117

Table 11. Experiment 3: JOL, Recall, Calibration, and Gamma Means.....118

Table 12. Reference item JOLs predicting transfer JOLs in Experiment 3.....119

List of Figures

Figure 1. Experiment 1: Research design 120

Figure 2. Experiment 1: Raw JOLs histogram 121

Figure 3. Experiment 1: Within-subject centered JOL histogram 122

Figure 4. Experiment 1: Raw JOL means 123

Figure 5. Experiment 1: Within-subject centered JOL means 124

Figure 6. Experiment 1: Percent recall 125

Figure 7. Experiment 1: Within-subject centered percent recall 126

Figure 8. Experiment 1: Calibration Bias 127

Figure 9. Experiment 1: Within-subject centered calibration Bias 128

Figure 10. Experiment 1: Gamma correlation 129

Figure 11. Experiment 1: Within-subject centered gamma correlation 130

Figure 12. Experiment 1: Within-subject JOLs predicting recall 131

Figure 13. Experiment 1: Reference item JOLs predicting transfer item JOLs.....132

Figure 14. Experiment 1: Reference item recall predicting transfer item JOLs 133

Figure 15. Experiment 1: Subjective JOL accuracy 134

Figure 16. Experiment 2: Research design 135

Figure 17. Experiment 2: Raw JOLs histogram 136

Figure 18. Experiment 2: Within-subject centered JOL histogram 137

Figure 19. Experiment 2: Raw JOL means 138

Figure 20. Experiment 2: Within-subject centered JOL means 139

Figure 21. Experiment 2: Percent recall 140

Figure 22. Experiment 2: Within-subject centered percent recall 141

Figure 23. Experiment 2: Calibration Bias 142

Figure 24. Experiment 2: Within-subject centered calibration Bias 143

Figure 25. Experiment 2: Gamma correlation 144

Figure 26. Experiment 2: Within-subject centered gamma correlation 145

Figure 27. Experiment 2: Within-subject JOLs predicting recall 146

Figure 28. Experiment 2: Reference item JOLs predicting transfer item JOLs 147

Figure 29. Experiment 2: Reference item recall predicting transfer item JOLs 148

Figure 30. Experiment 2: Subjective JOL accuracy 149

Figure 31. Experiment 3: Research design 150

Figure 32. Experiment 3: Within-subject centered JOL histogram 151

Figure 33. Experiment 3: Within-subject JOLs predicting recall 152

Figure 34. Experiment 3: Reference item JOLs predicting transfer item JOLs 153

Figure 35. Experiment 3: Reference item recall predicting transfer item JOLs 154

Abstract

Metacognitive judgments made during learning derive from several types of information. These metacognitive cues often reflect intrinsic properties of the to-be-learned material, such as encoding fluency and processing fluency. By contrast, delayed metacognitive judgments depend on internal indicators of memorability, mnemonic cues. When delayed judgments occur under conditions that allow for covert retrieval, retrieval fluency is a potent and reliable indicator of future memorability. Intrinsic cues of material judged during learning (immediate judgments of learning [JOLs]) lead to the illusion of competency. In contrast, mnemonic cues of delayed JOLs enlighten judgers on their current state of episodic memory—but only when covert retrieval is possible. Unfortunately, the generation of potent mnemonic cues is not possible during learning, for the very presence of to-be-learned material prevents its covert retrievability. In contrast, the retrieval of episodically related information, such as previously studied material, is possible. Three experiments explored the effects of retrieving episodically related information on making immediate JOLs. Presented with previously studied word pairs, participants considered them when making JOLs on newly encountered word pairs. Manipulating the presentation of the previously studied word pair varied the likelihood of covert retrieval. Presenting word pairs as cue-only promoted covert retrieval of the episodically related target word, but presenting word pairs as cue-target prevented this occurrence. In all three experiments, when covert retrieval of previously studied information was made possible, immediate JOLs were influenced by the successfulness of that retrieval.

How Judgments of Learning Can Create Illusions of Episodic Memory

When faced with the need to learn new material, gauging the extent to which information is mastered is a critical component to learning. Judgments of learning (JOLs) are used for two important purposes: they serve the purpose of monitoring memory and they can be used to control learning behavior (Koriat, Ma'ayan, & Nussinson, 2006). These two functions are usually thought of as directional such that monitoring memory directs subsequent learning behavior. For example, judgments on the degree to which material is mastered can be used to influence the allocation of resources in order to achieve mastery (Dunlosky & Metcalfe, 2008). Well-learned material receives high JOLs and is unlikely to be selected for restudy, whereas unlearned material receives low JOLs and is likely to be selected for restudy (Metcalfe & Kornell, 2005; Thiede & Dunlosky, 1999).

The directionality that metacognitive judgments govern regulation of study habits inspired the restudy selectivity hypothesis: increasing the accuracy and the validity of metacognitive judgments leads to better restudy decisions and, in turn, the final outcome is greater memory performance on a criterion test (Kimball, Smith, & Muntean, 2013). Appealing to researchers, the hypothesis motivated experiments to focus on discovering elements that make metacognitive judgments more accurate. In the typical JOL paradigm, participants study unrelated cue-target word pairs and then are tasked with judging their confidence in being able to produce the target when provided with the cue on a later memory test (i.e., make JOLs). There is a moderate correlation between JOLs and eventual memory performance when judgments are made immediately after the presentation of each studied word pair. However, JOLs can be

highly accurate if made after a delay and with only the cue word present (Nelson & Dunlosky, 1991). This is a robust finding termed the delayed-JOL effect¬ and is the most widely investigated metacognitive phenomenon (Dunlosky & Metcalfe, 2008).

There are two popular and actively researched explanations of the increased accuracy of delayed JOLs. Nelson and Dunlosky (1991) first proposed a monitoring-dual-memories hypothesis, which assumes JOLs are made by retrieving information from short-term and long-term memory. Judgments made immediately after study are based on information retrieved from short-term memory, whereas judgments made after a delay are based on information retrieved from long-term memory. Short-term memory information is no longer available and accessible during the final test. Thus, the information used to make JOLs is not the same information required on a final criterion test. A final test, temporally delayed from the studied phase, requires retrieval of information from long-term memory, which better matches the information used when making delayed JOLs. According to the monitoring-dual-memories hypothesis, covertly retrieving information from episodic memory leads to better metacognitive judgments.

An alternative account assumes that the delayed-JOL effect is a byproduct of the memory system itself, rather than an increase in a metacognitive skill (Kimball & Metcalfe, 2003; Spellman & Bjork, 1992). This position assumes that participants attempt to covertly retrieve the target word prior to making delayed judgments. Successfully retrieved items receive high JOLs and non-retrieved items receive low JOLs. The covert retrieval process is tantamount to distributed retrieval practice—a very potent memory enhancement technique (Karpicke & Roediger, 2007)—and only items that are successfully retrieved receive such benefits, which extend to the final test.

Although for a different reason than the monitoring-dual-memories hypothesis, the memory-system account also assumes that episodic retrieval leads to more accurate metacognitive judgments.

The current work focuses on a key commonality in the above explanations of the delayed-JOL effect: retrieval of information occurs with delayed judgments (e.g., Nelson, Narnes, Dunlosky, 2004) and it is this retrieval process that brings about more accurate JOLs. The argument advanced here postulates that covert retrieval enlightens participants on two mnemonic properties: episodic nature of the to-be-retrieved item and a general measure of episodic retrievability. The former property breeds item-specific metamnemonic indicators that are used to discriminate between recallable versus unrecallable words and the latter results in broad and holistic metamnemonic awareness that indicates the current state of episodic memory.

Episodically strong items are fluently retrieved from memory (Bjork & Bjork, 2011) and retrieval fluency is a powerful metamnemonic indicator (Koriat, 1997). Prior to making a delayed JOL, for example, a participant experiences a greater sense of retrieval fluency when successfully producing the word pair's target as compared to when they cannot produce the desired target. Retrieval fluency is a potent mnemonic signal that operates at the item level (Benjamin & Bjork, 1996). In addition to item-specific attributes, participants become aware of a more global retrieval property governing their episodic memory under the current conditions of retrieval. An instance of this effect is when a participant comes to realize that, after repeatedly failing to retrieve previously studied items, it is less likely they will retrieve newly encountered information in the future.

At the center of most metacognitive research questions is how to increase participants' sensitivity to item-specific attributes (Dunlosky & Metcalfe, 2008). Motivated by the spirit of the restudy selectivity hypothesis (e.g., metacognitive control regulates study habits; Nelson, 1996), exploiting item-level differences in metacognition allows participants to allocate resources in a manner believed to bring about ideal learning conditions (e.g., Metcalfe & Kornell, 2003, 2005). By and large, when participants can discriminate between learned and unlearned information, they prefer to restudy unlearned information (cf. Dunlosky & Thiede, 2004). Naturally, it would seem that, under most conditions, restudying unlearned information leads to better competency as a whole (but see, Metcalfe & Kornell, 2005). Therefore, increasing item-level metacognitive discrimination, one intention of the current research, appears to be a practically valid and reasonable goal.

However, a recent study by Kimball, Smith, and Muntean (2013) factorially manipulated metamemory accuracy and self-regulation of study and found that the efficaciousness of restudy decisions did not differ as a function of monitoring accuracy: The mnemonic benefit of restudy was shared across items, regardless of restudy choice. This suggests that the mnemonic benefit of restudy due to delaying cue-only JOLs is not because learners can better discern between which items will benefit from restudy, but rather because learners are simply led to select more items for restudy. In light of this finding, a potentially more valuable avenue to pursue is researching metacognitive manipulations that inspire participants to restudy more information. As Kimball et al. noted, one manipulation to increase the number of restudy choices is delaying judgments. This effect is attributed to the engagement of covert retrieval attempts prior

to making JOLs (Nelson, et al., 2004), which brings about an awareness of episodic memory—participants are aware of what they can and cannot retrieve from episodic memory. Ultimately, it is episodic awareness that removes the illusion of competency.

The research goal is to bring the same awareness of episodic memory to immediately studied and judged items. Such a task is not easy given that recently studied items are biased from recency effects; retrieval strength is at the upper asymptotic bound (for a theoretical framework resting upon two classifications of memory strengths, see Bjork & Bjork, 2011; for empirical support within a metamemory paradigm, see Koriat & Ma'ayan, 2005). On the one hand, the disentanglement of short-term and long-term memory may seem unlikely at an item level; after all, the monitoring-dual-memories hypothesis postulates that short-term memory dominates immediate metacognitive judgments. On the other hand, participants may become more aware of their general ability to retrieve information from episodic memory and thereby develop a global sense of how well they can learn information for future retrieval (Koriat & Bjork, 2006a). As such, the current set of experiments promote conditions that encourages episodic retrieval when making immediate JOLs with hopes that the consideration of episodically related information will produce greater metacognitive accuracy.

Metacognition

Direct-access hypothesis

Early theories of metacognition—direct-access hypotheses—assumed participants directly and accurately monitor the strength of memory traces. These theories stem from Hart's (1965) seminal work on feeling-of-know¬ing (FOK)

judgments, which are judgments solicited after a failed recall attempt and measure the ability to recognize the unretrieved information. Hart found that metacognition was moderately accurate under these conditions and proposed that participants had direct access to the strength of memory traces (Hart, 1967; Nelson & Narens, 1990; Burke, MacKay, Worthley, & Wade, 1991). Inspired by the popular search of association (SAM) model of Gillund and Shiffrin (1984), Nelson and Naren (1990) crafted a conceptual framework for the direct-access hypothesis. They implemented metacognition such that simulated participants directly access the parameter in SAM that governs the strengths of associations, which is largely responsible for retrieval effects. To that end, retrieval attempts on the to-be-judged item, even for an immediately encountered item, should promote relatively accurate metacognitive judgments, or at the very least, be correlated with metacognitive judgments. Unfortunately, this is not the case with JOLs: Retrieval of just-encountered items is not highly correlated with JOLs (see, e.g., Nelson et al., 2004; Koriate & Ma'ayan, 2005). This suggests that direct access to the memory trace itself is not the only basis for immediate JOLs, but instead, immediate JOLs are grounded in other information. Thus, the manipulations that provoke episodic retrieval in the current experiments are not centered on retrieving immediately encountered information. Rather, the experimental manipulations encourage retrieval of distant—but episodically related—information, much the same as is thought to occur with delayed JOLs.

The direct-access hypothesis predicts another finding that has seen mixed support: Manipulations that affect encoding will also affect metamemory in the same manner (Schwartz, 1994). For example, strengthening memory traces by encoding

manipulations (e.g., depth of processing, repetition, encoding time, etc.) should increase JOLs, which are based on the now-stronger traces. Begg, Duft, Lalonde, Melnick, and Sanvito (1989) challenged the direct-access hypothesis by collecting ease-of-learning (EOL) judgments for sets of items that vary on two characteristics: imageability or word frequency. The direct-access hypothesis predicts a correspondence between recognition and metamemory judgments across the levels of imageability and word frequency. The data empirically supported the prediction across the levels of imageability: lower EOL judgments and lower levels of recognition for low imagery words than for high imagery words. However, the data did not support the prediction across the levels of word frequency: lower EOL judgments but higher levels of recognition for low frequency words than for high imagery words. The authors postulated that metacognitive judgments were based on more information than that of just measuring objective strength of memory traces.

Data accumulated against a direct-access hypothesis (e.g., Cutting, 1975; Shaughnessy, 1981; Rabinowitz, Ackerman, Craik, & Hinchley, 1982; Mazzoni & Nelson, 1995; Schwartz, 1994; Koriat, 1997), and ultimately, the theory fell out of favor. The theoretical paradigm shifted as researchers proposed heuristic explanations of metamemory (Koriat, 1993; Metcalfe, Schwartz, & Joaquim, 1993; Reder & Ritter, 1992). The general theme of these theories is that judgments are naturally contextualized, either within the judgment task itself or within the to-be-judged items. Contextualization within the judgment task implies that different metacognitive assessments are influenced by different factors and contextualization within the to-be-judged item implies that idiosyncratic (or systematic) properties of the stimuli influence

metacognition. In the spirit of these theories, the current experiments explore heuristics that contribute to the high accuracy of delayed JOLs, attempting to generalize and broaden their reach so they can extend to, and benefit, immediately judged items.

Cue utilization

Rather than assuming direct access of memory traces, heuristic hypotheses assume that metamemory is inferential in nature. By analogy, a direct-access hypothesis maintains that metacognition operates like a thermometer, measuring temperature directly, whereas heuristic theories suggest metacognition operates like a speedometer, inferring speed by measuring axel rotations (Bjork, Dunlosky, & Kornell, 2013). The question becomes, what are inferences drawn from? What exactly is one measuring?

Motivated to determine what factors make metacognitive judgments accurate and inaccurate, Koriat (1997) proposed a general framework describing the sources of information used when making metacognitive judgments. He called the framework *cue utilization*, which reflects the core underlying assumption: Participants consider cues when making metacognitive judgments. Along with this assumption, the theory postulates that metacognitive cues are classified into three categories, and they originate from two sources. According to cue utilization, the three types of metacognitive cues are intrinsic, extrinsic, and mnemonic. *Intrinsic cues* are defined as item attributes, or characteristics, that reveal an item's a priori learning fluency. *Extrinsic cues* pertain to the conditions of learning, either internal factors to the learner (e.g., encoding styles) or external factors (e.g., repetition of stimuli). *Mnemonic cues* refer to internal indicators on how well information is learned.

The cue utilization framework assumes that mnemonic cues originate from

subjective experience (Bjork, Dunlosky, & Kornell, 2013) and generate internalized senses of competency. Take, for example, the retrievability of an item prior to making a delayed JOL. The participant *experiences* the retrieval attempt, and this experience breeds a sense of competency: Easily and successfully retrieved items produce a relatively strong sense of competency, whereas difficult to retrieve or unsuccessfully retrieved items produce a relatively weak sense of competency. In this example, the by-product of experience is a mnemonic cue; an item-specific indication of how well information is learned. However, intrinsic and extrinsic cues can also be born out of experience (see, e.g., Koriat & Ma'ayan, 2005); one experiences the ease that information is processed (i.e., an intrinsic cue) and the encoding strategy applied (i.e., an extrinsic cue).

Whereas mnemonic cues are always generated from subjective experience, intrinsic and extrinsic cues can originate out of beliefs. This is best exemplified in a study conducted by Koriat, Bjork, Sheffer, and Bar (2004). Prior to making immediate JOLs, participants were made aware of the retention interval between study and test, either immediate, a day, or a week. Mean JOLs did not differ as a function of the retention interval; participants neglected to incorporate the extrinsic cue of retention interval when making immediate JOLs (see also Carroll, Nelson, & Kirwan, 1997). Given that people generally have a good understanding of the relationship between forgetting and time (Mazzoni & Kirsch, 2002), this result is somewhat surprising.

Studies prior to Koriat et al. (2004), however, show that intrinsic cues dominate extrinsic cues, and thereby render extrinsic cues ineffective (Koriat, 1997; Koriat, Sheffer, Ma'ayan, 2002; Meeter & Nelson, 2003). This led Koriat et al. (2004) to

hypothesize that encoding created experience-based intrinsic cues (e.g., familiarity of items, pre-experimental associations, etc.) that prevented the use of belief-based extrinsic cues when making JOLs. In a follow-up experiment, presenting a brief synopsis of the experimental design to new participants and asking them to predict aggregated recall levels at each retention interval eliminated the use of intrinsic cues. That is, instead of making item-level JOLs, participants predicted what the recall levels would be for every retention interval: immediate, one day, and one week. This resulted in metacognitive judgments relying on belief-based cues, uncontaminated by intrinsic cues. Consequently, the judgments indicated a decline in recall as retention interval increased—a pattern not observed in the previous experiment, when intrinsic cues were available.

The results of Koriat et al. (2004) underscore a key prediction of cue utilization theory: Intrinsic cues impede the use of extrinsic cues (Kornell, 2010; Kornell, Rhodes, Castel, & Tauber, 2011). For example, rather than accounting for the memorial benefit of extra study time when making JOLs, participants base their judgments on idiosyncratic intrinsic cues (Kornell & Bjork, 2009; Kornell et al., 2011; Koriat, et al., 2004; Meeter & Nelson, 2003; Finn & Metcalf, 2008). Intrinsic cues are potent because they originate through subjective experience (Koriat, 1997; Schwartz, 1994; Benjamin, Bjork, & Schwartz, 1998; Bjork, et al., 2013). By contrast, extrinsic factors are generally less experienced. For instance, it is impossible to experience a future retention interval that has yet to happen. Therefore, conditions of learning that are yet to be realized will have relatively little impact on metacognitive judgments, especially if more dominant experience-based cues are available.

The experiments reported in this paper share a similar agenda with Koriat et al. (2004): to reduce the illusory effects that intrinsic cues have on metacognition. However, in contrast to Koriat et al., where intrinsic cues were eliminated altogether, our approach is to overshadow intrinsic cues with other richer experience-based mnemonic cues bred from retrieving previously encountered information. In doing so, we provide additional information, potential metamnemonic cues, to supplement the cues that participants would otherwise naturally use.

Encoding fluency

The ease with which information is memorized is most commonly considered the main basis for immediate JOLs (Begg, et al., 1989; Hertzog, Dunlosky, Robinson, & Kidder, 2003; Koriat, 1997; Mazzoni & Nelson, 1995; Koriat & Ma'ayan, 2005), and more formally known as encoding fluency (e.g., Undorf & Erdfelder, 2011). In the strict form, encoding fluency refers exclusively to the execution of controlled processes at encoding (but see Begg et al., 1989 for a more general ease-of-processing hypothesis). Regardless of whether the encoding process is elaborative imagery, rote rehearsal, serial association, or any other form, the ease with which the process is executed constitutes encoding fluency (Hertzog et al., 2003).

Encoding fluency is typically investigated by measuring the time spent required to encode material (Koriat, 2008; Koriat, Ackerman, Lockl, & Schneider, 2009; but for an alternative method see, Undorf & Erdfelder, 2011). Using the response time approach, Hertzog et al. (2003) investigated encoding fluency and immediate JOLs by controlling acquisition strategies. In Experiment 1, participants were instructed to engage in imagery during encoding and notify the experimenter immediately after

formulating an image. Upon that indication, participants were allotted 0, 2, or 6 extra seconds of study time, after which JOLs were solicited. While there was not a monotonic increase in global JOLs as a function of additional study time, there was a significant correlation between encoding latencies and JOLs; quicker generated images were associated with higher JOLs. The authors replicated this finding across various encoding strategies. Again, these findings support the key prediction of cue utilization: Experience-based intrinsic cues (e.g., encoding fluency) are favored over extrinsic cues (e.g., increased study time). Furthermore, these data suggest that participants base immediate JOLs primarily on encoding fluency, an experience-generated intrinsic cue.

Immediate JOLs are only moderately accurate at discriminating between recallable and unrecallable information (Nelson & Dunlosky, 1991). In addition, Koriat et al. (2002) found that participants' immediate JOLs (trial 1 in multi-trial learning) exhibit a high degree of overconfidence, which can be detrimental to self-regulation of learning (e.g., Son & Metcalf, 2000). If encoding processes result in using metamnemonic cues that produce illusions of competency, then it is important to understand how and why this occurs; especially when an objective is to reduce the illusory effects, as in the current experiments.

Borrowing from theories in the feeling-of-knowing paradigm, Koriat (1997) suggested that a viable heuristic when making metacognitive judgments is the accessibility of information. When making FOK judgments after a failed retrieval, participants evaluate the amount of accessible information related to the to-be-retrieved item (see e.g., Costermans, Lories, & Ansay, 1992; Nelson, et al., 1984; Koriat, 1993; Dunlosky & Nelson, 1992; Nelson, et al., 2004). Participants might engage in a similar

process during encoding. For example, upon encountering a to-be-studied word pair, participants activate traces related to the associations between the cue word and the target word. Activation of semantic traces is most likely when studying cue-target word pairs. Activating more traces, and thereby increasing access to related information, results in a richer experience-originated intrinsic cue, which manifests in higher subjective metacognitive ratings. Indeed, accessibility of information has a similar influence on feelings of familiarity and source judgments in recognition memory (see, e.g., Kelley & Lindsay, 1993; Kelley & Jacoby, 1998).

Unfortunately, accessibility can lead to systematic biases (Eakin, 2005; Eakin & Hertzog, 2012), much the same as occurs with proportion and frequency judgments made by a similar mechanism, the availability heuristic proposed by Tversky and Kahneman (1973). Additionally, participants have no way of knowing whether the semantically activated traces reflect valid information (e.g., test-enhanced false memory, Kimball, Muntean, & Smith, 2010). The relationship shared between activated traces and the to-be-judged material may be superficial, spurious, and misrepresentative. To some degree, then, relying on semantically accessible information during immediate JOLs will produce inaccurate and inflated judgments—especially if the semantic information has no bearing on eventual recall (Koriat & Bjork, 2006a, 2006b). Thus, metacognitive cues used to make JOLs may not be diagnostic of eventual recall. Under these conditions, cue utilization theory predicts poor metacognition.

One method to manipulate semantically accessible information in word pairs is to vary the association strength between the cue word and the target word (Koriat & Bjork, 2005, 2006a, 2006b). Immediate JOLs are inflated under specific conditions of

association strength, and this occurs regardless of whether association strength is diagnostic of recall. Forward association strength is conceptualized as the likelihood of a cue word eliciting a target word in a word association task, and is diagnostic of cued recall. The reverse is true for backward association strength; the likelihood of the target word eliciting the cue word is not diagnostic of cued recall.

In two separate papers, Koriat and Bjork (2006a, 2006b) investigated metacognitive biases in immediate JOLs. As we do in the present research, Koriat and Bjork focused on mitigating illusions of competency. In their first paper (Koriate & Bjork, 2006a, Experiment 4), participants studied word pairs that varied in forward and backward association strength and then made immediate or delayed JOLs. Based on previous research, they predicted that, when delaying JOLs, participants would minimize the use of experience-based intrinsic cues generated during encoding (Koriat & Ma'ayan, 2005). Instead, delaying JOLs and presenting only the cue word (i.e., cue - ____) would lead participants to covertly retrieve targets and generate richer mnemonic cues—ones based on experience and episodic retrieval. Accordingly, the predicted pattern was such that both types of associations (forward and backward) would lead to similar immediate JOLs; semantic traces would activate regardless of the directionality of association. However, because only forward associations are useful for cued-recall, delaying judgments will result in lower JOLs for backward associations. Covert retrieval enlightens participants on the contents of their episodic memory, removing the illusion of competency. The predicted pattern materialized, delaying judgments removed the illusion of competency for backward associated word pairs. Thus, supporting the author's theory that episodic retrieval produces potent, and accurate,

experience-based mnemonic cues.

The experiments reported in the present research generalize the concept that episodic retrieval reduces the illusion of competency. These experiments are further motivated by the research question of Koriat and Bjork's (2006b) second study on competency illusions: Can participants apply the use of metacognitive cues to new learning situations in order to prevent competency illusions from intrinsic cues? Koriat and Bjork (2006b) explored transfer effects of two potentially bias-mitigating cues: experience-based mnemonic cues generated implicitly from repeated study-test trials and belief-based cues generated explicitly from information about the nature of association strength and cued recall.

The experiment employed a two study-test trial procedure where the second study list consisted of the same items from the previous trial or consisted of new items (i.e., a transfer condition). One set of participants received information about the directionality of association and how only forward association strength was predictive of final cued recall. This was considered the belief-based cue. The other set of participants did not receive any information, and instead, the authors explored whether, through the testing experience, participants could generate a global mnemonic cue that only forward association strength was predictive of recall. This was considered the experience-based cue. Only the belief-based cue had an impact on JOL ratings in the transfer condition. Participants either did not create mnemonic cues during the between-trial testing or did create mnemonic cues and were unable to apply them in the presence of intrinsic cues.

In Experiment 1AB and 2, we use a method that increases the potential to create

and apply mnemonic cues during immediate JOLs. More precisely, we promote the use of covert episodic retrieval of previously studied items during the time of making immediate JOLs on new items. Whereas, Koriat and Bjork (2006b) temporally separated the creation of mnemonic cues and metacognitive judgments, we join these two processes to better represent transfer conditions that are likely to occur under common learning conditions. For example, episodic retrieval of previously encountered information is common when learning material from a textbook. Each new concept builds on previous information, and thus, to determine whether a just-read concept needs to be reread, immediate JOLs are made in relation to concepts previously encountered. If retrieving episodically related information is difficult (or perhaps unsuccessful if you dozed off while reading) then a relatively low implicit JOL is made and the material needs to be reread. Therefore, methods that encourage participants to retrieve episodic information are explored in order to reduce the illusion of competency.

Retrieval fluency

Retrieval fluency, the ease with which information is retrieved from memory (Benjamin & Bjork, 1996; Benjamin, et al., 1998; Matvey, Dunlosky, & Guttentag, 2001; Koriat & Ma'ayan, 2005), is closely related to the accessibility heuristic. In a previous study, Benjamin et al. (1998) explored JOLs under a counterintuitive condition whereby, after a delay, difficult general knowledge questions are better remembered than easy general knowledge questions. A simple explanation for this phenomenon is that participants engage in longer searches of memory where they build stronger and more elaborative associations, which are then useful for retrieval after substantial delays (Gardiner, Craik, & Bleasdale, 1973). Participants were given general knowledge

questions and required to produce answers prior to making JOLs. The authors found that longer answer response times were associated with the lower JOLs, and this effect was taken as strong evidence that participants base JOLs on retrieval fluency.

It seems reasonable to assume that a similar effect occurs for episodic retrieval—easy to retrieve information occurs quickly and is given a high JOL. Koriat and Ma'ayan (2005) tested this assumption. With a rationale identical to Benjamin et al. (1998), latencies for retrieval attempts were recorded just prior to making JOLs. Additionally, the timing of retrieval/JOLs varied across items: Retrieval attempts and judgments occurred either immediately after study, after a short delay, or after a longer delay. Varying the timing of JOLs allowed the researchers to explore the reliance on retrieval fluency for immediate and delayed JOLs. A strong correlation between retrieval latencies and JOL magnitude would indicate that participants use retrieval fluency as a mnemonic cue, much the same as in Benjamin et al. (1998). Furthermore, participants studied word pairs at their own pace, and encoding latencies were measured. Just as with retrieval fluency, a correlation between encoding latencies and JOLs implicates its usage.

The correlation between pre-JOL retrieval latencies and JOLs increased as a function of JOL delay. The correlation was very weak for immediate JOLs and very strong for delayed JOLs. As a further result, there was a stronger correlation between encoding latencies and JOLs when JOLs were made immediately than when made at delays, collectively suggesting that 1) retrieval fluency is used when making delayed JOLs but not immediate JOLs, and 2) encoding fluency is used when making immediate JOLs but not delayed JOLs. While this result is somewhat unsurprising, as others have

postulated or tested the same principle (Kimball & Metcalf, 2003; Spellman & Bjork, 1992; Nelson et al., 2004; Benjamin et al., 1998; Matvey, et al., 2001; Serra & Dunlosky, 2005), it adds support to the proposition that JOLs are based on different information as a function of delay.

Koriat & Ma'ayan (2005) suggest their data help address the answer of whether delayed JOLs are a byproduct of the memory system or a result of heightened monitoring of memory. They advance the argument that differences found in accuracy of metamemory judgments are attributable to using different cues when making judgments, supporting the cue utilization theory (Koriat, 1997). We take their point further and postulate that there need not even be any *metacognitive mechanism*, per se. Instead the man behind the curtain is a standard decision-making process and it is no less or more faulty when making delayed JOLs, but the information it bases decisions upon becomes more valid. This view is compatible with a memory account of the delayed-JOL effect, where successfully retrieved items prior to JOLs are strengthened and are better remembered on a later test. Participants use the same decision system when making immediate and delayed JOLs, but the cues used when making delayed JOLs are more valid (for a discussion on the relationship between memory and decision-making see, Thomas, Dougherty, Sprenger, & Harbison, 2008).

Because retrieval fluency is by and large responsible for the increased accuracy in the delayed-JOL effect, promoting conditions that tap into the same retrieval mechanism seems like a viable option to facilitate the usage of more valid information when making immediate JOLs. Furthermore, ancillary benefits of delaying JOLs, such as a reduction in overconfidence, could extend to immediately judged items. However,

the results of Koriat & Ma'ayan (2005) suggest that retrieval of the to-be-judged item is not predictive of memorability in immediately judged items. Therefore, the methods of the current experiments rely on episodic retrieval of previously studied and judged items. The underlying proposition advanced here is that a global mnemonic cue develops through the collection of episodic retrieval attempts, which then can be used as a basis to make more reliable immediate JOLs. However, unlike in the experiments reported by Koriat and Bjork (2006a), where mnemonic cues were temporally removed from the judgment phase, participants in Experiment 1AB and 2 are faced with a previously studied item, either presented as cue-only or cue-target, and told to consider that item when making immediate JOLs on new word pairs. When the previously studied item is presented as cue-only, participants are predicted to make covert retrievals and evaluate the dynamics of their episodic memory. As a result of this methodology, participants begin generating a global mnemonic cue that is updated and available during every immediate JOL.

Goals of the current research

A metacognitive bias represents a discordance between judgments and performance. In the JOL paradigm, immediate judgments are plagued with bias. The typical pattern is such that participants are overconfident in their performance; the condition produces illusions of competency. Unfortunately, unless metacognitive judgments are perfectly calibrated at an item level, bias exists in other common JOL conditions, such as delaying JOLs. However, whereas immediate JOLs result in overconfidence, delaying JOLs results in underconfidence (see, e.g., Koriat & Ma'ayan, 2005). While a bias nonetheless, underconfidence removes the illusion of competency

and promotes conditions that are desirable for metacognitive control, namely the realization that mastery requires restudying more items.

Regarding explanations of the delayed-JOL effect, the cue utilization framework postulates that retrieval fluency serves as a potent mnemonic cue—an internal indicator of how well information is learned. Furthermore, retrieval fluency is a *valid* mnemonic cue in that it has a more accurate relationship with recall (i.e., predicts recall). By extension, retrieval fluency serves as a valid cue to trigger a more general internal metamnemonic awareness. Thus, although retrieval fluency produces mnemonic cues for individual items, influencing individual item-JOLs, this does not preclude the development of other experience-based cues that apply globally (see, Dunning, Johnson, Ehrlinger, & Kruger, 2003). It is reasonable to believe holistic mnemonic cues contribute to the observed underconfidence concomitant with delaying JOLs.

One primary goal of the current paper is to produce desirable biases in immediate JOLs, biases that reduce illusion of competency (see, e.g., Koriat & Bjork, 2006) and promote conditions desirable for restudy selections. Furthermore, we take steps to increase the ecological validity of metacognition, whereby judgments of learning are not made in isolation from one another, but instead are made in relation to one another (Koriat, 1997; Dunlosky & Matvey, 2001; Castel, 2008). By way of illustration, imagine reading a current behavioral methods article and learning about a new statistical concept. Progressing through the article, you covertly make judgments of learning on newly acquired information; after all, your intention is long-term proficiency of this new statistical method. However, consecutive judgments of learning (or more broadly, comprehension) are not isolated from one another. The retrieval of

previously encountered information, and hence, a shared episodic relationship, occurs when making each new judgment. In this illustration, semantic overlap of information facilitates covert episodic retrieval. Here, and in most common learning situations, episodic retrieval happens naturally.

When making JOLs, episodic retrieval is important. As outlined in previous sections, retrieval fluency is responsible for the delayed-JOL effect and removes the illusion of competency. However, immediate episodic retrieval of just-encountered material is contaminated by recency. Mnemonic cues are unable to adequately form, or worse, develop into foresight bias (Koriat & Bjork, 2005), *increasing* the illusion of competency. This concept is similar in spirit to the conclusion drawn from a study by Kelley and Jacoby (1996). Participants predicted the difficulty of solving anagrams either in the presence or absence of the solution. Predictions were much more accurate when the solution did not accompany the anagram, an effect attributable to subjective experience. The lack of subjective retrieval experience when making immediate JOLs is detrimental to metacognition.

In the statistical method example outlined above, episodic retrieval is not of the just-encountered information. Instead, retrieve occurs for earlier encountered information. Covert retrieval replenishes subjective experience, creates mnemonic cues, and is used as a source when making JOLs. In this case, immediate JOLs reflect the extent that prior information is difficult to retrieve. Ultimately, the effects of delaying JOLs carryover to immediate JOLs: If one cannot covertly retrieve the equation for the density of a Weibull distribution, then it is unlikely that the newly encountered equation for the density of a 3-parameter lognormal distribution will be retrieved in the future.

Common experimental designs investigating immediate JOLs do no promote episodic retrieval of previous information. Typical procedures include studying cue-target word pairs that do not share any a priori relationship, and more importantly, the word pairs across the list do not share any relationships. Furthermore, during the judgment phase, only a single isolated word pair is presented. Both the lack of relatedness across word pairs and making judgments in isolation drastically reduce the likelihood of episodic retrieval when making immediate JOLs. To produce conditions that are more similar to everyday learning situations, experimenters can manipulate either one or both of these factors.

Manipulating semantic associations between a word pair's cue and target has been done in several studies (see, e.g., Koriat & Bjork, 2005). As mentioned above, semantic relatedness produces a heightened feeling of competency. This presumably happens because of the ease of access to semantically related information and encoding fluency. A greater amount of (semantic) information is retrieved for related word pairs than for unrelated word pairs, giving the illusion of strong episodic memory traces. Similarly, creating associations between related cue-target word pairs is easier than for unrelated cue-target word pairs, and therefore, elaborative encoding is perceived as superficially efficient (but see, Koriat & Bjork, 2006a for examples when semantic information is diagnostic of eventual recall).

To promote episodic retrieval through semantic associations, relationships between different word pairs must develop throughout the study list. Relational processing of this type is not the typical encoding strategy employed by participants. Of course, the main reason this encoding strategy is uncommon is because relational

processing across word pairs does not transfer to better retention as measured by a cued recall task (see, e.g., Mulligan & Lozito, 2004; McDaniel & Waddill, 1990). Once participants are made aware of the retrieval task (e.g., being able to recall the target word when prompted with a cue word) they engage in cue-target relational encoding. One method to increase list-wide relational encoding is to increase semantic associations between different word pairs in a study list. While this methodology seems fruitful, there is no assurance that participants will notice the relationships across word pairs and engage in episodic retrieval.

Experiments 1AB and 2 use a method much more certain to promote episodic retrieval during immediate JOLs. To be exact, when making immediate JOLs, we presented a previously studied cue-only or cue-target word pair to participants and instructed them to consider that word pair when making immediate JOLs. This procedure creates a discrete classification of items. The previously studied word pair serves as a *reference item*: It brings about a mnemonic awareness by referencing episodic memory. The item given immediate JOLs serves as a *transfer item*, an item to which the mnemonic cues available from the reference item are intended to transfer.

Manipulating the presentation format of the reference item (cue-only versus cue-target) varies the degree and type of episodic awareness. Given that covert retrieval is thought to occur when making delayed cue-only JOLs (Nelson et al., 2004), we assume that instructing participants to consider a cue-only reference item will promote a similar covert retrieval attempt. This results in a rich recollective-based awareness. Much like in the example of learning a new statistical method, we anticipate the memorability of reference items to influence metacognitive judgments of transfer items, but only when

reference items have a strong influence on episodic awareness—when they are presented as cue-only.

Presenting an intact reference item (i.e., cue-target) deprives the learner of the subjective experience of retrieval (e.g., Kelley & Jacoby, 1996) and affords little recollective-based episodic awareness, even after a delay (Dunlosky & Nelson, 1992). Rather, a cue-target reference item offers a recognition-type awareness of episodic memory: The reactivated episodic traces are a mixture of recollection and familiarity processes (see latent recognition models of memory, e.g., Wixted, 2007; Yonelinas, Dobbins, Szymanski, Dhaliwal, & King, 1996; DeCarlo, 2002). Importantly, recognition-based traces are not efficacious (Dunlosky & Nelson, 1992)—they do not increase the accuracy of metamemory. As such, any effects of cue-target reference items are attributed to a mixture of processes (e.g., cue-target familiarity, processing fluency, etc.). In short, the relationship between reference and transfer items addresses the question of whether episodic insight can transfer to new items, and more specifically, whether episodic awareness can reduce the illusion of competency.

We used a less direct and more implicit approach in Experiment 3: We presented multiple word pairs to be judged during the same timeframe. In Experiment 3, a previously studied word pair appeared above an immediately studied word pair, and participants were instructed to give JOL ratings in the order in which the word pairs appeared (i.e., top to bottom). Hence, participants made delayed JOLs on *reference* items just before making immediate JOLs on *transfer* items. Compared to making JOLs in isolation (as is typical in JOL experiments), this methodology was intended to increase the propensity to use episodically related information. This is specifically

because the two simultaneously presented to-be-judged word pairs are episodically related. As in Experiment 1AB and 2, the presentation of the reference item was manipulated. Cue-only reference items provoke covert episodic retrieval, generating an experience-based cue. However, cue-target reference items activate different types of episodic information, namely through a familiarity process. This methodology differs from Experiment 1AB and 2 in that no requirement is placed on considering the use of prior information. Instead, mnemonic cues were generated by a delayed JOL and their natural effect on immediate JOLs was evaluated.

Experiment 1AB

Participants in Experiment 1 made JOLs on three classifications of items: *reference items, transfer items,* and *control items*. The first third of the study list consisted of *reference items*, for which judgments were made immediately following study presentation, with only the cue word present (see Figure 1). Reference items served the purpose of referencing subjects' current state of episodic memory when they subsequently made immediate cue-only judgments on *transfer items*. This classification label reflects the general research question of whether insight into episodic memory can transfer to newly studied items. The methodology of Experiment 1 presented a previously studied reference item during the judgment phase of a transfer item. By design, transfer items were always judged in the presence of a reference item. While the first third of the study list consisted of only reference items, the last two-thirds of the study list contained transfer items as well as *control items* that received immediate cue only judgments.

The controlled presentation of reference items manipulated the subjects'

awareness of their episodic memory. Reference items were presented in either cue-only or cue-target form. It was reasoned that providing the reference item as cue-only would strongly enhance episodic awareness through covert retrieval of the target word, much the same as is thought to occur in the delayed-JOL effect (Nelson, et al., 2004, Son and Metcalf, 2005; Kimball & Metcalf, 2003). By contrast, presenting cue-target, or intact, reference items weakens the subjective experience of episodic retrieval (Kelley & Jacoby, 1996). A potent mnemonic cue, according to cue utilization theory, is the retrieval fluency that emerges during the target's absence. However, presenting the target alongside the cue interferes with the use of internal mnemonic signals and thus provides relatively little insight into episodic memory.

There is potential for two reference item characteristics to influence transfer item JOLs, encoding fluency and retrieval fluency. Regarding the latter, participants are likely to make covert retrieval attempts when faced with cue-only reference items (Nelson et al., 2004). A failure to produce the correctly paired target gives rise to strong mnemonic cues and we predict that, when retrieval failure occurs, participants are more likely to select lower JOLs for transfer items. This is not because the transfer item was less well learned than other items but because participants are enlightened; they understand the properties (e.g., fallibility) of episodic memory. It alters, and perhaps better calibrates, the perspective participants take when evaluating new data. However, this only occurs under conditions that allow for recollective episodic retrieval attempts (i.e., cue-only reference items).

A test of this prediction entails comparing transfer item JOLs as a function of whether the reference item was recalled on the final test. If a participant can produce the

correct target on a final test then it is reasonable to assume they can successfully produce it during time of transfer item JOLs (e.g., Nelson et al., 2004). Accordingly, transfer item JOLs will be higher for reference items that are recalled on the final test and lower for reference items that are not recalled on the final test. But most importantly, this effect will only emerge in the presence of cue-only reference items. Observing this pattern would provide strong evidence that covert retrieval of previously studied items, which is possible with cue-only reference items, affects immediate JOLs.

The second characteristic that has potential to influence metacognition for transfer items is encoding fluency of the reference item. Easy to process reference items may elicit false senses of competency (Metcalf, et al., 1993) and carry over to transfer item JOLs. In the current experiment, immediate JOLs given to reference items in the beginning of the study list serve as the best proxy for their encoding fluency (Koriat & Ma'ayan, 2005). Therefore, to test whether the encoding characteristics of reference items influence transfer item metacognition, we regress transfer item JOLs on reference item JOLs. On the one hand, easily processed reference items may give participants the impression that transfer items are relatively more difficult to process. To the extent that this occurs, transfer items would be negatively correlated with reference item immediate JOLs. On the other hand, easily processed reference items may produce a general sense of competency, which can be misattributed to the transfer item. The observed pattern, in this case, would be positively correlated immediate JOLs between the two sets of items.

The general prediction in the current experiment is that cue-only reference items increase the propensity for participants to adopt the belief that their memory is fallible. As a whole, this manifests in a reduction in overconfidence, reducing the illusion of

competency. Adopting this perspective implies the potential for an ancillary effect. A participant's a priori belief will determine the utility of new evidence. The new evidence may be one, or the combination of several, intrinsic or mnemonic cues. For example, consider two participants who either give low or high JOLs for the first set of items (i.e., the reference items), personifying their a priori belief about their memory. When judging transfer items, the two participants—who differ in subjective beliefs about their memory—will view easy-to-process reference items differently. For instance, one potential pattern would be that the participant who thinks they have poor memory (mistakenly) interprets the intrinsic cue generated by the easy-to-process reference item as convincing evidence of memorability for the transfer item. By contrast, the participant who believes they have a strong memory discounts the intrinsic cue—it is consistent with their hypothesis (e.g., that they have a good memory), and therefore has less of an impact on JOLs. This account would predict a stronger correlation between reference item JOLs and transfer item JOLs for participants who give lower immediate JOLs to reference items than participants who give higher immediate JOLs to reference items.

However, the reverse pattern is equally probable. For example, those who give high JOLs to the first set of items may be more persuaded by superficial intrinsic cues generated during encoding of the reference items. Then, the reappearance of the reference items during the judgment phase of transfer items produces a similarly persuasive but superficial intrinsic cue of competency. The opposite effect would occur with those participants who rate the first set of items with lower JOLs; they are insensitive to intrinsic cues, reducing the correlation between reference item and

transfer item JOLs. Finally, neither pattern could emerge, and instead, the effects of referencing previously encountered studied items operate similarly across participants who have different subjective beliefs about their memory.

A further prediction entails desirable effects carrying over to control items, which were studied and judged in the same temporal proximity as transfer items. In particular, control items will exhibit lower JOLs as compared to earlier judged reference items. If the awareness of episodic fallibility is global then items rated after enlightenment will share the same fate. This would demonstrate the powerfulness of a global mnemonic cue generated from covert episodic retrieval. The mnemonic cue alters the judger's inherent perspective, affecting subsequent ratings on new items and under different conditions. Such an effect might be contradictory to the results reported by Koriat and Bjork (2006b). They found that only belief-based, and not mnemonic-based, debiasing transferred to new items. In light of their results, we asked participants at the end of the study phase to rate their global accuracy in making the different types of JOLs (i.e., transfer item ratings versus control and reference item ratings). If participants believe they are more or less accurate under different conditions, then this can be taken as strong evidence that reference items brought about belief-based cues (i.e., explicit awareness that one condition leads to more accurate predictions) in addition to experience-based mnemonic cues (i.e., encoding or retrieval fluency of reference items). The presence of both, lower JOLs for control items and differences in global accuracy predictions, would be consistent with Koriat and Bjork's findings: debiasing transfers to new situations with the existence of theory-based cues.

Finally, the last prediction explored in the current experiment tests the core

assumption of cue utilization framework: intrinsic, extrinsic, and mnemonic cues are relative in nature (Koriat, 1997). This implies that ratings are, to some degree, temporally related and therefore not independent from one another (see, e.g., Dunlosky & Matvey, 2001). This is inline with all our earlier predictions (e.g., participants' perspective drifts, reference items affect transfer item JOLs, underconfidence for control items). However, we take an additional and more explicit approach by testing sequential relationships—we test whether JOL residuals are correlated as a function of serial position. This method is discussed in greater detail in the method of analysis section.

Method

Participants

Seventy University of Oklahoma undergraduate students enrolled in the introductory psychology course participated for partial course credit (n = 37 for Experiment 1A and n = 33 for Experiment 1B; 4 participants out of the 37 in Experiment 1A were excluded for failing to follow instructions, they inaccurately described the instructions in a debriefing question, see below). The experiments were run during the same timeframe, with participants from the same population, and with only a single manipulation change—the presentation of the reference item during transfer item JOLs—and therefore the experiments were collapsed and the presentation of the reference item was treated as a between-subject factor hereafter.

Design

Item type (*reference*, *transfer*, and *control*) was manipulated within-subjects and the likelihood of episodic retrieval (likely retrieval, unlikely retrieval; e.g., the

effectiveness of the reference item promoting episodic retrieval) was manipulated between-subjects. Participants in both conditions made immediate cue-only JOLs on the first third studied word pairs (i.e., reference items). Metacognitive judgments were made in a similar fashion for half of the remaining items (i.e., immediate cue-only JOLs; control items). Judgments for the other half of the remaining items (i.e., transfer items) were made in the presence of a previously studied item (i.e., reference items) from the first third of the study list. The presentation of the reference item at the time of transfer item JOLs was manipulated between-subjects; one group of participants viewed cue-only reference items and the other group viewed cue-target reference items when making transfer item JOLs.

Materials

The stimuli comprised 180 four-letter common nouns, with high word frequency, familiarity, concreteness, and imageability (cf. Kimball, Smith, & Muntean, 2012). Word pairs were randomly paired anew for each participant, with 30 pairs assigned to each of the three within-subject conditions (reference, transfer, and control items). All 90 word pairs were studied and tested.

Apparatus

The experiment was programmed in the JavaScript programming language and implemented in Qualtrics' framework, an Internet survey provider. While this method was fully *Internet* capable (i.e., able to collect data through internet participants) we chose to administer the experiment on our on-campus lab PCs through the Chrome web browser. Our tests indicated that Chrome's JavaScript engine rendered stimuli much quicker and much more stable than either Internet Explorer or Mozilla/Firefox web

browsers.

Procedure

The experiment consisted of an instruction phase, followed by a single study/cued-recall test session. After participants consented to the experiment, the computer displayed an interactive instruction set, in which the participants could navigate through and whereby the computer revealed the process of the experiment in a step-by-step nature, including examples of each task the participant would encounter. The instructions began with a general description of the study list—that participants were to study 90 cue-target word pairs, each presented for 4-seconds, and then make judgments on their confidence in recalling the target word when provided the cue on a subsequent memory test (e.g., memory judgments).

The instructions then elaborated the exact nature of the memory judgments. Participants were told that following some word pairs they would see the statement, "Indicate your confidence in being able to recall the target word studied with cue - _____ on a later memory test", and then instructed to indicate their confidence on a scale shown below. The scale ranged from 0 (*not confident at* all) to 100 (*very confident*), with single digit increments. The cue word was shown left of the scale (i.e., "cue - _____") and participants were informed the word "cue" would be replaced with an actual studied cue during the experiment.

Participants were then told some memory judgments were to be made in the presence of additional information. The likely-retrieval condition was told the computer would display the phrase, "Consider the following previously studied word pair: cue - _____ when judging your confidence in recalling the missing target BELOW: cue -

____". Participants in the unlikely-retrieval condition were told the computer would display the phrase, "Consider the following previously studied word pair: cue - target when judging your confidence in recalling the missing target BELOW: cue - ____". As can be seen, the only difference between the conditions was whether the reference item was presented as cue-only or cue-target. The same scale was displayed below the prompts for all judgments.

Following the delivery of instructions, participants were posed with an open-ended question requiring a description of the two *different* memory judgment tasks. Four participants in the reference-effective failed to accurately describe a difference between the two tasks and were removed from all subsequent analyses.

The experiment began by presenting each word pair at the top of the screen for 4-seconds. Immediately following each of the first 30 word pairs, the prompt requesting cue-only JOLs was displayed. The order of conditions for the remaining 60 word pairs (i.e., the transfer and control items) was randomized anew for each participant. In a similar fashion, the prompt soliciting JOLs immediately followed each studied word pair. All judgments were self-paced.

The computer posed two questions to each participant following the last studied item and before the final cued-recall test began. The first question asked, "How accurate do you think your memory predictions were when considering additional information?", and the second asked, "How accurate do you think your memory predictions were when you DID NOT have to consider additional information?". The same scale was used as before and the questions were self-paced. The cued-recall test required participants to

type the target studied with the provided cue. Participants continued testing each studied word pair at their own pace.

Results

Method of analysis

In the section that follows, we briefly described several statistics and methods of analysis, offering our rationale and justification. Specifically, we introduce the use of (generalized) linear mixed-models (GLMM; LMM) as our main analysis for inferential testing, which comprise both null hypothesis significance testing of parameter estimates and model comparison.

Gamma coefficient

Nelson (1984) proposed using the nonparametric measure of association, gamma coefficient (*G*; Goodman & Kruskal, 1954), to quantify relative accuracy—the likelihood that an item given a higher JOL will be recalled versus an item given a lower JOL. Gamma is the most used measure in metamemory and is formulated by calculating all possible pairwise combinations of *J* items and taking the ratio of concordant pairs (where the ordering of the predictor is consistent with ordering of the outcome variable for the pair) minus discordant pairs (where the ordering of the predictor is inconsistent with ordering of the outcome variable for the pair) over concordant plus discordant pairs. All pairwise ties are discounted and not used in the calculation (i.e., any pair of items in which either both items were recalled, not recalled, or given identical JOLs). Subjects without variance in either JOLs or recall accuracy cannot be calculated. Gamma takes on values between negative one and positive one, with positive one indicating perfect resolution.

Generalized linear models

The traditional method of inferentially testing effects is through an ANOVA. Typically, the design matrix of ANOVAs consists of classifier variables. The variation in (batches of) coefficients determines whether the observed data are more extreme than would be expected given the null model. We use this traditional method to analyze aggregated level data, such as gamma correlation, calibration, and JOL means. More complex and powerful models analyze item-level data. Before describing those models we outline several prerequisites.

An ANOVA is a member of the linear model family—a categorical model. Assessing association with a continuous independent variable is also accomplished through a linear model. Equation 1 shows the simplest formulation.

$$y_i = \alpha_0 + \beta_1 x_i \qquad 1$$

Outcome variable y_i indicates the response for the ith item, α_0 is the parameter estimate for the intercept, with interpretation depending on the coding structure of the continuous independent variable (e.g., centering, scaling, reference point) or the categorical predictor (e.g., effect coding, simple coding, treatment coding, etc.), and β_1 is the parameter estimate representing changes in the outcome variable per each unit change in X, the predictor.

The association between the outcome and the predictor, then, is quantified by the coefficient, β_1, which is theoretically unbounded but practically restricted by the scale and range of the data. As a consequence, discussing the coefficient in terms of

strength can be somewhat misleading in metamnemonic paradigms. For example, perfect item-level calibration (i.e., greater recall for items receiving greater JOLs than items receiving lower JOLs) would be akin to a slope of 1, interpreting of slopes greater than 1 as *too strong* of an association between JOLs and recall is a dissatisfying characterization. Thus, our recommendation is to avoid—to the extent possible—describing the relationship in terms of strength, despite the traditional terminology in linear models, and instead, describe the coefficient relative to an identity slope. This latter point poses some challenges and will be discussed shortly.

Regarding metamnemonic paradigms, the predictor variables are typically continuous (e.g., JOLs) and the outcome variable binary (e.g., recall accuracy). As such, model residuals are not normally distributed but rather binomially distributed. To accommodate these properties, a link function is placed on the linear component and results in Equation 2.

$$
\begin{aligned}
&y_i \sim \text{Binomial}(n, \pi) \\
&f_{\text{logit}}(\pi) = \eta = \alpha_0 + \beta_1 x_i \\
&\pi = f^{-1}{}_{\text{logit}}(\eta)
\end{aligned}
\qquad 2
$$

Equation 2 represents an inverse logit link placed on η which represents the linear combination of variables in Equation 1.The dependent variable, y_i, is binomially distributed with probability π in *n* trials, with π equaling the inverse link function of the linear combination of the fixed components in Equation 1. The variance of the model reduces to π(1- π).

Several link functions exist for binary data. Among the most common are, a

probit link, which utilizes a cumulative normal distribution and is the most common distribution in signal-detection theory (Wickens, 2002), a generalized logit link for multinomial data, and a log link for count data modeled through a poisson distribution. A link function generalizes the linear model. Consequently, these models are part of the *generalized linear model* (GLM) family.

Two link functions make most sense for JOL data, probit and logit links. The latter has the advantage with respect to the ease of which coefficients can be interpreted. Specifically, converting log-odds into an odds ratio make for an intuitive interpretation. For example, investigating item-level recall using an intercept-only model might yield a logit coefficient of -1.1. The odds of recalling an item can be inferred by taking the exponential of the estimate. Thus, there is a one-third chance of recalling any given item.

The predictability of metacognitive judgments on recall can now be analyzed succinctly in terms of a generalized linear model. The simplest model is shown below,

$$\eta = \alpha_0 + \beta_1 * JOL_i \qquad 3$$

with the intercept representing the log-odds of recalling an item given a zero JOL value and B1 representing the change in log-odds of recalling an item with each increase in JOL. The model can be refit with mean-centered JOLs, leading the intercept to represent the log-odds of recalling an item at the average JOL in the sample. For interpretation purposes, it is important to center continuous variables when including categorical variables in a linear model, especially when including interaction terms.

When doing so, the betas are interpreted as main effects (i.e., the effect of the categorical variables at the average value of the continuous variable).

Despite the straightforwardness of this approach, several concerns, some of which are shared with the gamma coefficient, restrict the use of logistic regressions in metacognitive experiments. The first problem concerns the interpretability of the log-odds slope relative to an *ideal* log-odds slope. An ideal slope in a standard linear model is the identity slope. Rescaling JOLs to range from 0 to 1 results in an identity slope of 1; increases in JOLs are accompanied by equal increases in recall. An identity line of this sort is not possible in logit space because of the mapping to the latent space via the link function. The parameter estimate of the ideal slope depends on the dataset—namely, the proportion of values at the boundaries—and cannot be easily compared across different datasets.

Luckily, JOLs can be treated as likelihoods of recalling items on a later test. Consequently, a solution entails linking the (rescaled) JOLs themselves onto the same scale as the parameter coefficients through the following equation:

$$\log(p/1\text{-}p) \qquad 4$$

The identity slope of the logistic regression now becomes 1, and can be used as a proxy for perfect item-level calibration. This method only works under the constraints that the predictors are probabilities (e.g., likelihoods) and that no values in the dataset are on the boundary. The latter constraint is due to undefined values—which is the same culprit in the previous method. Of course, it is quite common for participants in

metamnemonic experiments to use the extreme endpoints of a scale (e.g., Dunlosky & Nelson, 1992).

An alternative approach avoids the issue of unidentifiable values. Rather than converting the predictors to the logit scale, a model's fitted values can be converted to the identity space and then refit with a linear model (for a similar concept in signal-detection theory, see Yonelinas, Dobbins, & Szymanski, 1996). This approach still requires that predictors be probability or likelihood judgments; the identity slope makes little sense otherwise. The refit slope of the fitted values is invariant to centering variables, a favorable property when analyzing complex models with many predictors. While this method eases interpretations, the refit slope should not be used for inferential testing. Besides, the slope was already inferentially tested in the linked model. Therefore, value of the refit slope should be used as a descriptive reference point and logistic regression used as the inferential testing model.

The second issue with using logistic regressions to analyze item-level data in metacognitive experiments is they neglect individual differences, a shared concern with the gamma coefficient. Entering each recall-JOL pair into a regression and assumes all observations come from the same distribution. Consequently, this implies that the relationship between the recall and JOLs variables is identical across scale-strata, neglecting individual usages of the JOL scale (e.g., participants who use the upper end versus participants who use the lower end). This can be remedied—to some extent—by a centering scheme (Raudenbush & Bryk, 2002; Snijders & Bosker, 1999) that accounts for within-subject and between-subject variability, separately.

A participant's JOLs can be reparameterized into two orthogonal variables,

within-subject differences and between-subject differences. The former is constructed by centering the participants JOLs on their mean JOL (i.e., subtracting the subject mean JOL from each of the subject's JOLs). The latter is constructed via a new variable that represents the difference between a participant's mean JOL rating and the sample mean JOL (i.e., subtracting the sample mean JOL from a subject's mean, creating a participant-level covariate for each of their observations). The within-subject centered variable represented the relationship between recall and JOL regardless of individual differences, representing the within-subject variability. The group-centered variable represents the relationship between recall and an individual's overall JOL, relative to the sample mean and thus representing between-subject variability. It addresses the question of whether participants who use the upper end of the JOL scale have greater recall than those who use the lower end of the scale. Additionally, an interaction term of the two centering variables can be added to the model. This third coefficient represents the relationship between a subject's item-level calibration (i.e., whether differences between JOLs are associated with meaningful differences in recall levels) and scale usage.

Generalized linear mixed models

Centering schemes help linear models tease a part different sources of variance but individual differences are still not fully accounted. Observations nested within participants are more likely to be similar than observations across participants. Such clustering violates the assumption of independence and is problematic when estimating within-subject coefficients (e.g., the relationship between recall and JOLs) because it systematically leads to over-estimating statistical significance (Laird & Ware, 1982).

Modeling the dependency, and thus modeling the structure of the data, provides a natural way of incorporating nested and/or crossed data.

Standard linear models only accommodate a single random effect and thus a single error term. Usually, subjects are modeled as the random effect (but see, Raaijmakers, 2003), and in this context, random effect means sampling subjects from a larger population from which you want to generalize to. Whereas the definition of a random effect is noncontroversial in the linear model, defining what constitutes as a random variable in mixed models is not always straightforward (Gelman, & Hill, 2006). Inheriting from the definition in the linear model, a variable is said to be fixed if the researcher is interested in the variables themselves (or the levels of the factor) and random if the researcher is interested in the underlying population (Searle, Casella, & McCulloch, 1992). Here, the research question distinguishes between fixed and random factors.

An alternative distinction between random and fixed factors is to model all *true* random factors as random and all others as fixed. The nature of fixed versus random factors is determined by the experimental design (Lamotte, 1983). In the JOL paradigm, a participant's recall and JOLs are not experimentally controlled and then, by design, should be modeled as random factors. Recall can be modeled as a random factor by including an additional error term associated with the intercept. This change results in the intercept representing the grand mean and the new error term representing a subject's deflections from that mean. The error term is assumed to be normally distributed, though can take on other distributions. JOLs are modeled as random variables in the same manner. Green and Tukey (1960) offer a slightly different

perspective on the differences between fixed and random factors. According to them, a fixed factor is one whereby levels within exhaust the population and a random factor where the levels are sampled from a larger population. This also implies the experimental design determines when factors are fixed or random.

Yet another alternative distinction between the two types of factors is to resort to random factors when variables are known to violate the assumptions of an ANOVA (Maxwell & Delaney, 2004). One of the most common violated assumptions of an ANOVA is variance-covariance structure of the levels within a factor[1]. Factors in an ANOVA take on a special variance-covariance structure called compound symmetry: All levels within a factor have the same variance (i.e., homogeneity of variances) and a constant covariance (Maxwell & Delaney, 2004). Violating this assumption leads to spurious effects or, depending on the correction applied, reductions in power (Mendoza, Toothaker, & Nicewander, 1974). Compound symmetry is often violated in cognitive experimental designs, especially designs that involve timing variables (e.g., response deadline paradigm, Reed, 1973). In the ANOVA paradigm, violations of compound symmetry can be dealt with by adjusting the F-test (e.g., Greenhouse & Geisser, 1959). By contrast, modeling the troublesome variable as random in a mixed model allows the researcher to specify the variance-covariance matrix of their choosing; mixed models make no assumptions about the variance-covariance structure. Hence, fixed and random factors are not distinguished by the purpose of generalizability but rather by their statistical requirements.

Another statistical perspective on modeling random factors is to only include

[1] In the point that follows, the assumption of compound symmetry apples to more complex ANOVA designs than just levels within a factor.

random components when statistically significant (Pinheiro & Bates, 2000). This approach borrows from the traditional model building perspective that emphasizes the use of caution when building increasingly more complex models. From this view, the researcher should prefer parsimony to complexity.

There is value in each of the different definitions of fixed and random factors. As such, we do not attempt to distinguish between any of them; doing so is beyond the scope of this paper. Instead, we consider them as additional motivation for utilizing mixed models for our analyses. First, mixed models accomplish the goal of generalizing our findings to a greater population. Second, mixed models provide a natural way of incorporating subject-specific responses (e.g., JOLs, RTs, etc.) as predictors, which are not exhaustive but drawn from a larger population of possible responses. And third, mixed models provide a method to test different variance-covariance structures. This last point is critical to testing a core assumption of cue utilization theory: relativity in judgments.

The variance-covariance structure that makes the least assumptions is a completely unstructured or unrestricted matrix. In terms of parameters, however, it is the most costly structure. Consider a random effect with three levels. The variance of each level is estimated (3 parameters) in addition to a covariance between each level (covariance of level-1 and level-2, level-1 and level-3, and level-2 and level-3). A total of 6 different parameters are estimated. This contrasts the number of parameters estimated under a compound symmetry matrix, which is two, single variance parameter and a single covariance parameter. There are several methods of testing the structure of random effects. The simplest method is to calculate Wald-type statistic by dividing the

parameter estimate by the standard error (see, e.g., Singer, 1998). This value is then compared to a critical value on a distribution table (*z*-table for random effects). However, this method often inaccurate (Pinheiro & Bates, 2000) and instead a deviance test is preferred, especially in generalized linear mixed models (Bolker, et al., 2009). A deviance test, commonly known as a log likelihood-ratio test, takes the difference between twice the negative log likelihood of two models and compares it to a chi-square distribution. The degrees of freedom is the difference in the number of parameters.

The Wald test and likelihood-ratio test are asymptotically equivalent in large samples. However, the Wald test is more biased in small samples (Johnston & DiNardo, 1997), leading to the endorsement of the likelihood-ratio test. Despite gaining acceptance, the likelihood-ratio test has unfavorable properties when testing values on the boundary of the parameter space (i.e., random effects close to zero). The test is ultraconservative when the random effect falls close to the boundary (Self & Liang, 1987). Under known cases, the true asymptotic distribution of random effects takes on a mixture of chi-square distributions (see, e.g., Self & Liang, 1987; Silvapulle & Silvapulle, 1995; Verbeke & Molenberghs, 2003). The popular statistical software, SAS, detects whether the tested random effects falls under one of the known cases and uses the correct chi-square mixture distribution[2]. We rely on this approach to test our random effects.

A further consideration must be made when using likelihood-ratio tests: the

[2] An alternative approach is a parametric bootstrap. The general procedure involves getting a critical likelihood-ratio value from a standard likelihood-ratio test. Then repeatedly generating data from the null model and recalculate the likelihood-ratio. The ratio of the number repetitions that produce a likelihood-ratio value greater than the critical likelihood-ratio value constitutes as a *p*-value: The probability of observing a dataset as extreme given a true null model (for a comparison of bootstrapping to an alternative approach, see, Crainiceanu & Ruppert, 2004).

method of parameter estimation. Two common techniques are maximum likelihood (ML) and restricted maximum likelihood (REML). The former simultaneously integrates over both fixed and random effects and is known to produce bias estimates of the random effects (Little, 2006; Pinheiro & Bates, 2000; Bolker et al., 2009). In contrast, REML first integrates out the fixed effects before estimating the random components. This is known to produce more accurate estimates of the random components and reduce bias in small samples (Little, 2006). However, because REML integrates out the fixed factors first, a likelihood-ratio test is only permissible when the fixed components between compared models remain the same (Pinheiro & Bates, 2000).

There are two classes of random effects that can be estimated, grouped random effects and residual random effects. These terms are made popular by SAS in order to distinguish between modeling random effects grouped on a factor (e.g., on participants; termed *G-side* random effect), and residual random effects (*R-side* random effects). For an illustration, imagine a three-level repeated factor that comprises 10 observations each (e.g., a participant offers a total of 30 data points). G-side random effects model the variance and covariance of the three factors, which is grouped on the participants. Ending the modeling endeavor at this point implies one very important assumption: The residuals are independent and identically distributed. Continuing the illustration, imagine that the 30 observations are JOLs that collected randomly throughout an experiment (e.g., randomization of stimuli and conditions to serial position). Cue utilization assumes that participants perceive cues—the sources of information used for JOLs—in a relative manner, especially intrinsic and mnemonic cues (Koriat, 1997).

Accordingly, this implies that, even after integrating out the effects of condition, a residual dependency remains. We test this assumption by modeling the residual effects.

In the hypothetical JOL example above, the *true* repeated measure is JOL and the observations are repeated across serial position. After integrating out the fixed and G-side random effects, the variance of the residuals is 1 and the covariance is 0, indicating uncorrelated residuals. A common covariance structure when modeling time-related variables is the first-order autoregressive structure (see, e.g., Box & Pierce, 1970). When using an autoregressive structure to model residuals, the variance remains unchanged but the correlation (covariance) between observations is now estimated rather than assumed to be zero. Observations one unit apart take on the correlation of rho, which decreases exponentially with each unit apart in time. This effectively models observations closer in time as more closely related than observations farther apart. After fitting our final model, we add an autoregressive structure to the residuals and test whether this constraint statistically supports our, and cue utilization's, theoretical prediction.

Testing fixed effects is more controversial than testing random effects in mixed models. Like random effects, a Wald-type test can be constructed to test single parameters[3]. This test is not controversial for Gaussian models with known degrees of freedom (Pinheiro & Bates, 2000), but calculating degrees of freedom is challenging when random effects take on more complex variance-covariance structures. The effective number of parameters used by a random factor lies somewhere between 1 and

[3] Also like random effects, fixed effects can be inferentially tested through log likelihood-ratio tests. However, whereas random effects likelihood-ratio tests are ultra conservative, fixed effects likelihood-ratio tests are ultra liberal (cf. Bolker et al., 2009). Therefore, we do not consider this a valid approach for testing fixed effects. Ultimately, a bootstrap is required to obtain accurate probability values for fixed effect likelihood-ratio tests.

N - 1 (i.e., all nested levels/observations within a random factor). Kenward and Rogers (1997) extended the Satterthwaite method, which calculates residual degrees of freedom for single parameter tests (Satterthwaite, 1941), to account for more complex nesting structures and for simultaneously testing multiple parameters (i.e., factors with more than two levels). Essentially, Kenward-Roger correction accounts for correlations among the levels nested under a random factor. This is a suitable approach for hypothesis testing (Bolker, 2009) and is the more conservative than applying traditional within-between degrees of freedom (Little, 2006). Therefore, we use Kenward-Roger corrected degrees of freedom when inferentially testing fixed effects.

Relationship between JOLs and recall

We first report analysis on JOLs and recall, separately. These analyses consist of group-level statistics, which collapse over item-specific factors (e.g., JOLs). We then investigate the relationship between JOLs and recall, analyzing participant calibration effects (e.g., average differences between JOLs and recall), gamma coefficients, and finally regressing recall on JOLs by a generalized linear mixed model. In particular, we investigate whether metacognitive resolution and calibration differ between strong and weak references to episodic memory.

JOLs

All judgments were collapsed into bins spanning 5 points and analyzed as such for the remaining analyses. The overall frequency of JOLs is displayed in Figure 2. Descriptively, the JOLs for reference items are more uniformly distributed than those given for control or transfer items. To help tease apart subject-level JOLs across the three item conditions, Figure 3 displays the subject-centered frequency of JOLs. Figures

displaying subject-centered variables facilitate the visual comparisons on factors manipulated within-subjects. In this case, for example, Figure 3 shows that the JOL distributions for control and transfer items are shifted towards lower values. It furthermore shows that there does not appear to be any major signs of polarization. This is the case when JOLs are made immediately after studying a word pair (Dunlosky & Nelson, 1994; Kimball & Metcalf, 2003). Polarization is much more prevalent when delaying JOLs, and is taken as evidence that participants are covertly retrieving the targets, with successfully retrieved targets being given high JOLs and unsuccessfully retrieved targets being given low JOLs.

A mixed-factor[4] analysis of variance was used to investigate differences in mean JOLs as a function of the item type (manipulated within subjects: reference; control; transfer) and episodic retrieval (manipulated between subjects: likely retrieval; unlikely retrieval). There was a main effect of item, $F(2, 128) = 108.5$, $MSE = 50.6$, $p < 0.0001$, $\eta_p^2 = 0.63$, $\omega_G^2 = 0.12$[5]. As can be seen in Figure 4 and Figure 5, JOLs for reference items were much higher than for control or transfer items. Further simple comparisons were tested to examine the hypothesis that explicitly referencing episodic memory would lead to a reduction in the illusion of competency. Regarding the solicitation of JOLs, the only procedural difference between reference and control items was the

[4] The term mixed-factor should not be confused with the term mixed model. A mixed-factor ANOVA indicates that at least one factor was manipulated between subjects and at least one factor was manipulated within subjects (e.g., mixed designs, split/strip-plot designs). This confusion is perpetuated in the literature by inappropriately using the term "mixed ANOVA". A *mixed-effects* ANOVA refers to an ANOVA with a random effects, which is different than a mixed-factor ANOVA. In the context of an ANOVA, the term *mixed* should not appear without clarification.

[5] We report two measures of effect size in the aggregated data inferentials, partial-eta squared and generalized-omega squared. Partial-eta squared measures the variance explained by the factor of interest (Cohen, 1973). However, it is not directly comparable across different experimental designs (Olejnik & Algina, 2003). As such, we report generalized-omega squared, which better accounts for the variance contributions of each factor in the omnibus model and is more comparable across different designs (Kellermann et al., 2013). Additionally, we report *F*-tests for all simple comparisons as to maintain consistency and comparability across the reported effects.

temporal occurrence in the study list. Reference items occurred in the beginning of the study list and control items occurred later on in the list and among transfer items. Thus, the comparison between reference item JOLs and control item JOLs would test whether the effect of referencing episodic memory would carry over to new items, items for which episodic retrieval was not experimentally required. This difference was reliable, $F(1, 64) = 112.92$, $MSE = 119$, $p < 0.0001$, $\eta_p^2 = 0.63$, $\omega_G^2 = 0.11$ and confirms that the illusion of competency, as measured by overall JOL level, was reduced for items in which no episodic retrieval was required. This finding lends support to the hypothesis that episodic retrieval produces a mnemonic cue used to assess global memorability.

Additional planned comparisons were conducted to test whether requiring participants to consider episodically related information would result in item-specific reductions in confidence levels. Here, and throughout the paper where we report aggregated data inferentials, we use the term *item-specific* loosely because we are, in fact, measuring the reductions in confidence on a group of items (transfer items vs. control; transfer items vs. reference). However, it is item-specific in the sense that reductions are specific to items in which episodic retrieval was enforced. Transfer items were found to have lower JOLs than for either control items, $F(1, 64) = 12.92$, $MSE = 39$, $p = 0.0006$, $\eta_p^2 = 0.16$, $\omega_G^2 = 0.0038$, or reference items, $F(1, 64) = 132.06$, $MSE = 144$, $p < 0.0001$, $\eta_p^2 = 0.67$, $\omega_G^2 = 0.1485$. This shows that, when making immediate JOLs, incorporating episodic information reduces confidence levels.

The main effect of episodic awareness was not significant, nor was the interaction of episodic awareness (Fs < 0.18). The lack of an interaction is somewhat surprising. It was predicted that episodic awareness, brought about by covert retrieval,

differs between cue-only reference items and cue-target reference items. Consequently, it was predicted that covert retrieval would lead to a greater reduction in confidence judgments because participants would realize the fallibility of memory. These data, however, suggest that, on the average, both cue-only and cue-target reference items affected transfer items similarly.

Recall

Participants' average recall was analyzed through a 3 (item type: reference; control; transfer) x 2 (episodic retrieval: likely; unlikely) mixed-factor ANOVA; the first factor, item type, was manipulated within subjects, and the second factor, episodic retrieval, was manipulated between subjects. There was a significant main effect of item type, $F(2, 128) = 26.34$, $MSE = 85.7$, $p < 0.0001$, $\eta_p^2 = 0.30$, $\omega_G^2 = 0.0544$, no main effect of episodic awareness, $p = 0.3254$, $\eta_p^2 = 0.02$, $\omega_G^2 = 0.0014$, and a significant interaction qualifying the main effects, $F(2, 128) = 4.16$, $MSE = 85.7$, $p = 0.0178$, $\eta_p^2 = 0.06$, $\omega_G^2 = 0.007$. Simple contrasts revealed a surprising effect. As seen in Figure 6 and Figure 7, reference item recall did not differ when presented as cue-only or cue-target during the judgment phase for transfer items. This is surprising because the double presentation of an intact word pair should lead to greater recall. It did not, $p = 0.8693$, $\eta_p^2 = 0.0$, $\omega_G^2 = 0.0$. As for the remaining contrasts, transfer item recall was marginally higher when the episodic retrieval was strong than when it was weak, $F(1, 64) = 2.93$, $MSE = 318$, $p = 0.0916$, $\eta_p^2 = 0.05$, $\omega_G^2 = 0.0308$. One potential reason for this difference is that presenting the reference item in a cue-target format may interfere with the encoding of the transfer item and thus cause lower recall. Finally, control items did not differ as a function of episodic retrieval, $p = 0.1535$ $\eta_p^2 = 0.03$, $\omega_G^2 = 0.0191$.

Calibration

To investigate mean calibration effects, we took participants' average JOL ratings and subtracted their average recall (see Figure 8 and Figure 9 for means). These values were used in a mixed-factor ANOVA with item type manipulated as a within-subject factor (reference, control, transfer) and episodic retrieval manipulated as a between factor (likely, unlikely). The omnibus ANOVA revealed a main effect of item type, $F(2, 128) = 5.6$, $MSE = 1.81$, $p = 0.0046$, $\eta_p^2 = 0.08$, $\omega_G^2 = 0.0092$. Transfer items showed less overconfidence than either control items, $F(1, 64) = 4.15$, $MSE = 0.899$, $p = 0.0458$, $\eta_p^2 = 0.06$, $\omega_G^2 = 0.0143$, and reference items, $F(1, 64) = 7.55$, $MSE = 2.66$, $p = 0.0078$, $\eta_p^2 = 0.11$, $\omega_G^2 = 0.0025$. Furthermore, control items showed marginally less overconfidence than the reference items, $F(1, 64) = 3.51$, $MSE = 0.14$, $p = 0.0654$, $\eta_p^2 = 0.05$, $\omega_G^2 = 0.0036$. Neither the main effect of episodic retrieval, nor the interaction was reliable (Ps > 0.14). These data suggest that the agenda of reducing the illusion of competency by considering previously studied items was accomplished. However, the hypothesis that cue-only reference items would promote better metamemory calibration than cue-target reference items was not supported. Because the subjective experience of covert retrieval is not possible with cue-target reference items, this implies that a different mechanism is responsible for this result—at least in the weak episodic condition.

Gamma correlation between JOLs and recall

The relationship between JOLs and recall was investigated first through a gamma correlation. Gamma was calculated for each subject, with 9 subjects having no variability in either recall or JOLs for at least one condition ($n = 5$ for the strong

episodic retrieval condition and $n = 4$ for the weak episodic retrieval condition). The group means are shown in Figure 10 and the subject-centered means are shown in Figure 11. The gamma correlations were then analyzed through a 3 (item type: reference; control; transfer) x 2 (episodic retrieval: likely; unlikely) mixed-factor ANOVA with item type as a within-subject factor and episodic retrieval as a between-subject factor. There was a marginal effect of item type, $F(2, 110) = 2.86$, $MSE = 0.12$, $p = 0.0616$, , $\eta_p^2 = 0.05$, $\omega_G^2 = 0.0179$, such that numerically, gamma was higher for reference items than for control items, and gamma for control items was higher than for transfer items. Neither the main effect of episodic retrieval nor the interaction of episodic retrieval and item type was significant (Fs < 1.25).

A planned 2 (item type: control; transfer) x 2 (episodic retrieval: strong; weak) mixed-factor ANOVA was used to investigate whether insight into episodic memory had an effect on resolution—whether participants could better discriminate between recallable and nonrecallable words. Neither of the main effects nor the interaction reached significant levels (Fs < 1.83). Consequently, providing access to episodic memory through a previously studied word did not increase JOL resolution compared to control items, regardless of how episodic memory was referenced, strongly or weakly.

GLMM: Regressing recall on transfer item JOLs

As previously described, random effects in mixed effects models can be modeled from different perspectives (e.g., population-level inferences). We take the approach of modeling random effects as a requirement of the experimental design. As such, we model experimentally controlled within-subject factors as random effects and *true* subject-level factors as random effects. Truly random factors, for example, the

JOLs provided by subjects, are random and not experimentally controlled. Therefore, we treated subject provided variables as random. The experimentally controlled within-subject factors are modeled as random on the bases of the variance-covariance of their levels. This implies that the levels are not entirely independent of one another (within the subject), whereas between-subject factors are, by design, independent. Furthermore, this also implies that observations within each level are clustered by subjects. We test these assumptions by enforcing simple variance-covariance structures or allowing for more ones on the data. Notwithstanding estimation techniques, using the simple structure, compound-symmetry, increases the comparability of a mixed model to an ANOVA. Therefore, we use this simple structure when possible.

The primary focus is on the relationship between recall and JOLs for the reference items: Does access to episodic memory increase the ability to distinguish between recallable and nonrecallable words? The gamma results reported above indicate that it is unlikely. As such, we model the differences between episodic retrieval strength (strong vs. weak) on transfer items. We begin the modeling approach with a baseline unconditioned model containing no random or fixed effects (Model 1). We used maximum likelihood to estimate parameters during the model selection phase. We added a random intercept representing individual differences on recall (Model 2). Table 1 shows the model fit increases substantially when adding individual differences. Next, we added the fixed effects of JOLs (Model 3). We centered the JOLs on the subjects (within subject JOLs), so that the slopes represent the item-level calibrations for participants, regardless of how they use the confidence scale. Additionally, we added an orthogonal variable centered on the group mean. This variable represents the differences

between participants (between subject JOLs). The variable addresses whether recall is better for those who tend to use the higher end of the scale than those who tend to use the lower end of the scale. The interaction of the two variables represents differences in item-level calibration as a function of scale usage. This improved the fit of the model.

Importantly, we modeled the within-subject JOLs as a random factor (Model 4). As can be seen from Table 1, this did not improve model fit. This indicates that participants did not differ substantially in their item-level calibrations. Regardless, on the basis of the experimental design, we retain the factor in the model. We added the remaining fixed factors to the model, episodic retrieval and the interactions among the fixed effects (Model 5 & 6). Neither these additions improved model fit, but because interpretations of the results do not change and because we took the experimental design approach to model building, we report the most complex model.

We refit Model 6 using restricted maximum likelihood to get the best estimates for the random effects. While this is not crucial because the random effect of within subject JOLs did not approach significance, we believe it is best practice to do so when there is interest in random effects. The final model estimates can be seen in Table 2. Only one effect was reliable, within-subject JOLs, $F(1, 31.96) = 40.63$, $p < 0.0001$, which indicates that JOLs are predictive of recall. Figure 12 shows the fixed effect of within-subject JOL on recall at the sample mean JOL (e.g., the main effect of within-subject JOLs). At the average JOL, a participant has a 25% chance of recalling a given item and that chance increases by a 2% with each increase in JOL. Item-level calibration did not differ as a function of episodic retrieval, supporting the gamma correlation results. Collectively, these data indicate that referring back to previously

encountered information does not sharpen one's metacognitive ability so that items are better separated into mastered and nonmastered categories. Instead, utilizing episodic memory lowers the over illusion of competency—regardless if covert retrieval is possible (cue-only reference item) or not (cue-target reference item).

LMM: Regressing transfer item JOLs on reference item JOLs

Thus far, the aggregate data results indicate that episodic awareness need not occur strongly. Simply reprocessing episodically related information is sufficient to influence metacognition, namely, to lower confidence. We take a closer look at what factors, if any, influence immediate judgments at the item level. Specifically, we look at whether previous JOLs for reference item have predictive power on transfer item JOLs. Reference items received immediate JOLs at the beginning of the study list. With immediate JOLs serving as a proxy for processing fluency (Koriat & Ma'ayan, 2005), we investigate whether processing fluency impacts transfer item JOLs on an item-by-item basis.

We begin model building in the same manner as above, estimating parameters via maximum likelihood, and starting with an unconditioned model that has no effects (see Table 3). This serves as our baseline model (Model 1). We added a random intercept in Model 2, which significantly improved the fit of the model. This indicates that participants use the JOL scale in different ways—individual differences. Next, we added two fixed factors, within-subject centered reference item JOLs and between-subject centered JOLs (Model 3). These were calculated in the same manner as before. A correlation with the within-subject JOLs is interpreted such that item-specific features of the reference items contribute to the ratings of transfer items. Between-subject JOLs

are interpreted as differences in scale usage of the reference items. An effect of between-subject JOLs would indicate that participants use the JOL scale similarly across reference items and transfer items. For example, those who generally provide high JOLs for reference items also provide high JOLs for transfer items. The interaction of the two variables is equally interesting: Are participants who more easily process information reference items (e.g., higher immediate JOLs for reference items) at a greater risk for item-level bias of transfer items? As before, we included the reference item JOLs as random effects (Model 4) and this improved the model fit. This also provides evidence that individual difference exist on the effect that reference item JOLs have on transfer items. We added the remaining fixed factor, episodic retrieval (Model 5), and the interaction of all the fixed factors (Model 6).

Finally, we tested an independent covariance structure between the intercept and the within-subject reference items. To do so, we fit Model 7 using restricted maximum likelihood (to better estimate random effects), and tested whether the covariance of the intercept and within-subject reference item JOLs was zero (e.g., independence). The results show that Model 6 with an unstructured covariance matrix fit the data better and therefore we retain the covariance of intercept and within-subject reference item JOLs. Model 6 is our final model and we explore the fixed effects next.

Table 4 displays the coefficients for the final model. The two types of reference item JOLs had a significant correlation with transfer item JOLs (within-subject: $F(1, 57.91) = 9.92$, $p = 0.0026$; between-subject: $F(1, 62) = 133.1$, $p < 0.0001$). These effects are qualified by two higher order interactions. First, there was a significant interaction of episodic retrieval and within-subject JOLs, $F(1, 57.91) = 6.66$, $p =$

0.0124, such that transfer item JOLs were more influenced by reference item JOLs when the reference item had been presented as cue-only than when presented as cue-target. However, there was a three-way interaction of episodic retrieval, within-subject JOLs, and between-subject JOLs, $F(1, 76.4) = 5.03$, $p = 0.0278$. In order to display the nature of the interaction, we plotted the fitted values of the relationship between episodic retrieval and within-subject JOLs and three different between-subject intervals, one standard deviation below the mean JOL ratings for reference items, at the mean JOL ratings, and one standard deviation above the mean JOL ratings. As seen in Figure 13, participants who gave lower JOLs to the reference items were less impacted by their re-presentation at the time of transfer item JOLs. This was the case regardless of whether the reference item was presented as cue-target or as cue-only. However, participants were more influenced by cue-only reference items when the reference items themselves received higher JOLs.

If immediate JOLs reflect the processing fluency of word pairs, then this effect is somewhat odd. It would be expected that previously studied intact word pairs would be processed more fluently than cue-only word pairs. Presumably, cue-only word pairs would promote covert retrieval and thereby slow down the processing fluency. The end result would be lower JOLs (see Benjamin, et al., 1998). These data suggest otherwise. Alternatively, immediate JOLs for reference items may have been made on the basis of cue familiarity (Koriat & Levy-Sadot, 2001). Then, later in the experiment during the time of transfer item JOLs, presenting the familiar cue without any interference from the target would lead to a greater sense of familiarity. That is to say that the target interferes with the recognition process of the cue. To the extent that a reference item's

target word interferes with the familiarity of its cue, we should expect to see the pattern we observed.

LMMR Regressing transfer item JOLs on reference item recall

Given the previous results, that reference item JOLs influence transfer item JOLs, it seems that metacognition is highly malleable, although this may not necessarily be fatal. If the ability to retrieve previous information is diagnostic on the future retrievability of related information, then that metamnemonic cue has predictive power. For instance, if I know I cannot recently studied mathematical proofs, it is unlikely that I will be able to retrieve newly learned equations in the future. To that effect, episodic retrieval can be a useful mechanism. We test whether participants rely on such a mechanism by analyzing transfer item JOLs as a function of final recall of the reference item. We use final recall as a proxy of successful retrieval during the time of JOLs. If episodic retrieval influences transfer item JOLs it should only occur for cue-only reference items, where covert retrieval is possible.

Because reference item recall is naturally a random factor, we treat it as such. We continue with using a mixed model and use the same unconditional model as before (Model 1). However, rather than adding a random intercept, we model the random effect of reference item recall, setting it as a classification variable. As such, the random component has 3 variances, one for unrecalled items, one for recalled items, and one for the covariance. This mode fit substantially better than did the unconditional model (see Table 5). We then added the fixed effect of reference item recall in Model 3, the fixed effect of episodic retrieval in Model 4, and the interaction in Model 5. Lastly, we tested homogeneity of variance between the two levels of reference item recall. This model fit

marginally poorer than the unstructured covariance. Because this test is ultraconservative, and because the two models do not change the interpretation of the results, we decided to endorse the more complex model, Model 5.

There was a main effect of reference item recall, $F(1, 64.6) = 21.79$, $p < 0.0001$, that was qualified by the two-way interaction, $F(1, 64.6) = 17.26$, $p < 0.0001$. Figure 14 shows the nature of the interaction and the parameter estimates can be found in Table 6. When participants are faced with a cue-only reference item, a condition that encourages episodic retrieval, they are much more likely to offer high JOLs when the reference item is eventually recalled on a final test. However, this effect does not occur in the condition that is does not promote episodic retrieval, when the reference item is presented as cue-target. Thus, as predicted, referencing episodic memory influences metacognition.

LMM: Regressing JOLs on serial position

Cue utilization theory postulates that JOLs are comparative in nature (Koriat, 1997). However, there are few studies that explore this dynamic (but see Dunlosky & Matvey, 2001; Castel, 2008). In the current experiment, explicitly instructed participants to consider episodically related information, which we believe occurs naturally in the course of learning. In this sense, we encouraged the relativity of JOLs in the latter portion of the experiments, when participants made use of reference items. However, in the beginning of the experiment, when judging the reference items themselves, participants were free to use any strategy when making JOLs. Accordingly, there should be some correlation among JOLs if participants made judgments comparatively. We test this in a mixed model by regressing reference item JOLs on

serial position. Then, once capturing the temporal relationship between we model the residuals and see whether there are any unaccounted correlations. Specifically, we employ a first-order autoregressive covariance structure where JOLs closer in time are more closely related than those further apart in time. To the extant that JOLs are related to each other, we should observe a residual correlation.

Reference item JOLs were analyzed in a mixed effects model with a fixed effect of serial position and two random effects, a random intercept and a random effect of serial position. Because the judgment phase of reference items did not differ procedurally between the two between-subject groups (episodic retrieval: likely retrieval; unlikely retrieval), we did not include that factor in the model. It is expected that, to a degree, JOLs decrease in magnitude over time. Participants might realize the increasing difficulty in memorizing word pairs as the study list increases. Consistent with this idea, there was a main effect of serial position, $F(1, 64.99) = 92.35$, $p < 0.0001$, such that JOLs decreased by an average of 0.7913 points with each serial position. The more important question is whether the residuals are correlated. In this case, correlated residuals would suggest a temporal relationship of the JOL ratings. To test this, we refit the serial position model with an R-side random effect, a first-order autocorrelation structure. Then we tested whether the additional parameter, rho, significantly improved the model fit. A likelihood-ratio test confirmed a model improvement, $x^2(1) = 22.56$, $p < 0.0001$, $\rho = 0.1187$, indicating a temporal relationship among JOLs, and thus, supporting a prediction of cue utilization theory.

Global predictions

Transfer item JOLs were affected by reference items in both the episodic

retrieval conditions. In the condition likely to provoke episodic retrieval, transfer items were largely impacted by the characteristics of the reference item. By contrast, in the condition unlikely to provoke episodic retrieval, transfer items were less impacted on an item-by-item basis, but were nonetheless influenced. To understand whether participants were aware of these differences, we asked them at the end of the study phase to indicate how accurate they thought the JOLs were. Specifically, we solicited accuracy judgments for control item JOLs and for transfer item JOLs. Given the empirical pattern, we expect the likely-retrieval condition to yield a greater difference between transfer and control items than in the unlikely-retrieval condition.

The global accuracy judgments were analyzed in a 2 (item type: transfer; control) x 2 (episodic retrieval: likely-retrieval; unlikely-retrieval) mixed-factor ANOVA. The analysis revealed a main effect of item type $F(1, 64) = 8.21$, $MSE = 337$, $p = 0.0056$, $\eta_p^2 = 0.11$, $\omega_G^2 = 0.03$, transfer item JOLs were thought to be less accurate than items that did not require consideration of a previously studied item (i.e., reference items and control items). No other effects reached significance (Fs < .10). One likely cause for the lack of the expected interaction may have been because of the relativity of JOLs and that episodic retrieval was manipulated between subjects. Participants have no reference point other than the control items, where the transfer items are perceived to yield relatively lower JOLs. In Experiment 2 we explore the subjective relativity further by manipulating episodic retrieval within subjects

Discussion

There are several key findings in Experiment 1. First, JOLs for reference items were much higher than for control items or transfer items. We believe the reason is

because participants are naive early on in the experiment. Then once engaged in making JOLs in the presence of previously studied items, participants become aware of the fallibility of their memory. As a result, they decrease the levels of their JOLs. This is not necessarily the result of covert retrieval prior to making immediate JOLs, as was hypothesized. Instead, it is a result of a complex interaction of item-specific factors that occur when making JOLs in the presence of a reference item. For instance, that reference item JOLs were predictive of transfer item JOLs suggests that the mere fluency or familiarity of previous items is enough to add source confusion when making JOLs. Nevertheless, the reduction in JOLs carried over to control item JOLs and this happened in both conditions between-subject conditions. These global effects can be desirable in that they reduce the illusion of competency overall (Koriat & Bjork, 2006b). This promotes conditions that increase the chances of selecting more items to restudy (Metcalf, 2002).

Considering previously studied information when making immediate JOLs did not improve the ability to discriminate between recallable and nonrecallable words. However, the gamma correlation was no worse than for the control items. In general, the sources of information used when making immediate JOLs does not lead to better metacognitive resolution, and this is true even when referring back to episodically related information. The important finding is that immediate JOLs are influenced by item-specific traces of episodic memory. For instance, retrieving easily accessible information during the time of making immediate JOLs will cause inaccurately inflated judgments. Thus, the extent that previously considered material is an indicator of successful retrieval will determine the accuracy of immediate JOLs. This is encouraging

in that this processes is likely to occur in everyday learning situations. In the lab, and as in the experiments reported here, we control the presentation of which item should be considered—chosen at random. However, in common real situations, the retrieved information is likely to share semantic and contextual relationships. Therefore, retrieval of this nature is much more likely to increase the reliability of immediate JOLs.

Experiment 2

The results from Experiment 1 are compelling. First, episodic information can be incorporated into formulation of immediate JOLs and this alters the metacognitive judgments made on an item-by-item basis. This awareness occurs regardless of whether one engages in covert retrieval and spreads to new items in which no episodic information is provided. However, there is a greater propensity of item effects when episodic retrieval is likely to occur. This underscores the potency that retrieval has on generating mnemonic cues to be used for JOLs. Unfortunately, participants did not realize this effect because the global accuracy judgments did not reflect differences between the episodic retrieval conditions. To investigate further, we manipulated episodic retrievability within subjects in Experiment 2. If JOLs are truly made in a relative manner, as the data suggest from Experiment 1, then this should result in 1) similar patterns as previously found regarding JOL magnitude, resolution, and item-by-item effects, and 2) differences in the participants' subjective global accuracy judgments. The latter point underscores the relativity of JOLs. Participants believed that referring back to previous information affected their JOLs, but this did not differ in the similar magnitude that JOLs were actually affected. That is, there were greater item-specific effects in JOLs when episodic retrieval was likely but not when episodic was

unlikely. The subjective beliefs did not reflect this pattern. Manipulating episodic retrievability within subjects is predicted to encourage the awareness of this pattern to participants. However, this would only occur if JOLs were made in a relative manner. Our goal, then, is to replicate the general pattern found in Experiment 1 using a within-subject design, and to test whether participants are aware of the influence of covert retrieval on immediate JOLs.

We made three key methodological changes in Experiment 2. First, because recall of the reference item predicted transfer item JOLs, we implemented a design feature to collect additional information about the retrievability of the reference item. Specifically, we delayed the JOLs of the reference items. As can be seen in Figure 16, reference item JOLs were delayed by 8 intervening items. This displaces the reference item from memory at time of JOL. According to leading theories of the delayed-JOL effect (e.g., Kimball & Metcalf, 2003), covert retrieval occurs prior to providing JOLs. Therefore, we predict to see the reference item JOLs correlate with transfer item JOLs when the reference item is later presented as cue-only. The second study block shown in Figure 16 demonstrates the second design change, a within-subject manipulation of the likelihood of episodic retrieval (i.e., retrieving the reference item target). This was in part motivated by the lack of a difference in participants' subjective ratings in JOL accuracy despite the fact that increasing the likelihood of covert retrieval of the reference item having a large influence on transfer item JOLs. The last change was the removal of control items. Experiment 1 demonstrated that a general underconfidence was brought about by considering related information from episodic memory during transfer item immediate JOLs. As such, the control items were eliminated from the

design so that a more key finding could be explored, the retrieval effects on immediate JOLs.

Method

Participants

Ninety-seven University of Oklahoma undergraduate students enrolled in the introductory psychology course participated for partial course credit.

Design

Item type (*reference* and *transfer*) and episodic retrieval (likely, unlikely) were manipulated within subjects. Participants made delayed cue-only JOLs on the first half of the studied word pairs (i.e., reference items). This occurred through bock randomization, in blocks of 8 word pairs. Judgments for the other half of the remaining items (i.e., transfer items) were made in the presence of a previously studied item (i.e., reference items) from the first half of the study list (see Figure 16). The presentation format of the reference item at the time of transfer item JOLs was manipulated; half of the transfer items were presented with cue-only reference items and the other half with cue-target reference items.

Materials

The stimuli comprised the same 180 four-letter common nouns from Experiment 1. Word pairs were randomly paired anew for each participant. Eighty word pairs were selected from that set; half of which served as reference items and the other half as transfer items. All 80 word pairs were studied and tested.

Apparatus

The same method of delivery was used as in Experiment 1.

Procedure

The experiment consisted of an instruction phase, followed by a single study/cued-recall test session. After participants consented to the experiment, the computer displayed an interactive instruction set, in which the participants could navigate through and whereby the computer revealed the process of the experiment in a step-by-step nature, including examples of each task the participants would encounter. The instructions began with a general description of the study list—that participants were to study 80 cue-target word pairs, each presented for 4-seconds, and then make judgments on their confidence in recalling the target word when provided the cue on a subsequent memory test (e.g., memory judgments).

As in Experiment 1, the instructions elaborated the exact nature of the memory judgments. Participants were told some memory judgments would be made in the presence of additional information. Participants were first introduced to the unlikely-retrieval instructions, "Consider the following previously studied word pair: cue - target when judging your confidence in recalling the missing target BELOW: cue - ____". Following those instructions, the computer presented instructions for the likely-retrieval condition, "Consider the following previously studied word pair: cue - _____ when judging your confidence in recalling the missing target BELOW: cue - ____". In addition, participants were informed that on some occasions there would be a delay between studying a word pair and judging the word pair. The same scale from Experiment 1 was used. Following the delivery of instructions, participants were posed with an open-ended question requiring a description of the two *different* memory judgment tasks, the delayed cue-only judgment task and the immediate judgment task

when required to consider a previously studied word pair. No participants were excluded for failure to understand instructions.

The experiment began by presenting each word pair at the top of the screen for 4-seconds. During the first 40 word pairs, after studying a block of 8 pairs, the computer reselected items from that block to be given delayed cue-only JOLs. This selection was randomized. After studying and judging the first 40 word pairs, participants engaged in studying and judging the transfer items. As in Experiment 1, transfer items were given immediate JOLs such that a prompt soliciting JOLs immediately followed each studied word pair. All judgments were self-paced.

The computer posed three questions to each participant following the last studied item and before the final cued-recall test began. In the same manner as in Experiment 1, we asked participants to indicate their accuracy in the three different types of ratings, reference items (which asked the accuracy of judgments when participants did not have to consider additional information), likely-retrieval transfer items (which asked the accuracy of judgments when considering a cue-only word pair), and unlikely-retrieval transfer items (which asked the accuracy of judgments when considering a cue-target word pair). The same scale was used as before and the questions were self-paced. The cued-recall test required participants to type the target studied with the provided cue. Participants continued testing each studied word pair at their own pace.

Results

JOLs

As in Experiment 1, all judgments were collapsed into bins spanning 5 points

and analyzed as such for the remaining analyses. The overall frequency of JOLs is displayed in Figure 17, and the within-subject centered frequency displayed in Figure 18. While difficult to see in the aggregated frequencies (Figure 17), the subject-centered JOLs clearly show that the reference items received polarized judgments. This was expected and suggests that participants judged items on the basis of covert retrieval, with successfully retrieved items receiving high JOLs and unsuccessfully retrieved items receiving low JOLs (Dunlosky & Nelson, 1994; Kimball & Metcalf, 2003). As such, this methodology—delaying reference item JOLs—allows for reference item JOLs to be used as a proxy for retrievability during transfer item JOLs. This, along with reference item final recall, will be used to test whether covert episodic influences transfer item JOLs.

A repeated measures ANOVA used to investigate differences in mean JOLs as a function of the item type (reference, transfer) and episodic retrieval of the reference item (likely, unlikely). There was a main effect of item, $F(1, 96) = 8.92$, $MSE = 591$, $p = 0.0036$, $\eta_p^2 = 0.08$, $\omega_G^2 = 0.01$, and episodic retrieval of the reference item, $F(1, 96) = 12.51$, $MSE = 43$, $p = 0.0006$, $\eta_p^2 = 0.12$, $\omega_G^2 = 0.016$. The interaction was not reliable ($F = 1.27$). Figures 19 and 20 show that episodic retrieval reduces JOLs, however, the pertinent comparison is the simple effect for transfer items. Indeed, that effect is reliable, $F(1, 96) = 17.98$, $MSE = 26$, $p < 0.0001$, $\eta_p^2 = 0.16$, $\omega_G^2 = 0.023$, indicating that, just like in Experiment 1, covert retrieval reduces the illusion of competency, a measured by JOL levels.

Recall

Participants' average recall was analyzed through a 2 (item type: reference;

transfer) x 2 (episodic retrieval of the reference item: likely; unlikely) repeated measures ANOVA. There was a significant main effect of item type, $F(1, 96) = 190.6$, $MSE = 0.02$, $p < 0.0001$, $\eta_p^2 = 0.66$, $\omega_G^2 = 0.241$, and episodic retrieval of the reference item, $F(1, 96) = 9.37$, $MSE = 0.01$, $p = 0.0029$, $\eta_p^2 = 0.08$, $\omega_G^2 = 0.01$. Which was qualified by an interaction, $F(2, 96) = 8.58$, $MSE = 0.01$, $p = 0.0043$, $\eta_p^2 = 0.08$, $\omega_G^2 = 0.01$. The nature of this interaction is described in Figures 21 and 22. Reference items shown as intact word pairs during the time of transfer item JOLs received a memorial advantage, attributed to an additional study opportunity. Thus, unlike in Experiment 1 where intact reference items were not recalled better, the expected recall pattern emerged in Experiment 2. Transfer item recall did not differ as a function of the likelihood to engage in episodic retrieval of the reference item ($F < .001$).

Calibration

In a similar fashion as before, we took participants' average JOL ratings and subtracted their average recall to study calibration effects (see Figure 23 and Figure 24 for means). These values were used in a repeated measures ANOVA. However, we only investigate the differences for transfer items. The calibration results for the reference items are contaminated by their double exposure, which was not accounted for when participants made their JOLs for those items. As such, it is impossible to interpret the calibration results for reference items. Instead, we continue to focus on the main comparisons in this experiment, differences in transfer item JOLs when covert retrieval of the reference item is likely versus when it is not likely. This difference was reliable for calibration effects, $F(1, 96) = 5.21$, $MSE = 0.01$, $p = 0.0247$, $\eta_p^2 = 0.05$, $\omega_G^2 = 0.0132$, supporting the previous findings that covert retrieval of the reference items

influences overconfidence. However, Experiment 1 indicates that these results are conditional on the retrievability of the reference items. When covert retrieval of the reference item is likely and successful, transfer item JOLs receive a boost. Likewise, when covert retrieval of the reference items is likely but unsuccessful, transfer items JOLs are reduced. We explore these effects in later sections.

Gamma correlation between JOLs and recall

Gamma was calculated for each subject, and analyzed in a repeated measures ANOVA. The same caveated applies when interpreting reference item gamma as does when interpreting the calibration results. The double exposure of reference items was not accounted for at the time of JOLs. Thus, we only report that gamma was much higher for reference items than for transfer items ($p < 0.0001$). This is expected because reference items were given delayed cue-only JOLs. Transfer item gammas did not differ as a function episodic retrieval ($F = 0.15$; see Figures 25 & 26). Participants could no better discriminate between recallable and nonrecallable words when they made judgments in the face of a cue-only reference item or a cue-target reference item. Covert retrieval of episodically related information has little effect on metamemory resolution. We suspect that this is because the relationship between what is being retrieved (the reference item) and what is currently being judged (transfer item) is random—the retrieved information is not diagnostic of eventual recall. That is, the reference item is randomly selected among the previously studied items and thus the episodic relationship is random. In more realistic situations, the relationship between episodic retrieval and metacognition of immediately judged information are likely nonrandom and have meaningful relationships. However, before we investigate more meaningful

relationships between episodically related information, we first need to determine whether episodic retrieval has any role in metacognitive judgments. According to the item-level analyses in Experiment 1, covert retrieval of related information is quite influential on immediate JOLs. We turn to item-level analyses for Experiment 2 next.

GLMM: Regressing recall on transfer item JOLs

We continue on with the modeling approach described in Experiment 1, first focusing on the item-level calibration effects and then moving on to the association between reference items JOLs and transfer item JOLs. Given the gamma results, it is unlikely that we will observe any differences in the predictive power of transfer item JOLs on recall as a function of episodic retrieval of the reference item. Table 7 shows the modeling approach taken for these data. An unconditioned model was improved upon by adding the random effect of the within-subject manipulated factor of episodic retrieval (Model 2). The fixed effects of JOLs were then added, followed by the random effect of within-subject centered JOLs (Model 4). Much the same as in Experiment 1, this random effect did not improve the model. This indicates that participants did not differ substantially in their item-level calibrations. Regardless, on the basis of the experimental design, we retain the random factor in the model. We added the remaining fixed factors to the model, episodic retrieval and the interactions among the fixed effects (Model 5 & 6). Neither these additions improved model fit. The interpretations between these models do not change and because we are interested in the differences between episodic retrieval, we endorsed the more complex model, as in Experiment 1.

We refit Model 6 using a penalized restricted maximum likelihood to get better estimates of the random effects. The model parameters are listed in Table 8 and Figure

27 plots the fitted values for the effect of the within subject JOLs at the sample mean JOL. Two main effects were reliable in Model 6: within-subject JOLs, $F(1, 68.52) = 77.62$, $p < 0.0001$, and between-subject JOLs, $F(1, 94.89) = 5.19$, $p < 0.0249$. The within-subject JOLs indicate that participants JOLs are predictive of recall at an item level. Because this factor did not interact with episodic retrieval of the reference item, the benefit of gaining a better perspective of what item features constitute as diagnostic of recall did not develop as a function of episodic retrieval of the reference item. At the average JOL, a participant has a 21% chance of recalling a given item and that chance increases by 2% with each increase in JOL. This is quite similar to the item-level calibration effects observed in Experiment 1. The second main effect, between-subject JOLs, indicates that those who generally give higher JOLs have better recall than those who have give generally lower JOLs. While this main effect is reliable, Figure 27 does not plot the fitted values at different between-subject JOLs because this factor did not interact with any other factor. To be specific, the relationship between recall and within-subject JOLs described in Figure 27 maintains across the scale usage of JOLs. Along with the gamma results, the GLMM supported the finding that episodic retrieval of the reference item did not produce greater precision in metacognition on immediate JOLs.

LMM: Regressing transfer item JOLs on reference item JOLs

The results from Experiment 1 indicate the item-level effects are prevalent when making immediate JOLs in the presence of other information. To be exact, using episodically related word pairs as mnemonic cues impact JOLs. The extant that episodic information is diagnostic of eventual recall will determine the usefulness of considering such information when making immediate JOLs. To investigate whether the use of

episodically related information has an effect on immediate JOLs, we employ a mixed effects model and begin by regressing transfer item JOLs on reference item JOLs.

As seen in Table 9, variables were added in the same order as in Experiment 1. Notably, all the random effects increased the fit of the model. Moreover, fitting a simpler covariance structure in Model 7 was rejected, much like the regression in Experiment 1. Therefore, we endorsed and explored the fixed effects using Model 6, and the coefficients can be found in Table 10. First, there was a main effect for within-subject reference item JOLs, $F(1, 99.91) = 24.17$, $p < 0.0001$, between-subject reference item JOLs, $F(1, 95.11) = 23.24$, $p < 0.0001$, and episodic retrieval of the reference item, $F(1, 95.1) = 16.39$, $p < 0.0001$. These effects are qualified by two higher order interactions. First, there was a significant interaction of episodic retrieval of the reference item and within-subject JOLs, $F(1, 3668) = 28.13$, $p < 0.0001$, such that transfer item JOLs were more influenced by reference item JOLs when the reference item had been presented as cue-only than when presented as cue-target. However, there was a three-way interaction of episodic retrieval of the reference item, within-subject JOLs, and between-subject JOLs, $F(1, 3668) = 5.11$, $p = 0.0238$. These results mirror those obtained in Experiment 1, thus we use an identical method to explore the nature of the three-way interaction. We plotted the fitted values of the relationship between episodic retrieval of the reference item and within-subject JOLs and three different between-subject intervals, one standard deviation below the mean JOL ratings for reference items, at the mean JOL ratings, and one standard deviation above the mean JOL ratings. As seen in Figure 28, participants who gave lower JOLs to the reference items were less impacted by their re-presentation at the time of transfer item JOLs. This

was the case regardless of whether the reference item was presented as cue-target or as cue-only. However, participants were more influenced by cue-only reference items when the reference items themselves received higher JOLs.

These results are identical to Experiment 1. The key difference between the experiments, however, is that the reference item JOLs are delayed. Because delayed JOLs encourage the use of covert retrieval prior to making judgments, the interaction between within-subject JOLs and episodic retrieval of the reference item was expected. This interaction increased as participants offered higher JOLs. Collectively, this strongly indicates that being able to covertly retrieve an episodically related item from memory during the time of immediate JOLs will lead to inflated JOLs. We provide additional support for this proposition next.

LMM: Regressing transfer item JOLs on reference item recall

Thus far, we have replicated all the effects of Experiment 1. Supporting our hypothesis that episodic retrieval of the reference item alters metacognition for immediate JOLs. To provide another test for this hypothesis, we use the final recall of reference items as a proxy for successful retrieval during the time of transfer item JOLs. If episodic retrieval of the reference item influences transfer item JOLs it should only occur for cue-only reference items, when covert retrieval is possible. Thus far, we have retained the most complex models for the mixed effects analysis. We continue to do so for the remaining results. We analyzed the effect of reference item recall on transfer item JOLs in a mixed-effects model with an unstructured covariance matrix on the random factor, reference item recall. Again, we model reference item recall as a random factor because it is not experimentally controlled.

There was a main effect of reference item recall, $F(1, 99.43) = 19.39$, $p < 0.0001$, that was qualified by the two-way interaction, $F(1, 1390) = 24.48$, $p < 0.0001$. The nature of the interaction is slightly different than from Experiment 1. As can be seen in Figure 29, when participants were unable to recall a reference item it reduced their JOLs relative to the other three conditions (post-hoc contrast, $p < 0.0001$). Thus, unlike in Experiment 1 where recalling a reference item increases transfer item JOLs, manipulating episodic retrieval within-subjects causes a relatively different subjective use of metamnemonic cues. These data suggest that participants become accustomed to the high fluency of processing previously encountered information and that only in the absence of such information are JOLs reduced.

Global predictions

To test whether participants were aware that considering previously studied items impacted their JOLs, we analyzed their subjective accuracy ratings for the different item types. In the previous experiment, subjects did not rate transfer item JOLs any differently when presented with cue-only reference items (and thereby promoting covert retrieval) and cue-target reference items. The lack of a difference may be attributed to the relative nature of JOLs and manipulating the likelihood of retrieving the reference item between subjects prevented the subjective experience between the two types of judgments. In the current experiment, we manipulated the likelihood of retrieving the reference items at the time of transfer item JOLs within subjects and thus if JOLs were relative in nature we expect that participants would detect differences between these ratings.

The difference between subjective accuracy of transfer item JOLs differed as a

function of episodic retrieval of the reference item, $F(1, 96) = 15.205$, $MSE = 583$, $p = 0.0002$, $\eta_p^2 = 0.13$, $\omega_G^2 = 0.0327$, such that the prediction was confirmed: Participants felt their transfer item JOLs were less accurate when the episodically related item was presented as cue-only, which is consistent with the item-level analyses.

Discussion

The results from Experiment 2 by and large replicate the previous experiment. Recall levels increased for reference items when presented as cue-target compared to when presented as cue-only during the time of transfer item JOLs, although the double exposure resulted in better memorability overall. Essentially, this is a practice effect, a potent method of increasing memorability of practiced information. Additionally, reference items presented as cue-only during the transfer item JOLs received a memorial boot compared to transfer items. One cause for this effect is a potential testing effect afforded by the covert retrieval of the reference item. This explanation adds support to the methodological design, that participants are attempting covert retrieval attempts with the presentation of a cue-only reference item.

The JOL pattern in Experiment 2 was also partially supported. In the first experiment, the transfer item JOLs were lower than reference and control items but did not differ between the episodic retrieval conditions. While the same pattern emerged regarding the main effect of reference items and transfer items, the simple effect emerged for the transfer items. Transfer items were given lower JOLs when the reference item had been presented a cue-only. Initially, these data can be taken as support that covert retrieval leads to a greater awareness of episodic memory where the byproduct is a reduction in the illusion of competency. This hypothesis is also supported

by the calibration data, where the episodic-retrieval transfer items were better calibrated.

However, as the item-level data show, the reduction in the illusion of competency is contingent on the item characteristics of the reference items. To be exact, the item-level data indicate that the retrievability of reference items, measured either through the delayed JOL or by final recall of the reference item, play a key role in transfer item JOLs. To the extent that the previous item recall is indicative of future recall for the to-be-judged item, the JOLs immediate items receive will be accurate. For example, when overall recall is poor for the reference items then the same pattern would be expected for the transfer item recall. The cue-only reference items enlighten participants on this global effect and thereby resulting in the reduction in the illusion of competency.

Finally, the data for participants' subject accuracy of the different item type JOLs show that participants are aware of the differential effects of presenting the reference item in different formats. Specifically, participants believe their ratings are poorer when the reference item is presented in cue-only format as compared to a cue-target presentation format. This represents a paradox. On the one hand, they are correct in that there is greater influence on the transfer item JOLs. On the other hand, this greater influence is beneficial overall and reduces overconfidence. Furthermore, because participants were able to detect the differential effects of the episodic retrieval of the reference item during immediate JOLs, it could be the case that participants may not naturally adapt such a strategy when giving immediate JOLs. We explore this general idea in Experiment 3.

Experiment 3

The main objective in Experiment 3 is to use a more subtle method of inducing covert retrieval. In the previous experiments, participants were required to consider a previously studied word pair (i.e., a reference item) when making transfer item JOLs. This requirement is removed and replaced with a methodological design that is hypothesized to bring about the same covert retrieval of episodically related information. That design change is presenting a previously encountered word pair (reference item) along with an immediately encountered word pair (transfer item) so that both items receive JOL ratings during the same timeframe (see Figure 31). The reference item always appears above the transfer item and participants are required to rate the word pairs in the order they appear (i.e., top-to-bottom). Because the reference items receive delayed JOLs, we hypothesize that the dynamics of covert retrieval will be the same as in the previous experiments. Covert retrieval is likely to occur when the reference items are presented as cue-only, but not likely to occur when the reference items are presented as cue-target. We test whether the natural covert retrieval when making delayed JOLs will affect transfer item immediate JOLs, much like in the previous experiments.

Method

Participants

Seventy-one University of Oklahoma undergraduate students enrolled in the introductory psychology course participated for partial course credit.

Design

The experiment used a two study-test trial design. In the previous experiments,

recall levels were moderate and thus to ensure the previous effects found generalize to learning conditions with higher recall levels, we use two study-test trails with shorter study lists. Reducing the lag between study and test is a common method for increasing recall. We implement an additional procedure to generalize the previous findings to more realistic conditions of immediate JOLs where judgments are made in a temporally relative manner. This was achieved by having participants make JOLs on two word pairs that were presented simultaneously (See Figure 31). The word pair presented on the top took on the role of the reference item and received delayed JOLs, blocked by 10 intervening items. Thus, 10 reference items were studied in the first block and the following block contained the judgment phase for the reference items, along with study and judgment phase for transfer items. During the judgment phase, the transfer items were always presented below the reference items. That is, after studying each transfer item, participants rated a reference item that was previously studied in one of the 10 prior serial positions just before rating the transfer item. This process was repeated twice within each study list. The presentation of the reference item at the time of transfer item JOLs was manipulated; half of the transfer items were presented with cue-only reference items and the other half with cue-target reference items.

Materials

The stimuli comprised the same 180 four-letter common nouns as in the previous experiments. Word pairs were randomly paired anew for each participant. Eighty word pairs were selected from that set; half of which served as reference items and the other half as transfer items. Then these word pairs were split in half and assigned to either the first or second study trial. All word pairs were studied and tested.

Apparatus

The same method of delivery was used as in the previous experiments.

Procedure

The experiment consisted of an instruction phase, followed by two study/cued-recall sessions. After participants consented to the experiment, the computer displayed an interactive instruction set, which began with a general description of the study list. Participants were told they would encounter two study-test trials with each study phase containing 40 cue-target word pairs, each presented for 4-seconds, and then make judgments on their confidence in recalling the target word when provided the cue on a subsequent memory test (e.g., memory judgments). Participants were told to rate items from top to bottom. This ensured that the reference item was always rated prior to the immediately studied item. Reference items always received delayed JOLs. The scale was similar to the previous experiments except that the scale had intervals of five units rather than single units.

The methodology of delaying ratings for reference items occurred in a similar fashion as in Experiment 2. The first block contained 10 reference items, which were studied sequentially. The block that followed contained 10 studied items transfer items. Immediately after studying each transfer item, the computer presented a judgment task in which a previously studied reference item from the previous block was randomly selected to be displayed above the transfer item (See study block 1b in Figure 31). The judgment phase was self-paced. The blocked study/judgment process occurred twice in each study list, totaling to 40 word pairs being studied and judged per list. A cued recall test followed each study list. As in the previous experiments, word pairs were tested in a

random order and unconstrained by time.

Results and Discussion

The analyses from Experiment 3 focus on the main findings in the previous studies: item-level effects of episodic retrieval. As such, we summarize the aggregated data in Table 11 and report the most relevant findings. In addition, we inspect item-level effects in a mixed-effects model treating all within-subject manipulated variables as random factors. As before, this is motived by experimental design rather than statistical necessity (e.g., testing random effects and only including effects that improve model fit significantly). With exception of the item-level calibration model, all random factors contributed significantly to model fit, and therefore, the two approaches would lead to similar conclusions.

Aggregated data analysis

The polarization of JOLs when delaying judgments is an indicator that participants are performing covert retrieval attempts prior to judgments (see, e.g., Kimball & Metcalf, 2003). The delayed judgments made for cue-only reference items are consistent with that pattern under conditions that are likely for retrieval (See Figure 32). Those conditions are such that only the word pair's cue is presented during the judgment phase. No other condition resembled a polarized pattern of JOLs. This supports our postulation that presenting delayed cue-only reference items would cause covert retrieval, much like the instructional manipulation in the previous experiments.

A common finding with delaying judgment is lower overall JOLs as compared to immediately judged items (see, Dunlosky & Metcalf, 2009). To test whether this pattern was found in Experiment 3, we analyzed JOLs in a two-way repeated measures

ANOVA with episodic retrieval of the reference item and item condition as factors. All three effects reached significance: item type, $F(1, 70) = 4.34$, $MSE = 260$, $p = 0.0409$, $\eta_p^2 = 0.05$, $\omega_G^2 = 0.0131$; episodic retrieval of the reference item, $F(1, 70) = 4.4$, $MSE = 69$, $p = 0.0395$, $\eta_p^2 = 0.06$, $\omega_G^2 = 0.0136$; item type by episodic retrieval of the reference item interaction, $F(1, 70) = 24.26$, $MSE = 43$, $p < 0.0001$, $\eta_p^2 = 0.26$, $\omega_G^2 = 0.0831$. The interaction was driven by the difference in JOLs for reference items between their cue-only and cue-target presentation (see Table 11). That simple effect was reliable, $F(1, 70) = 13.68$, $MSE = 91$, $p = 0.0004$, $\eta_p^2 = 0.16$, $\omega_G^2 = 0.0476$; delayed cue-only reference items had lower JOLs.

More importantly, however, is testing whether covert retrieval extended to transfer items and whether this would result in underconfidence. Unlike in the previous experiments, covert retrieval of the reference item did not result in lower transfer item JOLs. In fact, the opposite occurred: Transfer items were given *higher* JOLs when judged after a cue-only reference item, $F(1, 70) = 13.68$, $MSE = 91$, $p = 0.0004$, $\eta_p^2 = 0.16$, $\omega_G^2 = 0.0476$. While this finding is at odds with the previous results, the interpretation of these data depends on the characteristics of the reference items. Specifically, if recall of the reference item affects the magnitude of the transfer item JOLs, then these data can only be interpreted when controlling for differences in reference item recall. As the means in Table 11 indicate, there is a large difference in recall of the reference item as a function of the likelihood of episodic retrieval, $F(1, 70) = 19.87$, $MSE = 0.01$, $p < 0.0001$, $\eta_p^2 = 0.22$, $\omega_G^2 = 0.1$. If the retrievability of the reference item impacts transfer item JOLs in a similar fashion as in Experiment 1 and 2, then the increased transfer item JOLs is attributed to the increase in reference item

recall. As such, we reserve interpretation of these data until further investigating item-level effects.

Lastly, calibration effects and gamma correlation were analyzed in separate two-way repeated measures ANOVAs. The calibration effects revealed a main effect of item type, $F(1, 70) = 33.02$, $MSE = 0.02$, $p < 0.0004$, $\eta_p^2 = 0.32$, $\omega_G^2 = 0.1249$, such that delaying judgments (i.e., the reference items) resulted in better calibration. The transfer items did not differ as a function episodic retrieval of the reference item ($F = 0.2$). The last notable finding is the gamma results. As seen in Table 11, gamma was higher for cue-only reference items compared to each of the other three conditions (ps < 0.0001). This delayed-JOL effect is attributable to covert retrieval and is supported by the polarized JOLs described earlier. As in Experiment 1 and 2, gamma did not differ between the two types of transfer items ($F = 0.87$).

The aggregated data in Experiment 3 resemble, but are not identical to, the previous findings reported. Requiring participants to use episodically related information reduced immediate JOLs in Experiment 1, and when covert retrieval was likely to occur, reduced immediate JOLs in Experiment 2. However, when no requirement is placed on episodic retrieval, transfer items were not given lower JOLs. While this result is striking, the item-level data from the previous studies indicate that additional factors contribute to immediate JOLs. The remaining aggregated data tell a similar story as before. Overall calibration is similar between the transfer items under likely and unlikely covert retrieval of the reference item, as in Experiment 1. And finally, covert retrieval does not increase the ability to detect recallable word pairs from nonrecallable word pairs, as indicated by the gamma results.

Item-level data analysis

Despite equal mean gamma coefficients between transfer items rated alongside cue-only or cue-target reference items, we investigate item-level calibration effects by regressing recall on transfer item JOLs. We modeled all experimentally controlled within-subject variables and subject-produced variables as random effects, nested within subjects. We added the random effect of episodic retrieval of the reference item to the unconditioned model and this improved model fit $x^2(1) = 474$, $p < 0.0001$. The fixed effects of JOLs (within-subject centered and between-subject centered JOLs) further improved the fit of the model, $x^2(2) = 78$, $p < 0.0001$. Then, much like in the previous models investigate item-level calibration, the random effect of within-subject JOLs did not improve model fit, $x^2(3) = 5$, $p < 0.1497$. Despite the lack of improvement, we continued building the model based on the rationale offered in the above sections. Neither the fixed effect of episodic retrieval of the reference item, $x^2(1) = 0.09$, $p = 0.78$, nor the interaction of the fixed effects improved model fit, $x^2(4) = 5$, $p = 0.2665$. However, we analyze the fixed effects from the most complex model on the basis of the experimental design—although trimming non-significant effects does not change the interpretation of the results. The only effect that reached significance was the main effect of within-subject JOLs, $F(1, 56.26) = 51.28$, $p < 0.0001$; the probability of recalling an item increases by 2% with each increase in JOL. No differences were found as a function of episodic retrieval. The overall fitted model for the effect of within-subject JOLs is plotted in Figure 33.

In a similar fashion, we used a linear mixed model to investigate the effects of reference item JOLs on transfer item JOLs. The reference items were given delayed

JOLs and as such, cue-only reference item JOLs likely reflect recallability of the target word. By contrast, delayed JOLs for cue-target reference items are made on the basis of intrinsic cues specific to the word pair (e.g., processing fluency, see Mueller, et al., 2014). We built the model as before and with exception of the fixed and random effect of episodic retrieval of the reference item ($p = 1$; increased deviance), all variables significantly improved model fit (ps < 0.0001). Table 12 displays the coefficients for the final model. The two types of reference item JOLs had a significant correlation with transfer item JOLs (within-subject: $F(1, 70.54) = 50.82$, $p < 0.0001$; between-subject: $F(1, 69.03) = 79.96$, $p < 0.0001$). Reference item JOLs predict transfer item JOLs at both an item level and at a global level. The latter effect can be interpreted such that those who offer higher JOLs for reference items will also offer higher JOLs for transfer items. Additionally, there was a main effect of episodic retrieval of the reference item, such that transfer items had higher JOLs when episodic retrieval of the reference item was likely to occur. This pattern supports the finding in the aggregated data analyses. However, unlike in the previous experiments there was no interaction of any of the variables. The fitted main effect of within-subject reference item JOLs is plotted in Figure 34. The lack of an interaction is somewhat unexpected, but can be interpreted in terms of processing fluency involved with the representation of the intact reference item. This explanation is consistent with the results from Experiment 1 where immediate JOLs of reference items were correlated with transfer item JOLs. Thus, the conclusion is that both episodic retrieval of the reference item (i.e., a cue-only reference item) and the reprocessing of episodically related information (i.e., a cue-target reference item) affects metacognition.

To exam a more direct effect of episodic retrieval of the reference item on transfer item JOLs we fit a mixed model predicting transfer item JOLs as a function of recall of the reference item on final recall. The means can be seen in Figure 35, and show a pattern similar to Experiment 1 and 2: When covert retrieval of the reference item is likely, successful retrieval of the reference item leads to increases in transfer item JOLs. This was confirmed by a two-way interaction of item type and episodic retrieval of the reference item, $F(1, 687.2) = 10.55$, $p = 0.0012$. Thus, to the extent that retrieval of previous information is predictive of recall for the currently judged item, this mnemonic cue will be diagnostic. In the current experiment, however, the random assignment of stimuli to conditions is unlikely to provide beneficial retrieval effects on the final test. However, on the one hand, when mean recall is lower, such as in Experiment 1 and 2, this leads to a reduction in the illusion of competency. On the other hand, when recall levels are higher, such as in Experiment 3, this leads to inflated JOLs.

Collectively, these results support the premise put forth in this paper: Metamemory judgments are made in a relative manner, relying on features of episodically related information. Under the methods of Experiment 3, judgments were made in a natural manner; participants were not forced to retrieve or use previously encountered information. The same major patterns were found as when instructed to consider episodically related information in Experiment 1 and 2. Metacognitive judgments are made on retrieval fluency of related information, but only when covert retrieval is made possible. Successfully retrieved information leads to higher JOLs for items judged in the same temporal context. When covert retrieval is not possible, but episodic information is still available, metacognitive judgments rely on intrinsic cues of

the available information. In the end, episodic related information is naturally incorporated in metacognitive judgments. Supporting the proposition that metacognition is comparative in spirit.

General discussion

Summary

In Experiment 1, participants made immediate judgments on all sets of items, reference items, transfers items, and control items. The JOLs for reference items, which were judged prior to any experimental manipulations, showed a near normal distribution of ratings—typical of immediate JOLs (Dunlosky & Nelson, 1994). Compared to JOLs given to the other classes of items, reference items were rated much higher (i.e., more memorable). Once the administration of the experimental manipulations occurred, JOLs were reduced for the remaining items, the transfer and control items. To be exact, participants were required to consider previously studied information when making judgments on transfer items and doing so led to lower JOLs. The presentation of the previously studied information was manipulated to vary the degree to which covert episodic retrieval was likely to occur. This procedure reduced the magnitude of trasnger item JOLs compared to JOLs for the reference items and for control items. Control items received JOLs in the same manner as did the reference items; they were made without any additional information presented to the participants. These JOLs were lower compared to the reference items, providing strong evidence that the illusion of competency extended to metacognitive judgments made under a different condition, one in which consideration of previously studied items was no required. One interpretation of these results is that participants were able to form a global mnemonic cue that

represents an internal awareness of episodic memory. This cue may represent an internal baseline in which newly judged items are compared to. This interpretation is consistent with cue utilization and emphasizes the relativity of metacognitive judgments (Koriat, 1997; Schwartz, 1994).

The reduction in competency occurred regardless of the likelihood of covert retrieval, both for transfer items and for control items. Consequently, this suggests that multiple sources of information contribute to the development of global metamnemonic cues, much the same as is thought with metamnemonic cues of individual items (Serra & Dunlosky, 2005). Retrieval fluency is the most likely intrinsic cue that develops under conditions that promote the use of covert retrieval (Koriat & Ma'ayan, 2005). The collection of retrieval attempts over time breeds a general mnemonic cue that is applied list wide to all newly encountered information. In support of this view is the reduction in JOLs for control items compared to the JOLs for the reference items. The only difference between these conditions is the occurrence within the study list. Reference item judgments were made prior to experimental manipulations, whereas control item judgments were made afterward.

The reduction in JOLs is also attributed to the use of a different metacognitive cue, an intrinsic cue generated by the reference item. The processing fluency of the reference item is measured by their immediate JOLs. The correlation between immediate JOLs for reference items and transfer items was strong when covert retrieval was possible (likely retrieval condition for transfer items), but very weak when retrieval was not possible (i.e., presenting a cue-only reference item). This finding suggests that the fluency of the cue word, perhaps based on familiarity (Metcalf et al., 1993),

influences the magnitude of JOLs for transfer items. One potential explanation is that the fluency or familiarity of the cue is misattributed to the transfer item. This finding is somewhat surprising and seems like a promising effect for further experimentation.

The relative nature of metacognitive judgments was also explored in Experiment 1. Examining the first set of items prior to the experimental manipulations allows to test whether residuals were correlated once controlling for serial position effects. Two mixed effects models regressed JOLs on serial position, with each model differing in the residual covariance structure. The first model fit a typical independent and identically distributed residual covariance structure and the second model fit a first-order autoregressive residual covariance structure. The latter model provided a better fit. This is taken as evidence that judgments made closer in time are more related than those made further apart in time. To some degree then, participants were naturally basing judgments on episodically related information.

Finally, the last mnemonic cue that contributes to the reduction in JOLs is a global belief-based cue. Participants rated the subjective accuracy of their JOLs for the various types of items and the end of the study phase. These results indicate that participants believe their JOLs were less accurate for transfer item JOLs and these ratings did not differ as a function of episodic retrieval of the reference item. The lack of a difference seems to imply that the main contributor to the reduction in JOLs in the unlikely retrieval condition is a belief based mnemonic cue. Participants lack a general confidence in their predictions, and as a byproduct, error on the side of providing lower JOLs. Paradoxically, this results in the reduction of competency, as measured by JOL magnitude.

Two major design changes were made in Experiment 2, episodic retrieval of the reference item was manipulated within subjects and reference items were given delayed JOLs. The data from participants' subjective accuracy of JOLs motivated the first design change. As shown by the item-level data analysis, episodic retrieval of the reference item had a greater influence on transfer item JOLs, but participants rated the accuracy equal to that when episodic retrieval was not likely. Accordingly, we tested whether the lack of a difference is due to the relative nature of JOLs. Because retrievability of the reference item was manipulated between subjects, participants do not share the experience of both types of judgments. Instead, participants can only compare the retrievability condition to the control condition. Manipulating this factor within subjects allows for such a comparison to take place. It also provides an additional test of the assumption that metacognitive judgments are relative in nature. Accordingly, these data do show differences in subjective accuracy of JOLs as a function of retrievability of the reference item, with lower accuracy judgments given to transfer items when the reference item is likely to be covertly retrieved.

The second design change was intended to test the major finding in Experiment 1, which showed the retrievability of the reference item had a major influence on transfer items JOLs. Retrievability of the reference item was operationalized on the final recall accuracy for that item. The assumption being that if recall is successful during the final recall then it is also likely that retrieve is possible during the judgment phase (see Nelson, et al., 2004). Because the basis of delayed JOLs is primarily retrieve fluency, high JOLs reflect retrievability of that item. It follows then, that the same conclusion should result from a regression of transfer item JOLs and reference item

JOLs as a regression of transfer item JOLs and reference item final recall. Indeed, these predictions materialized providing converging evidence that retrieval of episodically related information influences metacognition.

The requirement to consider previously studied information was removed in Experiment 3, and instead, participants engaged in judging a pair of items simultaneously. The reference item appeared above the transfer item and was presented in either cue-only or cue-target format. Thus, a delayed-JOL was provided prior to making immediate JOLs on transfer items. Using the same rationale as in the previous experiments, covert retrieval is likely to occur prior to making cue-only delayed JOLs. In addition, any influence on the transfer item would be a result of a natural process. Thus, Experiment 3 used a more implicit approach to testing whether episodic retrieval influences immediate JOLs. The reference item JOLs were polarized as expected, supporting the notion that covert retrieval likely occurred. Most importantly, immediate JOLs were influenced as a function of successful retrieval of the reference item.

Utilization of episodic information

Metacognitive judgments are made through the use of various information. Immediately studied and judged items are based on item attributes, or characteristics, that reveal an item's a priori learning fluency, an intrinsic cue. The information used when making post-learning judgments (i.e., delayed JOLs) depends on whether covert retrieval is possible or not. When covert retrieval is possible then heavy reliance is placed on the retrieval fluency of the to-be-judged item. Successfully retrieved items are given higher JOLs and unsuccessfully retrieved items are given lower JOLs. The outcome is polarized JOLs. When covert retrieval is not possible in delayed judgments

(e.g., cue-target word pairs) then intrinsic cues are relied on when making JOLs, much like immediately judged items. Of these types of judgments, the most accurate is a delayed judgment with the potential for covert retrieval. This generates potent mnemonic cues that are predictive of future recallability of the judged item. The focus of the current paper was to bring the same type of awareness to immediately judged items. However, the mnemonic cue is not retrieval of the to-be-judged item, but of episodically related information. In the event that one realizes they cannot retrieve information from the past, in which encoding occurred in the same context, then this awareness should breed a global mnemonic cue. Experiment 1 demonstrated that such a global cue is likely to develop and be used for judgments. Control item JOLs were reduced when the episodically related information was provided to participants. However, the observed reduction occurred when episodic information could be retrieved (cue-only reference item) as well as when information could not be retrieved (cue-target reference item). It is a surprising finding that cue-target reference items can prevent the illusion of competency and future studies could explore the contributing factors to this phenomenon.

Collectively, the main objective of bringing insight into the current state of participants' episodic memory was by and large accomplished. Although, the consequences of such awareness did not transpire to a better metacognitive judgments as measured by a cued-recall test. In the event that episodically related information is indicative of future recall, then the ability to retrieve previous information will lead to more accurate judgments. In Bayesian terms, the retrievability of related information serves as the prior used to judge the posterior probability of recalling a new item in the

future. In more complex learning environments, the dynamics of recall are likely to be more contextual and semantic. Thus, it is reasonable to assume that using retrievability of semantically and episodically related information as the basis of metacognitive judgments will lead to more accurate judgments in free recall task that rely on such cues. Further studies are needed to understand the generalizability of the current effects to new situations, and in particular, to situations where retrieval of previous information share episodic and semantic traces to the to-be-judged item.

In the current experiments, participants increased or reduced transfer item JOLs depending on the recallability of the reference item. If a participant's recall is generally poor, then this transpires into lower JOLs. Recall was relatively low in Experiment 1 and 2 and a corresponding underconfidence was observed. Recall levels were higher in Experiment 3 (intended by the design changes; shortening the lag between study and test) and as a result, transfer item JOLs became inflated, which is attributed to increases in the number of successfully retrieved items during the judgment phase.

In sum, these findings have practical relevance to judgments made during everyday learning. Judgments made during learning are not only susceptible to intrinsic properties but also to what is currently active in memory. Because relational processing is likely to occur during new learning, there is a high chance that covert retrieval of information occurs. In the end, the current data suggest that the fluency of that retrieval (activation of episodically related traces) is used to make judgment ratings. Ultimately, the retrievability of related information determines whether judgments made during learning reflect illusions of competency.

References

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of memory and language*, *59*(4), 390-412.

Benjamin, A. S., & Bjork, R. A. (1996). Retrieval fluency as a metacognitive index. *Implicit memory and metacognition*, 309-338.

Benjamin, A. S., Bjork, R. A., & Schwartz, B. L. (1998). The mismeasure of memory: when retrieval fluency is misleading as a metamnemonic index.*Journal of Experimental Psychology: General*, *127*(1), 55.

Benjamin, A. S., & Diaz, M. (2008). Measurement of relative metamnemonic accuracy. *Handbook of memory and metamemory*, 73-94.

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*(5), 610-632.

Besken, M., & Mulligan, N. W. (2013). Easily perceived, easily remembered? Perceptual interference produces a double dissociation between metamemory and memory performance. *Memory & Cognition*, 41(6), 897–903. doi:10.3758/s13421-013-0307-8

Besken, M., & Mulligan, N. W. (2014). Perceptual fluency, auditory generation, and metamemory: Analyzing the perceptual fluency hypothesis in the auditory modality. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 40(2), 429–440. doi:10.1037/a0034407

Box, G. E., & Pierce, D. A. (1970). Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. *Journal of the American Statistical Association*, 65(332), 1509-1526.

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., & White, J.-S. S. (2009). *Generalized linear mixed models: a practical guide for ecology and evolution. Trends in Ecology & Evolution*, 24(3), 127–135. doi:10.1016/j.tree.2008.10.008

Burke, D. M., MacKay, D. G., Worthley, J. S., & Wade, E (1991). On the tip of the tongue: What causes word finding failures in young and older adults. *Journal Of Memory And Language*, 30(5), 542-579. doi:10.1016/0749-596X(91)90026-G

Bjork, E., & Bjork, R. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. *Psychology and the Real World: Essays Illustrating Fundamental Contributions to Society. New York: Worth Publishers*.

Bjork, R. A., Dunlosky, J., & Kornell, N (2013). Self-Regulated Learning: Beliefs, Techniques, and Illusions. *Annual Review Of Psychology*, 64(1), 417-444. doi:10.1146/annurev-psych-113011-143823

Carroll, M., Nelson, T. O., & Kirwan, A. (1997). Tradeoff of semantic relatedness and degree of overlearning: Differential effects on metamemory and on long-term retention. *Acta Psychologica*, *95*(3), 239-253.

Cohen, J. (1973). Eta-squared and partial eta-squared in fixed factor ANOVA designs. *Educational and Psychological Measurement*, 33, 107-112.

Costermans, J., Lories, G., & Ansay, C. (1992). Confidence level and feeling of

knowing in question answering: The weight of inferential processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*(1), 142-150.

Crainiceanu, C. M., & Ruppert, D. (2004). Likelihood ratio tests in linear mixed models with one variance component. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *66*(1), 165-185.

Cutting, J. E. (1975). Orienting tasks affect recall performance more than subjective impressions of ability to recall. *Psychological Reports*, *36*(1), 155-158.

DeCarlo, L. T. (2002). Signal detection theory with finite mixture distributions: Theoretical developments with applications to recognition memory. *Psychological Review*, *109*(4), 710-727.

Dunlosky, J., & Nelson, T. O. (1992). Importance of the kind of cue for judgments of learning (JOL) and the delayed-JOL effect. *Memory & Cognition*, *20*(4), 374-380.

Dunlosky, J., & Metcalfe, J. (2008). Metacognition. Los Angeles, CA: Sage.

Dunlosky, J., & Thiede, K. W. (2004). Causes and constraints of the shift-to-easier-materials effect in the control of study. *Memory & Cognition*, 32(5), 779-788.

Dunning, D., Johnson, K., Ehrlinger, J., & Kruger, J (2003). Why people fail to recognize their own incompetence. *Current Directions In Psychological Science*, 12(3), 83-87. doi:10.1111/1467-8721.01235

Eakin, D. K. (2005). Illusions of knowing: Metamemory and memory under conditions of retroactive interference. *Journal of Memory and Language*, *52*(4), 526-534.

Eakin, D. K., & Hertzog, C. (2012). Immediate judgments of learning are insensitive to implicit interference effects at retrieval. *Memory & cognition*,*40*(1), 8-18.

Finn, B., & Metcalfe, J. (2008). Judgments of learning are influenced by memory for past test. *Journal of memory and language*, *58*(1), 19-34.

Gardiner, F. M., Craik, F. I. M., & Bleasdale, F. A. (1973). Retrieval difficulty and subsequent recall. *Memory & Cognition*, 1(3), 213–216. doi:10.3758/BF03198098

Gelman, A. and Hill, J. (2006) Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological review*, *91*(1), 1.

Gonzalez, R., & Nelson, T. O. (1996). Measuring ordinal association in situations that contain tied scores. *Psychological bulletin*, *119*(1), 159.

Goodman, L. A., & Kruskal, W. H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association*, *49*(268), 732-764.

Green, B. F. and Tukey, J. W. (1960). Complex analyses of variance: General problems. *Psychometrika*. 25. 127–152.

Greenhouse, S. W., & Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika*, 2495-112. doi:10.1007/BF02289823Johnston, J. and DiNardo, J. (1997) Econometric Methods Fourth Edition. New York, NY: The McGraw-Hill Companies, Inc.

Hart, J. T. (1967). Memory and the memory-monitoring process. *Journal of verbal learning and verbal behavior*, *6*(5), 685-691.

Hertzog, C., Dunlosky, J., Robinson, A. E., & Kidder, D. P. (2003). Encoding fluency is a cue used for judgments about learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(1), 22.

Johnston, J., & DiNardo, J. (1972). Econometric methods. *New York*, *19*(7), 22.

Karpicke, J. D., & Roediger, H. L. (2007). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language*, 57, 151–162. doi:10.1016/j.jml.2006.09.004

Kellermann, A. P., Romano, J. L., de Gil, P. R., Pham, T., Rasmussen, P., Chen, Y. H., ... & Kromrey, J. D. GEN_OMEGA2: A SAS® Macro for Computing the Generalized Omega-Squared Effect Size Associated with Analysis of Variance Models.

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

Kelley, C. M., & Jacoby, L. L. (1996). Adult egocentrism: Subjective experience versus analytic bases for judgment. *Journal of Memory and language*, *35*(2), 157-175.

Kelley, C. M., & Jacoby, L. L. (1998). Subjective reports and process dissociation: Fluency, knowing, and feeling. *Acta Psychologica*, *98*(2), 127-140.

Kenward, M. G. and Roger, J. H. (1997), "Small Sample Inference for Fixed Effects from Restricted Maximum Likelihood," *Biometrics*. 53: 983-997.

Kimball, D. R., & Metcalfe, J. (2003). Delaying judgments of learning affects memory, not metamemory. *Memory & Cognition*, *31*(6), 918-929.

Kimball, D. R., Muntean, W. J., & Smith, T. A. (2010). Dynamics of thematic activation in recognition testing. *Psychonomic Bulletin & Review*, *17*(3), 355–361. doi:10.3758/PBR.17.3.355

Koriat, A. (1993). How do we know that we know? The accessibility model of the feeling of knowing. *Psychological review*, *100*(4), 609-639.

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. (2005). Illusions of competence in monitoring one's knowledge during study. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(2), 187.

Koriat, A., & Bjork, R. A. (2006a). Mending metacognitive illusions: a comparison of mnemonic-based and theory-based procedures. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(5), 1133-1145.

Koriat, A., & Bjork, R. A. (2006b). Illusions of competence during study can be remedied by manipulations that enhance learners' sensitivity to retrieval conditions at test. *Memory & Cognition*, *34*(5), 959-972.

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*(4), 643-656.

Koriat, A., & Ma'ayan, H. (2005). The effects of encoding fluency and retrieval fluency on judgments of learning. *Journal of Memory and Language*, *52*(4), 478–492. doi:10.1016/j.jml.2005.01.001

Koriat, A., Sheffer, L., & Ma'ayan, H. (2002). Comparing objective and subjective learning curves: judgments of learning exhibit increased underconfidence with practice. *Journal of Experimental Psychology: General*, *131*(2), 147-162.

Koriat, A., Ackerman, R., Lockl, K., & Schneider, W. (2009). The memorizing effort heuristic in judgments of learning: A developmental perspective. *Journal of experimental child psychology*, *102*(3), 265-279.

Koriat, A., Ma'ayan, H., & Nussinson, R. (2006). The intricate relationships between monitoring and control in metacognition: lessons for the cause-and-effect relation between subjective experience and behavior. *Journal of Experimental Psychology: General*, *135*(1), 36.

Kornell, N. (2010). Failing to predict future changes in memory: A stability bias yields long-term overconfidence. In A. S. Benjamin (Ed.), Successful remembering and successful forgetting: A festschrift in honor of Robert A. Bjork (pp. 365–386). New York, NY: Psychology Press.

Kornell, N., & Bjork, R. A. (2009). A stability bias in human memory: overestimating remembering and underestimating learning. *Journal of Experimental Psychology: General*, *138*(4), 449-468.

Kornell, N., Rhodes, M. G., Castel, A. D., & Tauber, S. K. (2011). The Ease-of-Processing Heuristic and the Stability Bias: Dissociating Memory, Memory Beliefs, and Memory Judgments. *Psychological Science*, 22(6), 787–794. doi:10.1177/0956797611407929

Laird, N. M., & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, 963-974.

Lamotte, L. R. (1983). Fixed-, random-, and mixed-effects models. *In Encyclopedia of Statistical Sciences*. (S. Kotz, N. L. Johnson and C. B. Read, eds.) 3 137–141. Wiley, New York.

Littell, R.C. et al. (2006) SAS for Mixed Models. (2nd ed.), SAS Publishing

Masson, M. E., & Rotello, C. M. (2009). Sources of bias in the Goodman–Kruskal gamma coefficient measure of association: Implications for studies of metacognitive processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(2), 509.

Matvey, G., Dunlosky, J., & Guttentag, R. (2001). Fluency of retrieval at study affects judgments of learning (JOLs): An analytic or nonanalytic basis for JOLs?. *Memory & Cognition*, *29*(2), 222-233.

Matvey, G., Dunlosky, J., & Schwartz, B. (2006). The effects of categorical relatedness on judgments of learning (JOLs). *Memory*, 14(2), 253–261. doi:10.1080/09658210500216844

Maxwell, S. E., & Delaney, H. D. (2004). Designing experiments and analyzing data: A model comparison perspective (Vol. 1). Psychology Press.

Mazzoni, G., & Kirsch, I. (2002). Autobiographical memories and beliefs: A preliminary metacognitive model. In T. J. Perfect & B. L. Schwartz (Eds.), Applied metacognition (pp. 121–145). Cambridge, UK: Cambridge University Press.

Mazzoni, G., & Nelson, T. O. (1995). Judgments of learning are affected by the kind of encoding in ways that cannot be attributed to the level of recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(5), 1263.

McDaniel, M. A., & Waddill, P. J. (1990). Generation effects for context words: Implications for item-specific and multifactor theories. *Journal of Memory and language*, *29*(2), 201-211.

Meeter, M., & Nelson, T. O. (2003). Multiple study trials and judgments of learning. *Acta Psychologica*, *113*(2), 123-132.

Mendoza, J. L., Toothaker, L. E., & Nicewander, W. A. (1974). A Monte Carlo comparison of the univariate and multivariate methods for the groups by trials repeated measures design. *Multivariate Behavioral Research*, 9(2), 165-177.

Metcalfe, J., & Kornell, N (2003). The Dynamics of Learning and Allocation of Study Time to a Region of Proximal Learning. *Journal Of Experimental Psychology: General*, 132(4), 530-542. doi:10.1037/0096-3445.132.4.530

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

Metcalfe, J., Schwartz, B. L., & Joaquim, S. G. (1993). The cue-familiarity heuristic in metacognition. Journal of Experimental Psychology: Learning, Memory, and Cognition, 19(4), 851-861.

Mueller, M. L., Dunlosky, J., Tauber, S. K., & Rhodes, M. G. (2014). Journal of Memory and Language. *Journal of Memory and Language*, 70, 1–12. doi:10.1016/j.jml.2013.09.007

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*. doi:10.3758/s13423-012-0343-6

Mulligan, N. W., & Lozito, J. P. (2004). Self-generation and memory. *The psychology of learning and motivation: Advances in research and theory*, *45*, 175-214.

Nelson, T. O., & Dunlosky, J. (1991). When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: The "delayed-JOL effect". *Psychological Science*, *2*(4), 267-270.

Nelson, T. O. (1996). Consciousness and metacognition. *American psychologist*, 51(2), 102-116.

Nelson, T. O., Gerler, D., & Narens, L. (1984). Accuracy of feeling-of-knowing judgments for predicting perceptual identification and relearning. *Journal of Experimental Psychology: General*, *113*(2), 282-300.

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. *The psychology of learning and motivation*, *26*, 125-141.

Nelson, T. O., Narens, L., & Dunlosky, J. (2004). A Revised Methodology for Research on Metamemory: Pre-judgment Recall And Monitoring (PRAM). *Psychological Methods*, *9*(1), 53–69. doi:10.1037/1082-989X.9.1.53

Olejnik, S., & Algina, J. (2003). Generalized eta and omega squared statistics: measures of effect size for some common research designs. *Psychological methods*, *8*(4), 434-454.

Pinheiro, J. C., & Bates, D. M. (2000). Mixed effects models in S and S-PLUS. Springer.

Raaijmakers, G. (2003). A further look at the language-as-a- fixed-effect fallacy. *Canadian Journal of Experimental Psychology*, 57, 141–151.

Raaijmakers, J. G. W., & Shiffrin, R. M. (1980). SAM: A theory of probabilistic search

of associative memory. In G. H. Bower (Ed.), The psychology of learning and motivation: Advances in research and theory (Vol. 14, pp. 207–262). New York: Academic Press.

Raaijmakers, J. G., & Shiffrin, R. M. (1981). Search of associative memory. *Psychological review*, *88*(2), 93-132.

Rabinowitz, J. C., Ackerman, B. P., Craik, F. I. M., & Hinchley, J. L. (1982). Aging and metamemory: The roles of relatedness and imagery. *Journal of Gerontology*, 37, 688-695.

Raudenbush, S.W., & Bryk, A.S. Hierarchical Linear Models. Applications and Data Analysis Methods. Newbury Park, CA: Sage, 2nd ed., 2002.

Reder, L. M., & Ritter, F. E. (1992). What determines initial feeling of knowing? Familiarity with question terms, not with the answer. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*(3), 435-451.

Reed, A. V. (1973). Speed–accuracy trade-off in recognition memory. *Science*, 181, 574-576.

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. doi:10.1037/a0013684

Rouder, J. N., & Lu J. (2005).An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin & Review*. 12, 573-604.

Satterthwaite, F. F. (1941). Synthesis of variance. *Psychometrika* 6, 309-316.

Schabenberger, O. (2005). Introducing the GLIMMIX procedure for generalized linear mixed models. *SUGI 30 Proceedings*, 196-30.

Scheck, P., & Nelson, T. O (2005). Lack of Pervasiveness of the Underconfidence-With-Practice Effect: Boundary Conditions and an Explanation Via Anchoring. *Journal Of Experimental Psychology: General*, 134(1), 124-128. doi:10.1037/0096-3445.134.1.124.

Schwartz, S. H. (1994). Are there universal aspects in the structure and contents of human values?. *Journal of social issues*, *50*(4), 19-45.

Searle, S. R., Caesella, G. and McCulloch, C. E. (1992). Variance Components. Wiley, New York

Self, S.G. and Liang, K.Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*. 82, 605-610.

Serra, M. J., & Dunlosky, J. (2005). Does retrieval fluency contribute to the underconfidence-with-practice effect?. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(6), 1258.

Silvapulle, M.J. and Silvapulle, P. (1995). A score test against one-sided alternatives. *Journal of the American Statistical Association*. 90, 342-349.

Shaughnessy, J. J. (1981). Memory monitoring accuracy and modification of rehearsal strategies. *Journal of Verbal Learning and Verbal Behavior*, *20*(2), 216-230.

Singer, J. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and residual growth models. *Journal of Educational and Behavioral Statistics*, 23, 323–355.

Snijders, T. A. B., & Bosker, R. J. (1999). Multilevel analysis: An introduction to basic and advanced multilevel modeling. London: Sage.

Son, L. K., & Metcalfe, J. (2000). Metacognitive and control strategies in study-time allocation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(1), 204-221.

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), 315-316.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4), 583-639.

Thiede, K. W., & Dunlosky, J. (1999). Toward a general model of self-regulated study: An analysis of selection of items for study and self-paced study time. *Journal of experimental psychology: Learning, Memory, and Cognition*, *25*(4), 1024.

Thomas, R. P., Dougherty, M. R., Sprenger, A. M., & Harbison, J. I. (2008). Diagnostic hypothesis generation and human judgment. *Psychological Review*, 115(1), 155–185. doi:10.1037/0033-295X.115.1.155

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive psychology*, *5*(2), 207-232.

Underwood, B. J. (1966). Individual and group predictions of item difficulty for free-recall learning. *Journal of Experimental Psychology*, 71, 673-679.

Undorf, M., & Erdfelder, E. (2011). Judgments of learning reflect encoding fluency: Conclusive evidence for the ease-of-processing hypothesis. *Journal of*

Experimental Psychology: Learning, Memory, and Cognition, *37*(5), 1264-1269.

Verbeke, G. and Molenberghs, G. (2003). The use of score tests for inference on variance components. *Biometrics*. 59, 254-262.

Wixted, J. T. (2007). Dual-process theory and signal-detection theory of recognition memory. *Psychological review*, *114*(1), 152-176.

Yonelinas, A. P., Dobbins, I., Szymanski, M. D., Dhaliwal, H. S., & King, L. (1996). Signal-detection, threshold, and dual-process models of recognition memory: ROCs and conscious recollection. *Consciousness and cognition*, *5*(4), 418-441.

Tables

Table 1
Goodness-of-Fit Statistics and LR x^2 Tests of Random Effects for Predicting Recall in Experiment 1

Model	Covariance Structure	AIC	BIC	-2LL	x^2 (Diff)	*df* (Diff)	*df* (Total)
1. Unconditioned	—	2233	2238	2231	—	—	1
2. Add random effect:							
Intercept	—	2055	2067	2051	180*	1	2
3. Add fixed effect:							
JOL (WS)							
JOL (BS)	—	1997	2019	1989	62*	2	4
4. Add random effect:							
JOL (WS)	Unstructured	2000	2033	1988	1	2	6
5. Add fixed effect:							
Episodic retrieval	Unstructured	1999	2039	1985	2	1	7
6. Add interactions							
All fixed factors	Unstructured	2003	2064	1981	4	4	11

Note. The LR tests are only valid for random effects. 2LL = -2 Log Likelihood (fit function); BIC = Bayesian Information Criterion; AIC = Akaike's Information Criterion; LR = likelihood ratio; WS = within subject; BS = between subject; Episodic retreival of reference item
*$P < 0.0001$

Table 2
Judgments of Learning Predicting Recall in Experiment 1

Effect	*df* (Residual)	*F*	*p*	Estimate	SE
Intercept	—	—	—	-1.3740	0.1350
Main effects					
JOL (WS)	31.96	40.63	<.0001	0.0245	0.0034
JOL (BS)	60.54	0.48	0.4914	0.0048	0.0070
Episodic retrieval	56.78	2.43	0.1244	0.2185	0.1350
Interactions					
JOL (WS) x JOL (BS)	43.9	1.24	0.2706	-0.0002	0.0002
Episodic retrieval x JOL (WS)	31.96	0.04	0.8339	-0.0006	0.0034
Episodic retrieval x JOL (BS)	60.54	2.74	0.1032	0.0120	0.0070
Episodic retrieval x JOL (WS) x JOL (BS)	43.9	0.1	0.7578	0.0001	0.0002

Note. Estimated using restricted maximum likelihood. Episodic retreival of reference item;
WS = within subject; BS = between subject. Estimates represent effect coding

Table 3

Goodness-of-Fit Statistics and LR x^2 Tests of Random Effects For Predicting Transfer Item JOLs in Experiment 1

Model	Covariance Structure	Residual	AIC	BIC	-2LL	x^2 (Diff)	*df* (Diff)	*df* (Total)
1. Unconditioned	—	—	18767	18778	18763	—	—	1
2. Add random effect:								
Intercept	—	361	17523	17540	17517	1245*	2	3
3. Add fixed effect:								
Reference JOL (WS)								
Reference JOL (BS)	—	357	17430	17458	17420	97*	2	5
4. Add random effect:								
Reference JOL (WS)	Unstructured	343	17398	17437	17384	36*	2	7
5. Add fixed effect:								
Episodic retrieval	Unstructured	343	17399	17444	17383	1	1	8
6. Add interactions								
All fixed factors	Unstructured	343	17396	17463	17372	11**	4	12
7. Change variance structure								
Intercept								
Reference JOL (WS)	Independent	344	—	—	—	-5***	-1	11

Note. The LR tests are only valid for random effects. 2LL = -2 Log Likelihood (fit function);
BIC = Bayesian Information Criterion; AIC = Akaike's Information Criterion; LR = likelihood ratio
Episodic retreival of reference item; *P < 0.0001; **P = 0.0212
*** Estimated with restricted maximum likelihood; p = 0.0265

Table 4
Reference Item JOLs Predicting Transfer Item JOLs in Experiment 1

Effect	*df* (Residual)	*F*	*p*	Estimate	SE
Intercept	—	—	—	34.6106	1.4031
Main effects					
Reference JOL (WS)	57.91	9.92	0.0026*	0.0945	0.0288
Reference JOL (BS)	62	133.1	<.0001*	0.8732	0.0734
Episodic retrieval	62	0.16	0.6912	0.5777	1.4031
Interactions					
Reference JOL (WS) x Reference JOL (BS)	76.4	0.71	0.4007	0.0015	0.0017
ER x Reference JOL (WS)	57.91	6.66	0.0124*	0.0772	0.0288
ER x Reference JOL (BS)	62	2.05	0.1570	0.1084	0.0734
ER x RI JOL (WS) x RI JOL (BS)	76.4	5.03	0.0278*	0.0039	0.0017

Note. Estimated using restricted maximum likelihood. WS = within subject; BS = between subject.
ER = episodic retrieval of reference item; RI = reference item; Estimates represent effect coding;
* Significant effect

Table 5

Goodness-of-Fit Statistics and LR x^2 Tests of Random Effects For Predicting Transfer Item JOLs In Experiment 1

Model	Covariance Structure	Residual	AIC	BIC	-2LL	x^2 (Diff)	*df* (Diff)	*df* (Total)
1. Unconditioned	—	—	18767	18778	18763	—	—	1
2. Add random effect:								
Reference recall	Unstructured	326	17386	17431	17370	1393*	4	5
3. Add fixed effect:								
Reference recall	Unstructured	326	17398	17411	17386	15*	1	6
4. Add fixed effect:								
Episodic retrieval	Unstructured	326	17400	17414	17386	0	1	7
5. Add interactions								
All fixed factors	Unstructured	326	17385	17403	17369	17*	1	8
6. Change variance structure								
Reference recall	Homogeneity	327	—	—	—	-2.9**	-1	7

Note. The LR tests are only valid for random effects. 2LL = -2 Log Likelihood (fit function);
BIC = Bayesian Information Criterion; AIC = Akaike's Information Criterion; LR = likelihood ratio
*P < 0.0001; ** Estimated with restricted maximum likelihood; p = 0.0887

Table 6
Reference item Recall Predicting Transfer item JOLs in Experiment 1

Effect	*df* (Residual)	*F*	*p*	Estimate	SE
Intercept	—	—	—	35.4469	2.4629
Main effects					
Reference recall	64.6	21.79	<.0001*	3.6068	0.7588
Episodic retrieval	63.17	0.37	0.5462	1.521	2.4629
Interactions					
Reference recall x Episodic retrieval	64.6	17.26	<.0001*	3.2106	0.7588

Note. Estimated using restricted maximum likelihood. Episodic reference of reference item
WS = within subject; BS = between subject. Estimates represent effect coding;
* Significant effect

Table 7

Goodness-of-Fit Statistics and LR x^2 Tests of Random Effects For Predicting Recall In Experiment 2

Model	Covariance Structure	Residual	AIC	BIC	-2LL	x^2 (Diff)	*df* (Diff)	*df* (Total)
1. Unconditioned	—	—	4183	4189	4181	—	—	1
2. Add random effect:								
Episodic retrieval	Unstructured	—	3788	3813	3780	400*	4	5
3. Add fixed effect:								
JOL (WS)								
JOL (BS)	Unstructured	—	3641	3678	3629	151*	1	6
4. Add random effect:								
JOL (WS)	Unstructured	—	3642	3699	3624	5	3	9
5. Add fixed effect:								
Episodic retrieval	Unstructured	—	3644	3707	3624	0	1	10
6. Add interactions								
All fixed factors	Unstructured	—	3648	3736	3620	4	4	14

Note. The LR tests are only valid for random effects. 2LL = -2 Log Likelihood (fit function);
BIC = Bayesian Information Criterion; AIC = Akaike's Information Criterion; LR = likelihood ratio;
WS = within subject; BS = between subject; Episodic retreival of reference item

*$P < 0.0001$

Table 8
Judgments of Learning Predicting Recall in Experiment 2

Effect	*df* (Residual)	*F*	*p*	Estimate	SE
Intercept	—	—	—	-1.5430	0.1165
Main effects					
JOL (WS)	68.52	77.62	<.0001	0.0252	0.0028
JOL (BS)	94.89	5.19	0.0249	-0.0106	0.0045
Episodic retrieval	98.53	0.69	0.4065	0.0282	0.0446
Interactions					
JOL (WS) x JOL (BS)	98.1	1.03	0.3118	-0.0002	0.0001
Episodic retrieval x JOL (WS)	3872	0.61	0.4356	-0.0017	0.0023
Episodic retrieval x JOL (BS)	106.6	1.41	0.2373	0.0020	0.0018
Episodic retrieval x JOL (WS) x JOL (BS)	3872	0.13	0.7178	0.0001	0.0001

Note. Estimated using restricted maximum likelihood. Episodic retreival of reference item; WS = within subject; BS = between subject. Estimates represent effect coding

Table 9

Goodness-of-Fit Statistics and LR x^2 Tests of Random Effects For Predicting Transfer Item JOLs in Experiment 2

Model	Covariance Structure	Residual	AIC	BIC	-2LL	x^2 (Diff)	*df* (Diff)	*df* (Total)
1. Unconditioned	—	—	38206	38218	38202	—	—	1
2. Add random effect:								
Episodic retrieval	Unstructured	425	34952	34983	34942	3259*	4	5
3. Add fixed effect:								
Reference JOL (WS)								
Reference JOL (BS)	Unstructured	415	34841	34885	34827	114*	2	7
4. Add random effect:								
Reference JOL (WS)	Unstructured	381	34708	34771	34688	138*	3	10
5. Add fixed effect:								
Episodic retrieval	Unstructured	381	34706	34775	34684	5**	1	11
6. Add interactions								
All fixed factors	Unstructured	384	34658	34752	34628	56*	4	15
7. Change variance structure								
Episodic retrieval								
Reference JOL (WS)	Independent	388	—	—	—	-12***	-2	13

Note. The LR tests are only valid for random effects. 2LL = -2 Log Likelihood (fit function);
BIC = Bayesian Information Criterion; AIC = Akaike's Information Criterion; LR = likelihood ratio
Episodic retreival of reference item; *P < 0.0001; **P = 0.0323
*** Estimated with restricted maximum likelihood; p = 0.0265

Table 10
Reference item JOLs Predicting Transfer item JOLs in Experiment 2

Effect	*df* (Residual)	*F*	*p*	Estimate	SE
Intercept	—	—	—	53.967	2.380
Main effects					
Reference JOL (WS)	99.91	24.17	<.0001*	0.096	0.019
Reference JOL (BS)	95.11	23.24	<.0001*	0.664	0.136
Episodic retrieval	95.1	16.39	0.0001*	-1.524	0.372
Interactions					
Reference JOL (WS) x Reference JOL (BS)	110.7	0.19	0.6639	0.001	0.001
ER x Reference JOL (WS)	3668	28.13	<.0001*	0.055	0.010
ER x Reference JOL (BS)	94.93	0.14	0.7051	0.008	0.021
ER x RI JOL (WS) x RI JOL (BS)	3689	5.11	0.0238*	0.001	0.001

Note. Estimated using restricted maximum likelihood. WS = within subject; BS = between subject.
ER = episodic retrieval of reference item; RI = reference item; Estimates represent effect coding;
* Significant effect

Table 11
Experiment 3: JOL, Recall, Calibration, and Gamma Means

Item Type Episodic retrievel	Recall	JOL	Calibration	Gamma
Reference item				
Likely retrieval of reference item	48.2 (3.2)	56.4 (2.5)	8.2 (2.1)	0.80 (0.03)
Unlikely retrieval of reference item	56.1 (3.0)	62.4 (2.5)	6.3 (3.3)	0.27 (0.05)
Transfer item				
Likely retrieval of reference item	45.0 (2.9)	64.3 (2.7)	19.3 (3.5)	0.25 (0.05)
Unlikely retrieval of reference item	44.0 (3.0)	62.5 (2.8)	18.5 (3.8)	0.35 (0.05)

Note. Mean standard errors are reported in parenthesis; Recall percent

Table 12
Reference item JOLs Predicting Transfer item JOLs in Experiment 3

Effect	*df* (Residual)	*F*	*p*	Estimate	SE
Intercept	—	—	—	63.336	1.855
Main effects					
JOL (WS)	70.54	50.82	<.0001	0.215	0.030
JOL (BS)	69.03	76.96	<.0001	0.826	0.093
Episodic retrieval	68.5	13.21	0.0005	1.466	0.423
Interactions					
JOL (WS) x JOL (BS)	84.86	1.37	0.2447	-0.002	0.002
Episodic retrieval x JOL (WS)	2754	0.96	0.3277	-0.016	0.016
Episodic retrieval x JOL (BS)	67.58	0.45	0.5032	-0.014	0.021
Episodic retrieval x JOL (WS) x JOL (BS)	2738	0.17	0.6792	0.000	0.001

Note. Estimated using restricted maximum likelihood.
WS = within subject; BS = between subject. Estimates represent effect coding

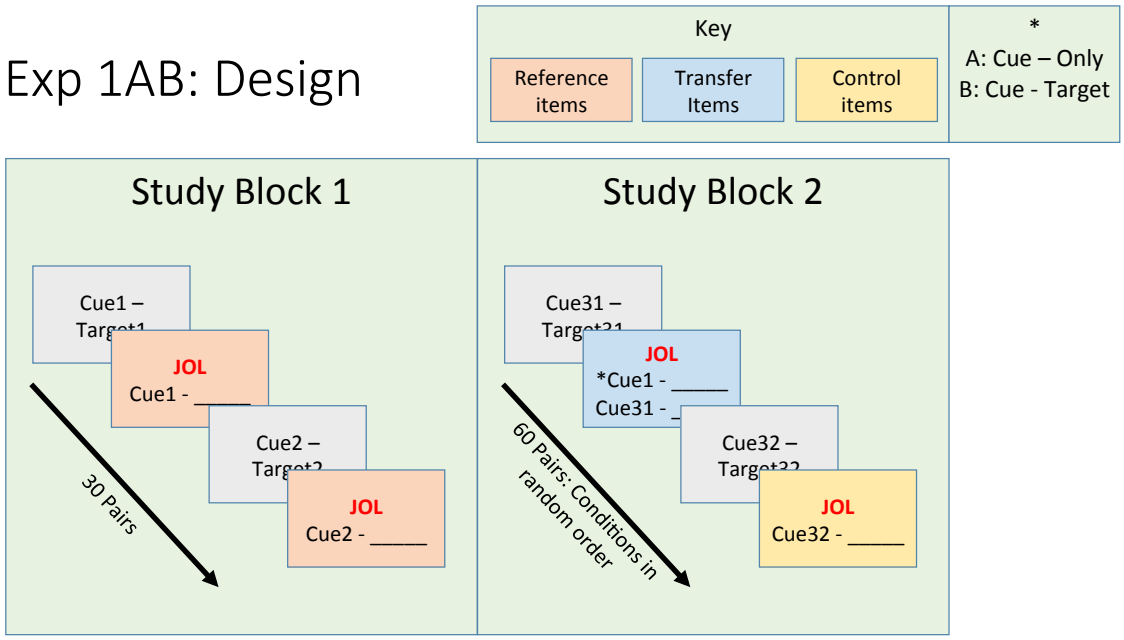

Figure 1. The design of Experiment 1. The first study block contains the reference items, which received immediate judgments of learning. The second block contains both control and transfer items. *Notice that the reference item was presented either as cue-only or cue-target, depending on the between-subject condition.

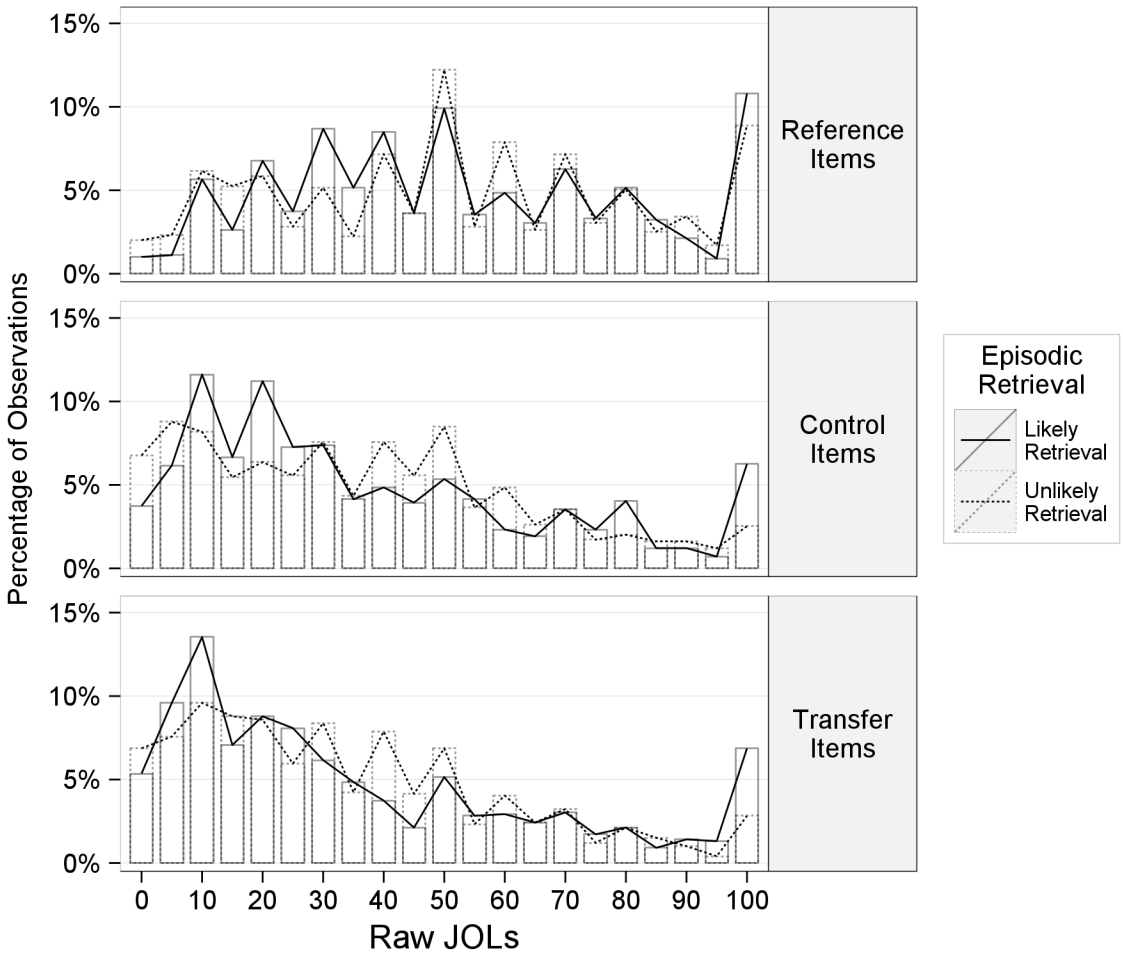

Figure 2. Experiment 1: A histogram displaying the percentage of observations within an item type (top: reference item; middle: control item; bottom: transfer item) given a particular raw judgments of learning. The superimposed lines indicate the between-subject episodic retrieval conditions (likely retrieval of the reference item, unlikely retrieval of the reference item).

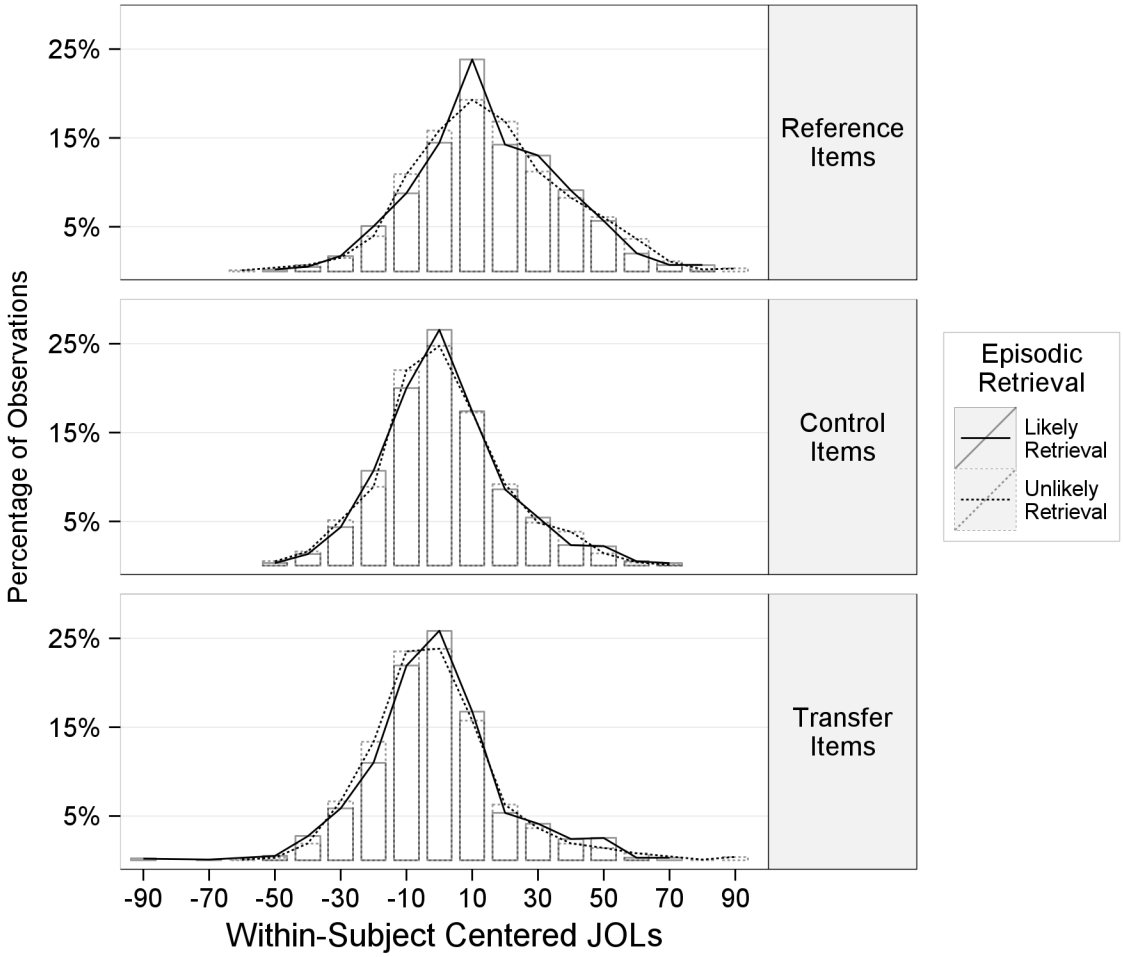

Figure 3. Experiment 1: A histogram displaying the percentage of observations within an item type (top: reference item; middle: control item; bottom: transfer item) given a particular within-subject centered judgments of learning. The superimposed lines indicate the between-subject episodic retrieval conditions (likely retrieval of the reference item, unlikely retrieval of the reference item).

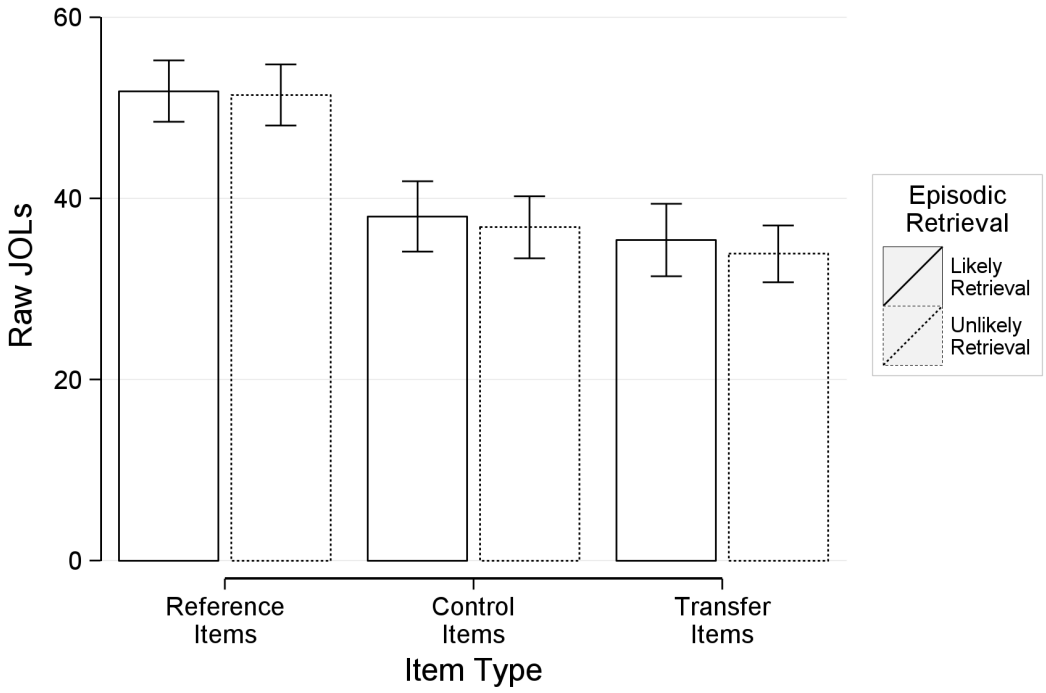

Figure 4. Experiment 1: Mean judgments of learning as a function of item type (reference, control, transfer) and episodic retrieval (likely retrieval of the reference item, unlikely retrieval of the reference item). The error bars represent standard errors of the mean.

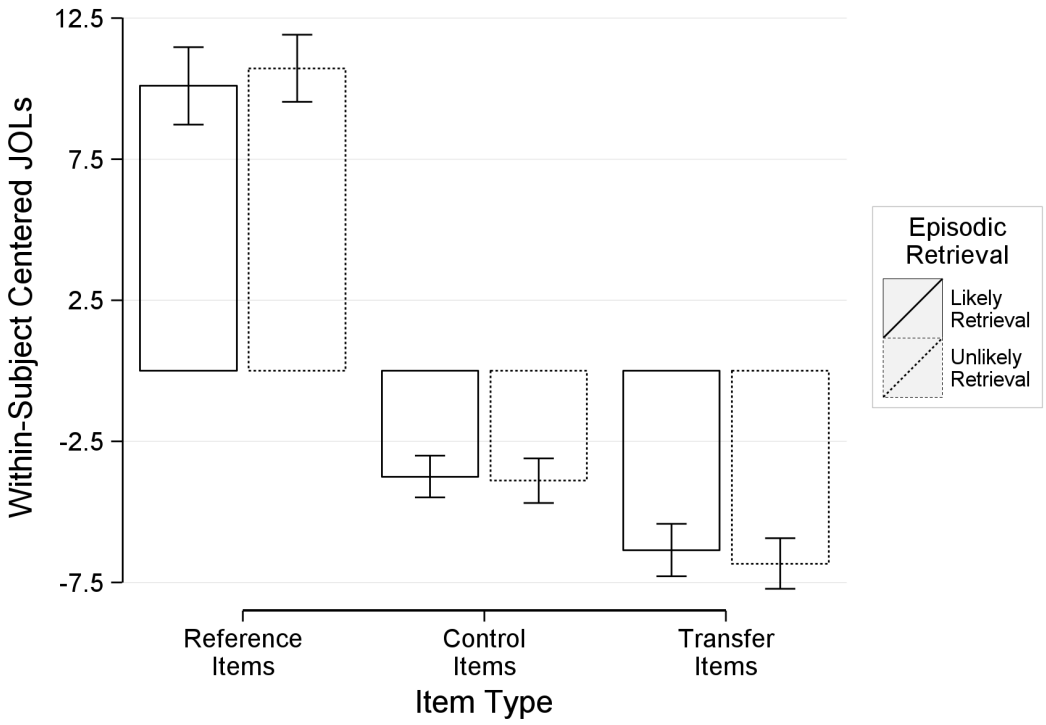

Figure 5. Experiment 1: Within-subject centered JOLs as a function of item condition (reference, control, transfer) and episodic retrieval (likely retrieval of the reference item, unlikely retrieval of the reference item). The values represent deflections from a subject's mean judgments of learning ratings and better represent within-subject differences (i.e., item type). The error bars represent standard errors of the within-subject means.

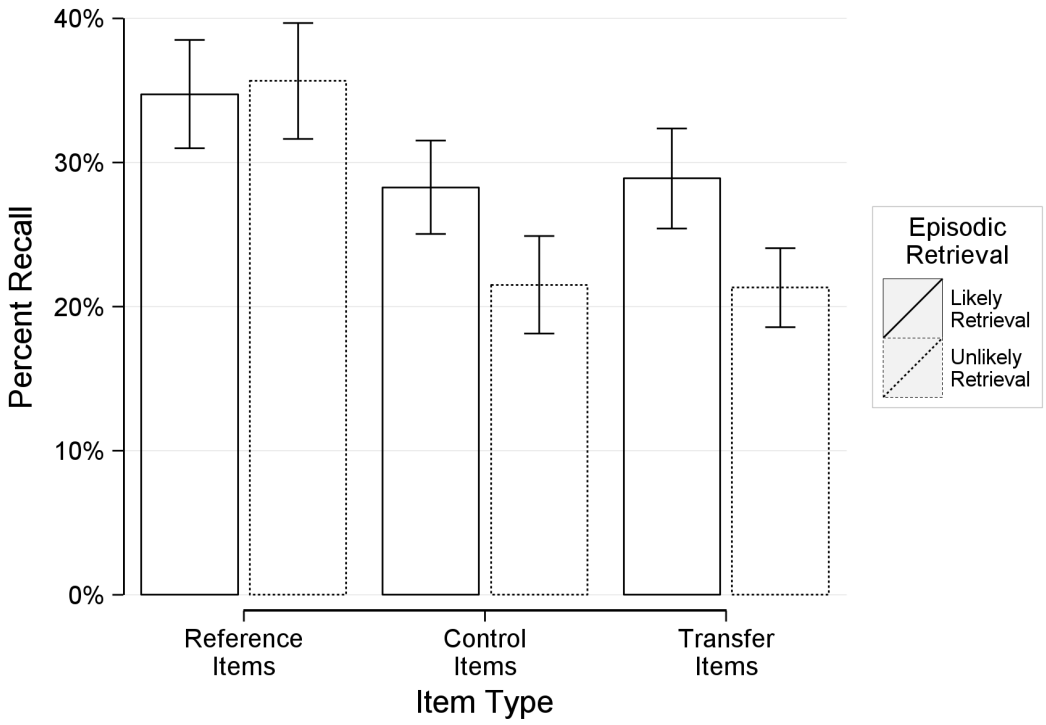

Figure 6. Experiment 1: Percent recall as a function of item condition (reference, control, transfer) and episodic retrieval (likely retrieval of the reference item, unlikely retrieval of the reference item). The error bars represent standard errors of the means.

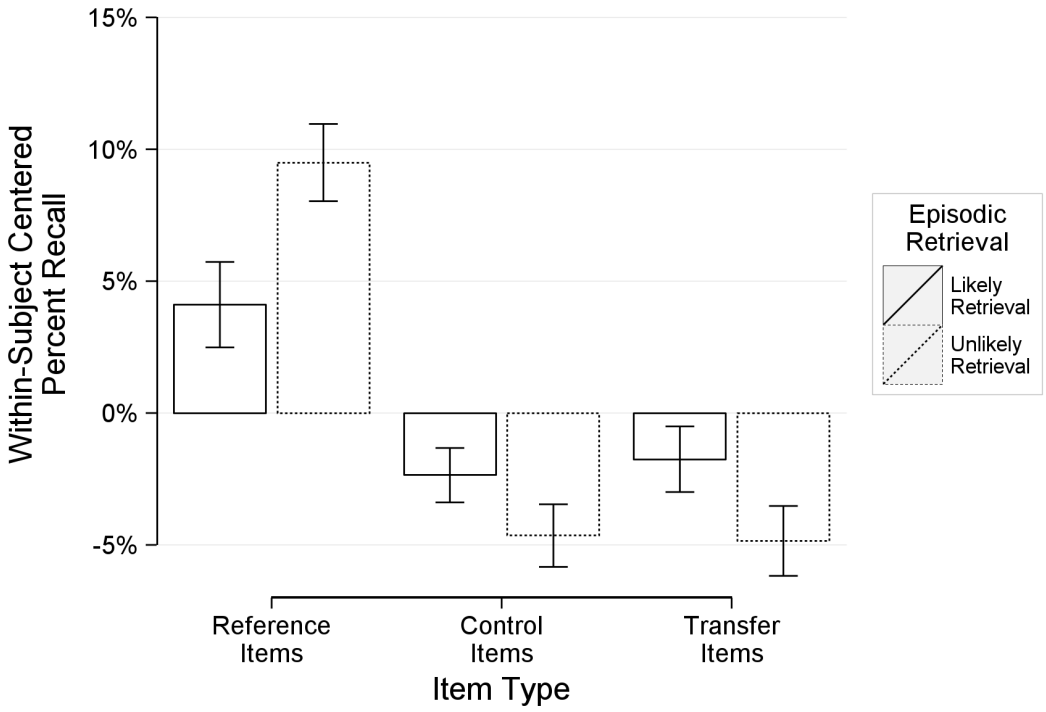

Figure 7. Experiment 1: Within-subject centered percent recall as a function of item condition (reference, control, transfer) and episodic retrieval (likely retrieval of the reference item, unlikely retrieval of the reference item). The values represent deflections from a subject's mean percent recall and better represent within-subject differences (i.e., item type). The error bars represent standard errors of the within-subject means.

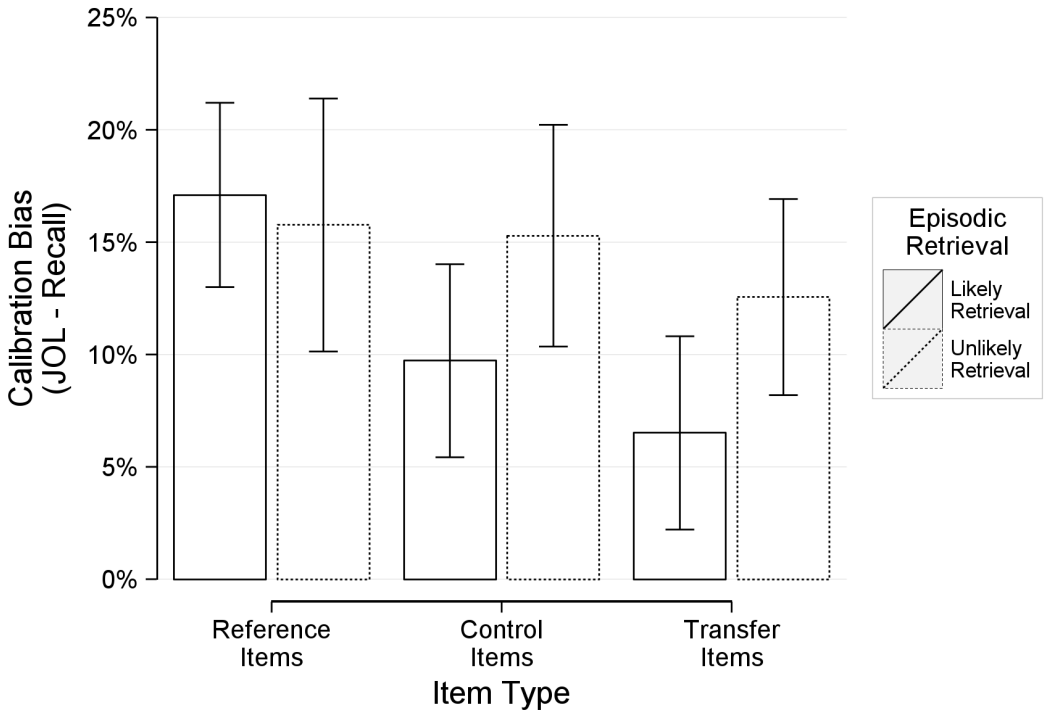

Figure 8. Experiment 1: Mean calibration bias as a function of item condition (reference, control, transfer) and episodic retrieval (likely retrieval of the reference item, unlikely retrieval of the reference item). Higher values indicate overconfidence. The error bars represent standard errors of the means.

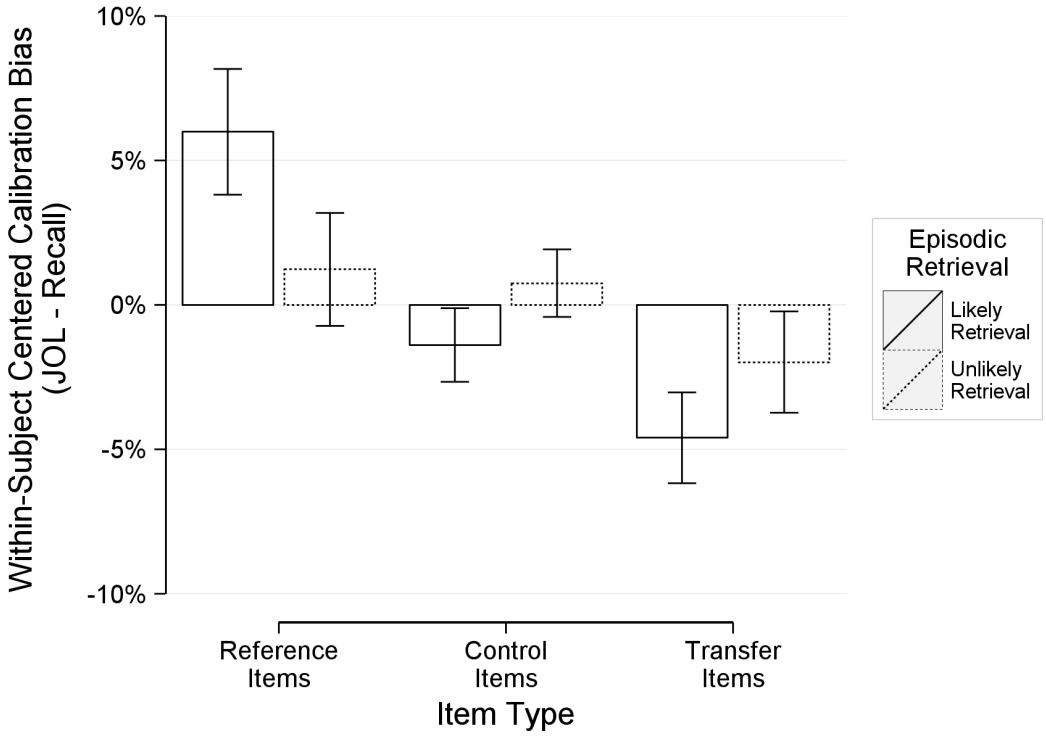

Figure 9. Experiment 1: Within-subject centered calibration bias as a function of item condition (reference, control, transfer) and episodic retrieval (likely retrieval of the reference item, unlikely retrieval of the reference item). The values represent deflections from a subject's mean calibration bias and better represent within-subject differences (i.e., item type). The error bars represent standard errors of the within-subject means.

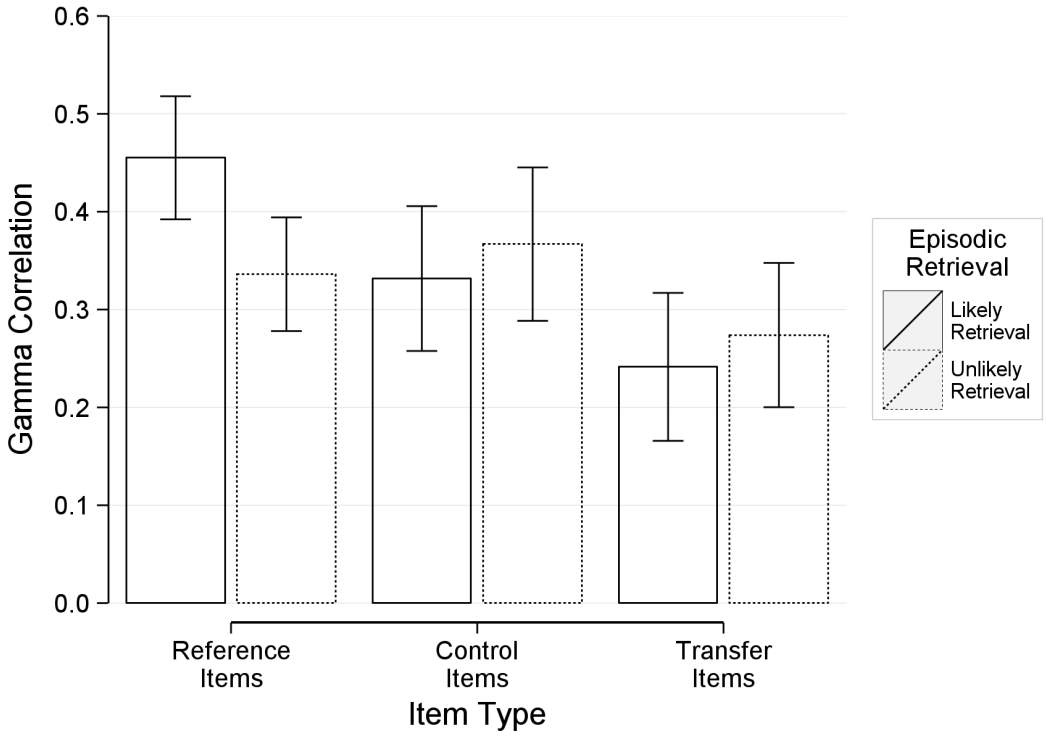

Figure 10. Experiment 1: Mean gamma correlation as a function of item condition (reference, control, transfer) and episodic retrieval (likely retrieval of the reference item, unlikely retrieval of the reference item). Higher values indicate better resolution. The error bars represent standard errors of the means.

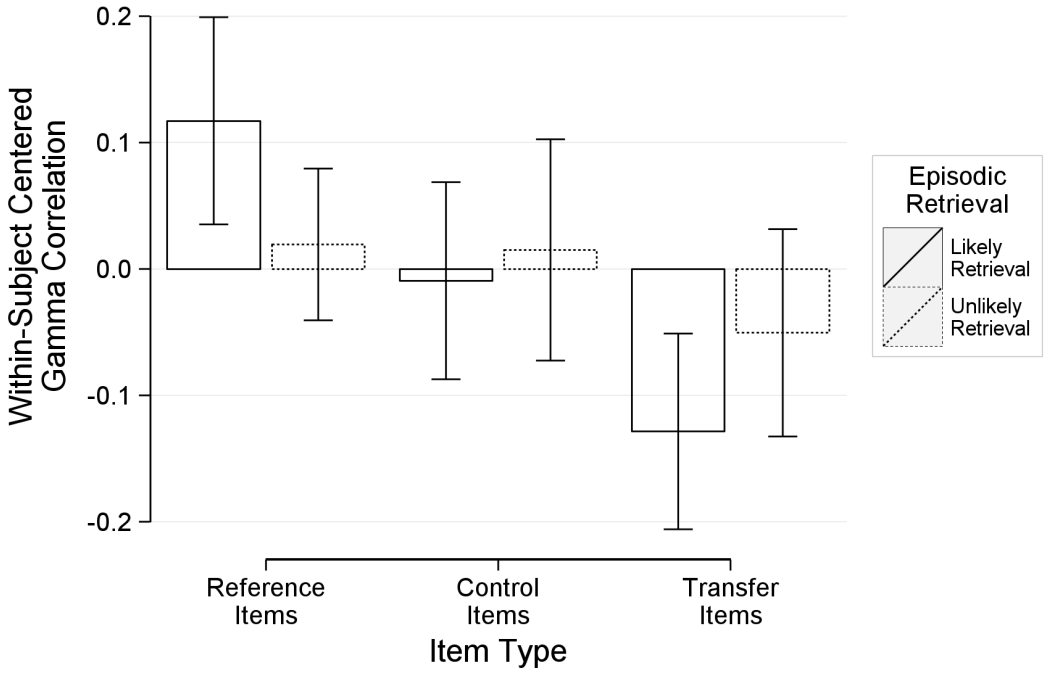

Figure 11. Experiment 1: Within-subject centered gamma correlation as a function of item condition (reference, control, transfer) and episodic retrieval (likely retrieval of the reference item, unlikely retrieval of the reference item). The values represent deflections from a subject's mean gamma correlation and better represent within-subject differences (i.e., item type). The error bars represent standard errors of the within-subject means.

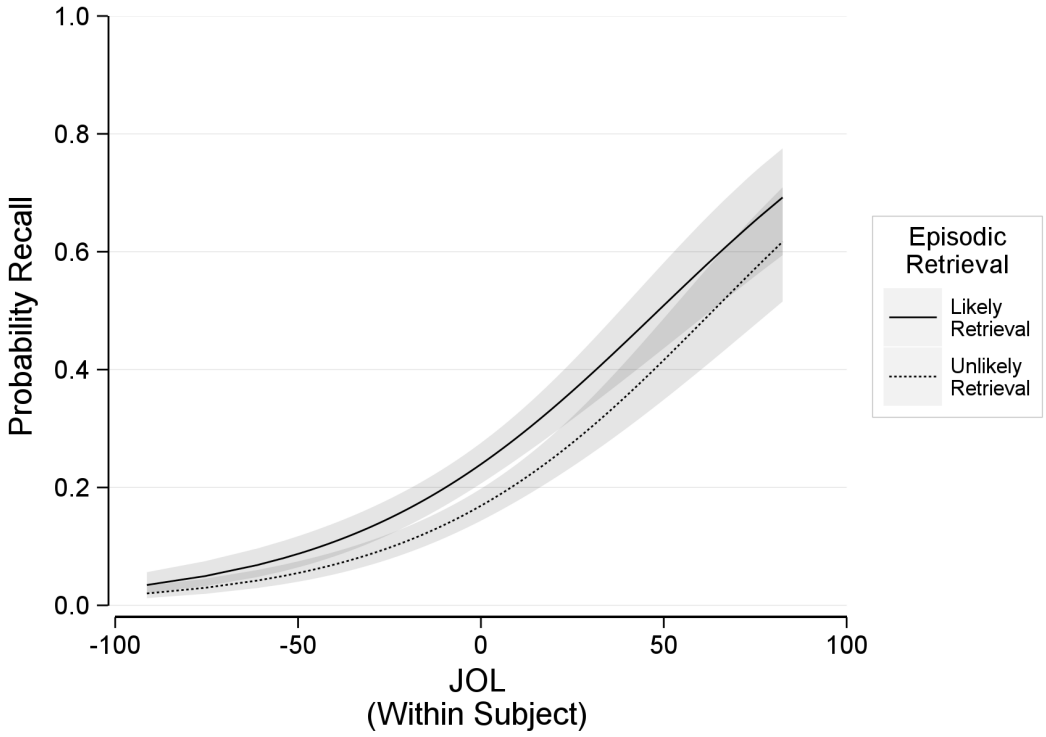

Figure 12. Experiment 1: The probability of recalling an item as a function of within-subject JOLs and episodic retrieval (likely retrieval of the reference item, unlikely retrieval of the reference item) at the sample mean judgments of learning. The error ribbon represents the combined standard errors of the fixed effects.

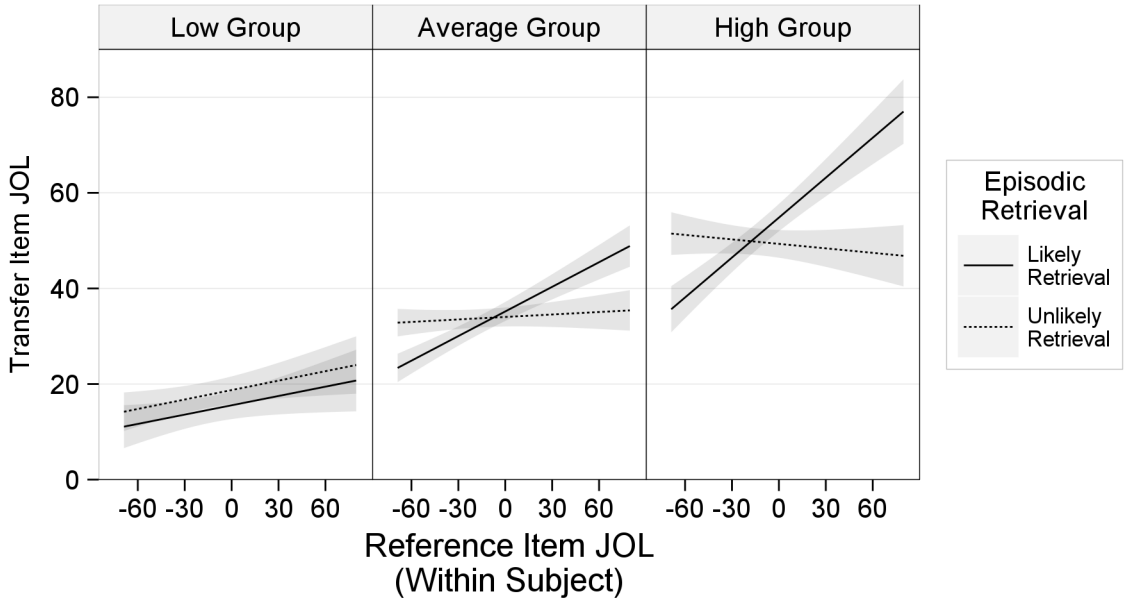

Figure 13. Experiment 1: Transfer item judgments of learning as a function of within-subject reference item judgments of learning and episodic retrieval (likely retrieval of the reference item, unlikely retrieval of the reference item). The three panels represent display the relationship at different levels of between-subject reference item JOLs. The low group is one standard deviation below the mean, the average group is at the mean, and the high group is one standard deviation above the mean. The error ribbon represents the combined standard errors of the fixed effects.

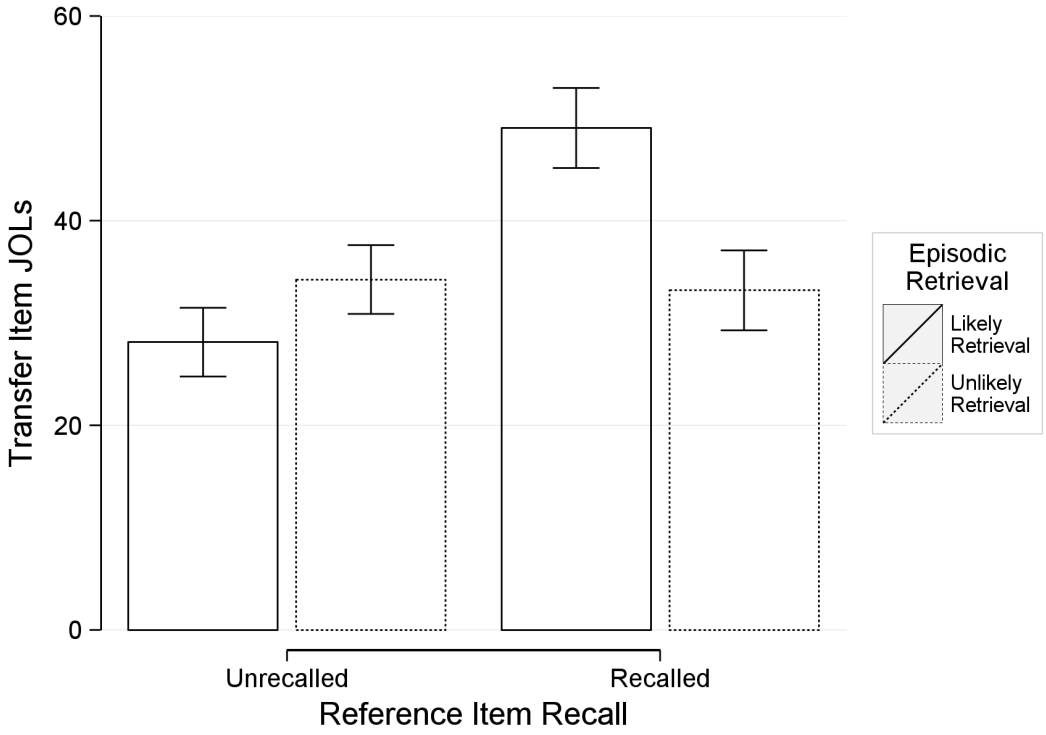

Figure 14. Experiment 1: Transfer item judgments of learning as a function of reference item final recall and episodic retrieval (likely retrieval of the reference item, unlikely retrieval of the reference item). The error bars represent the combined standard errors of the fixed effects.

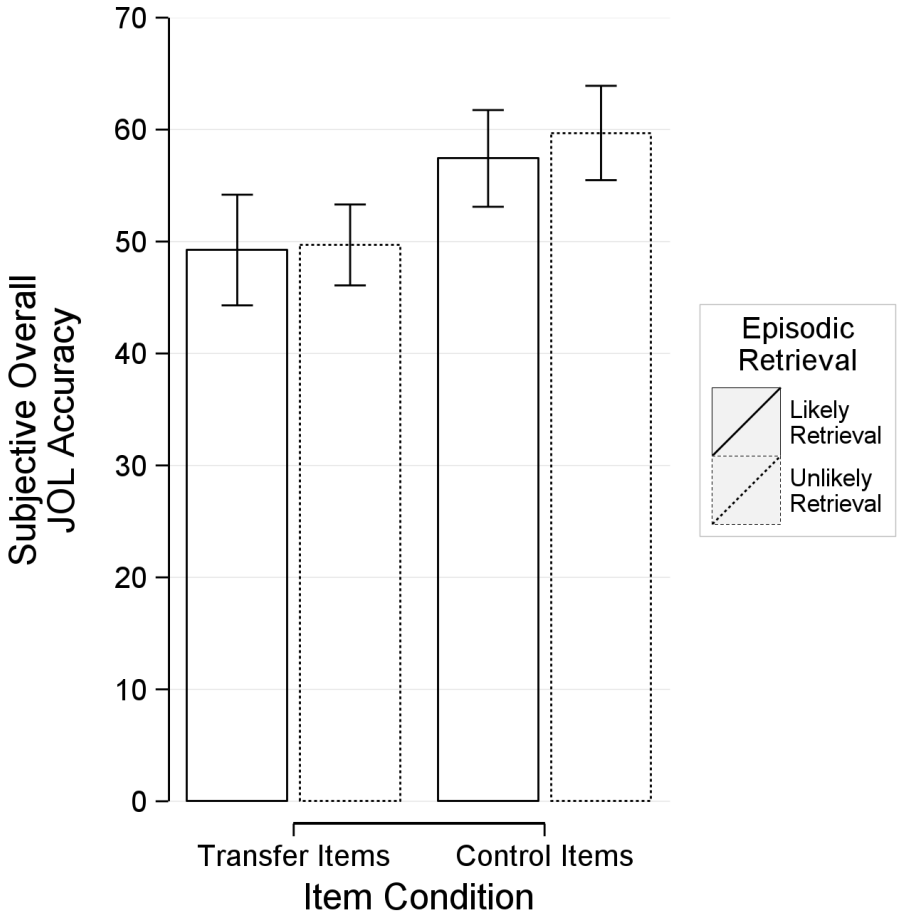

Figure 15. Experiment 1: The mean subjective accuracy of participants' judgments of learning as a function of item type and episodic retrieval (likely retrieval of the reference item, unlikely retrieval of the reference item). The error bars represent the standard error of the means.

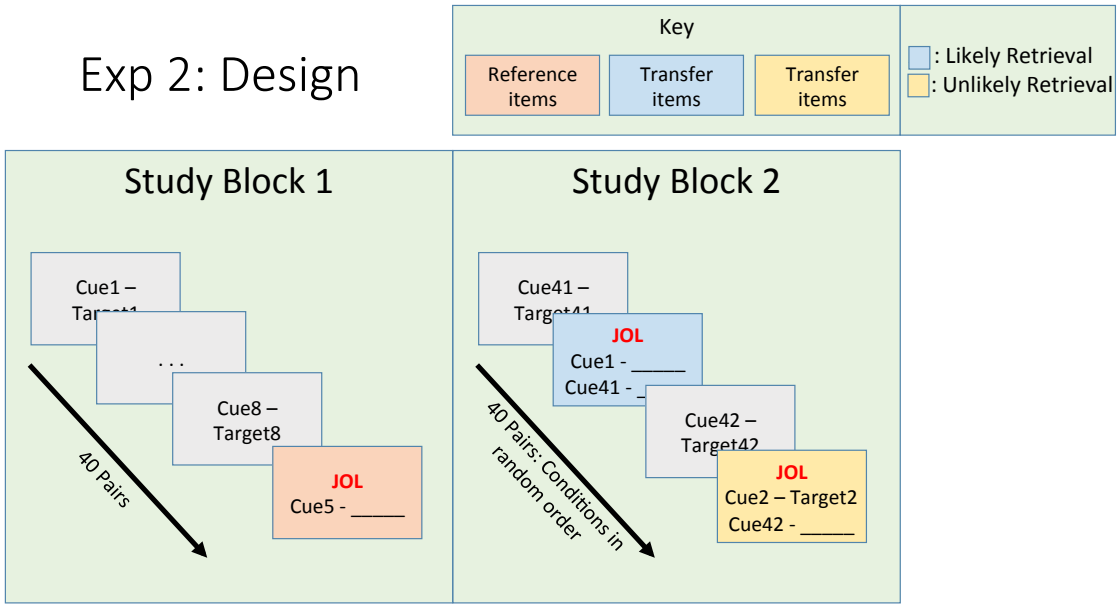

Figure 16. The design of Experiment 2. The first study block contains the reference items, which received delayed judgments of learning, delayed by 8 items. The second block contains two sets of transfer items. In one set, the reference item is presented in cue-only format and promotes episodic retrieval. In the other set, the reference item is presented in cue-target format and episodic retrieval is unlikely to occur.

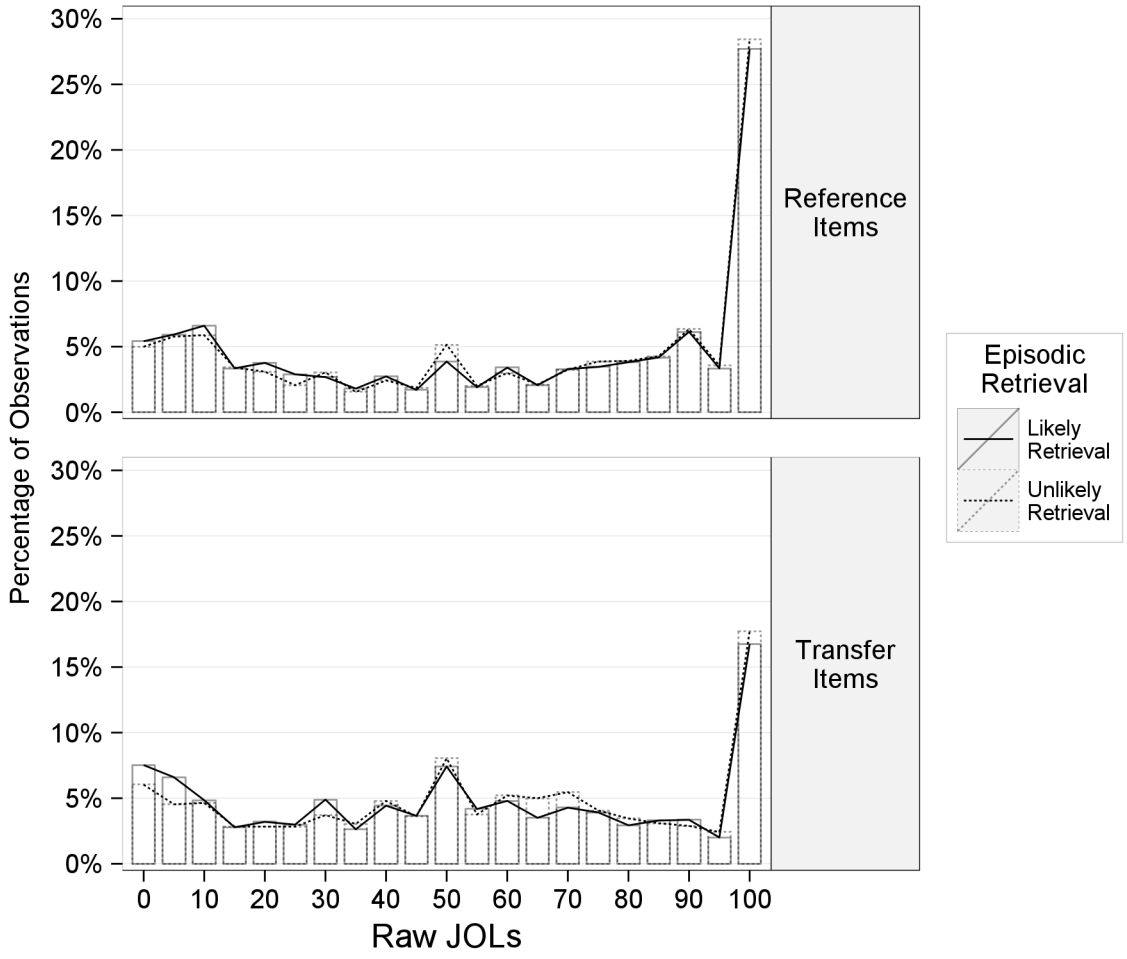

Figure 17. Experiment 2: A histogram displaying the percentage of observations within an item type (top: reference item; bottom: transfer item) given a particular raw judgment of learning. The superimposed lines indicate the episodic retrieval conditions (likely retrieval of the reference item, unlikely retrieval of the reference item).

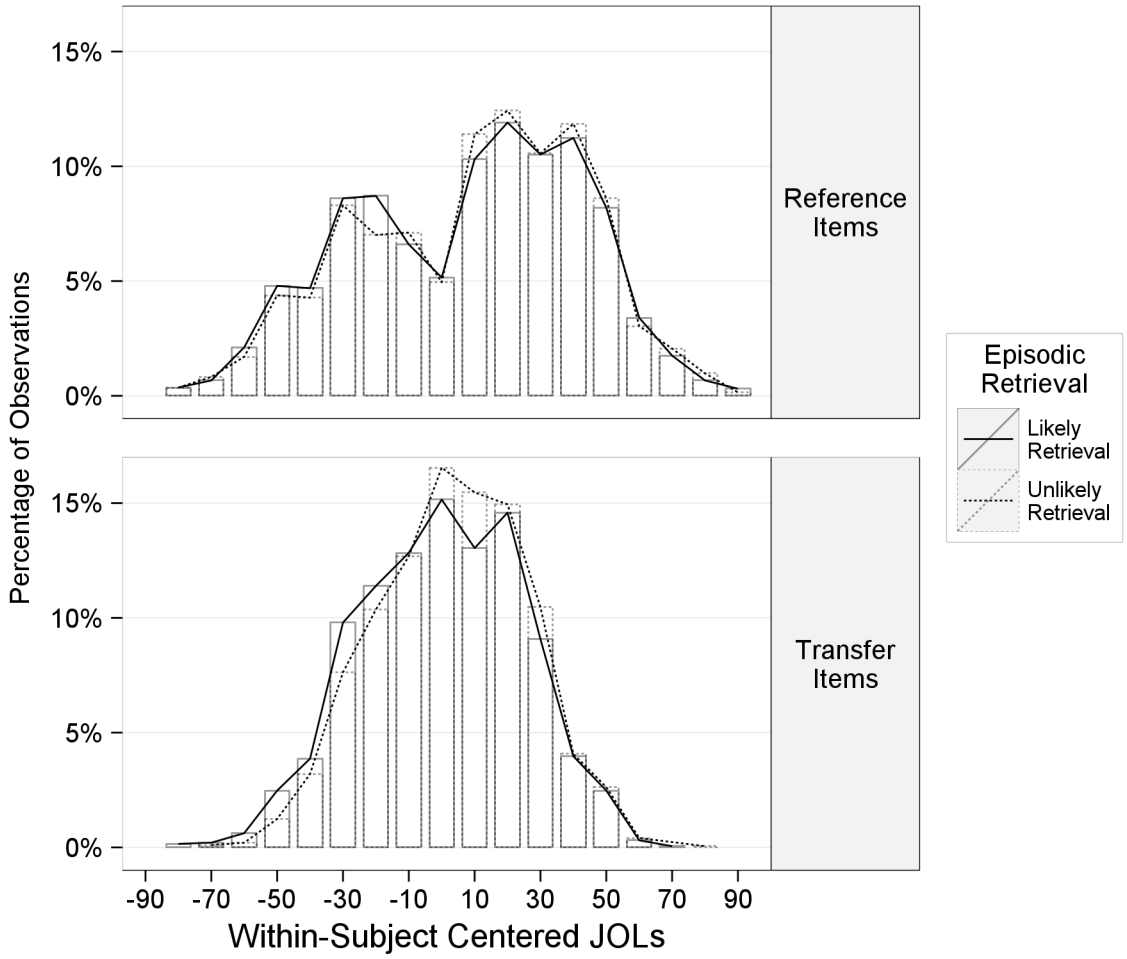

Figure 18. Experiment 2: A histogram displaying the percentage of observations within an item type (top: reference item; bottom: transfer item) given a particular within-subject centered judgment of learning. The superimposed lines indicate the episodic retrieval conditions (likely retrieval of the reference item, unlikely retrieval of the reference item).

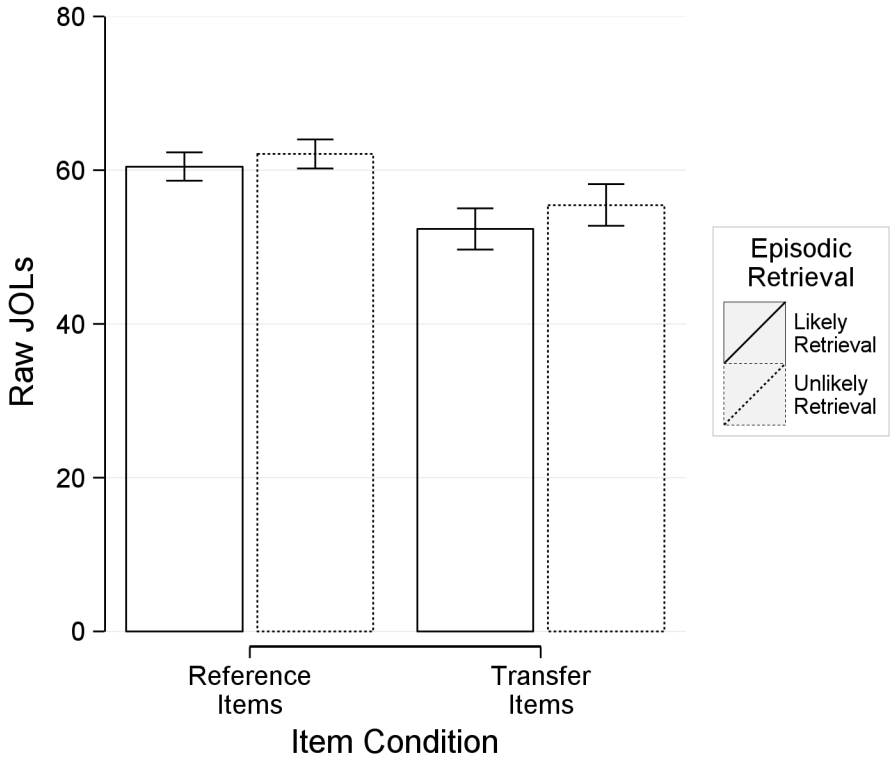

Figure 19. Experiment 2: Mean judgments of learning as a function of item type (reference, transfer) and episodic retrieval (likely retrieval of the reference item, unlikely retrieval of the reference item). The error bars represent standard errors of the mean.

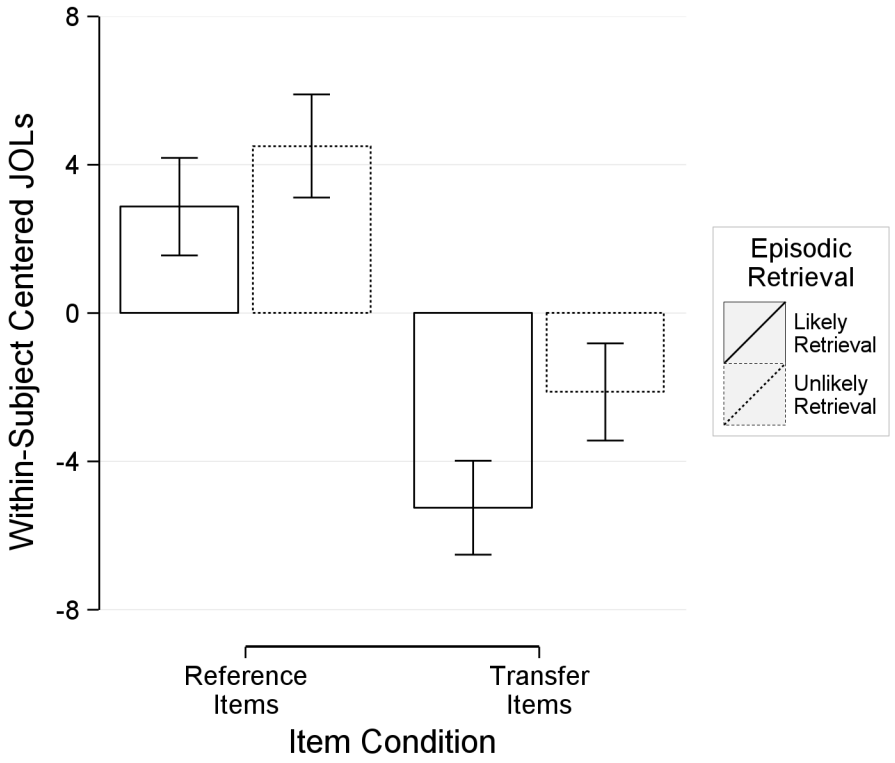

Figure 20. Experiment 2: Within-subject centered JOLs as a function of item condition (reference, transfer) and episodic retrieval (likely retrieval of the reference item, unlikely retrieval of the reference item). The values represent deflections from a subject's mean judgments of learning ratings and better represent within-subject differences (i.e., item type). The error bars represent standard errors of the within-subject means.

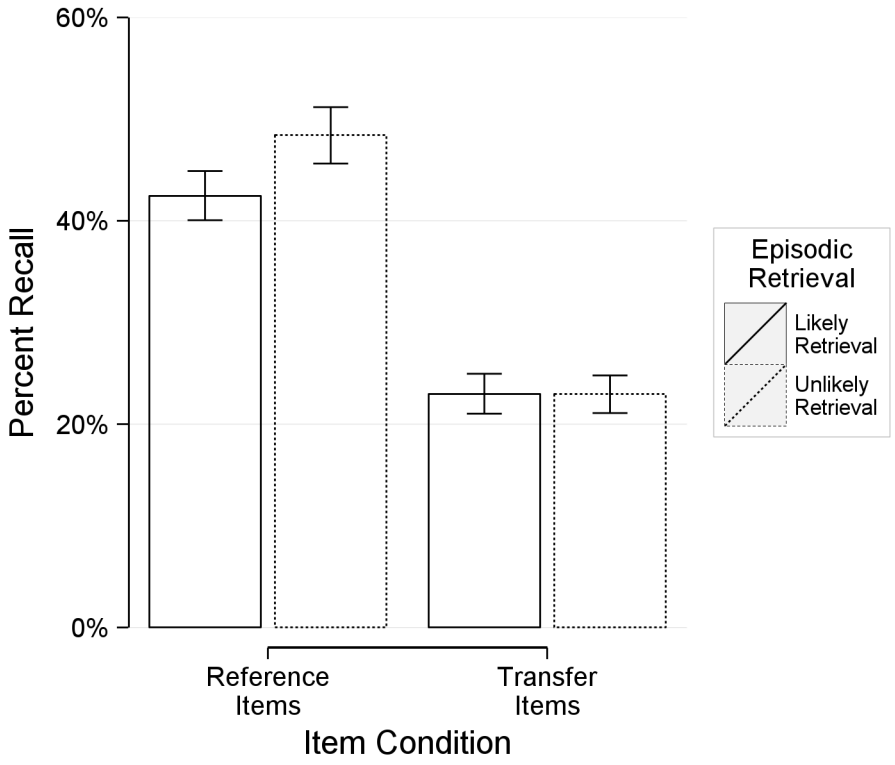

Figure 21. Experiment 2: Percent recall as a function of item condition (reference, transfer) and episodic retrieval (likely retrieval of the reference item, unlikely retrieval of the reference item). The error bars represent standard errors of the means.

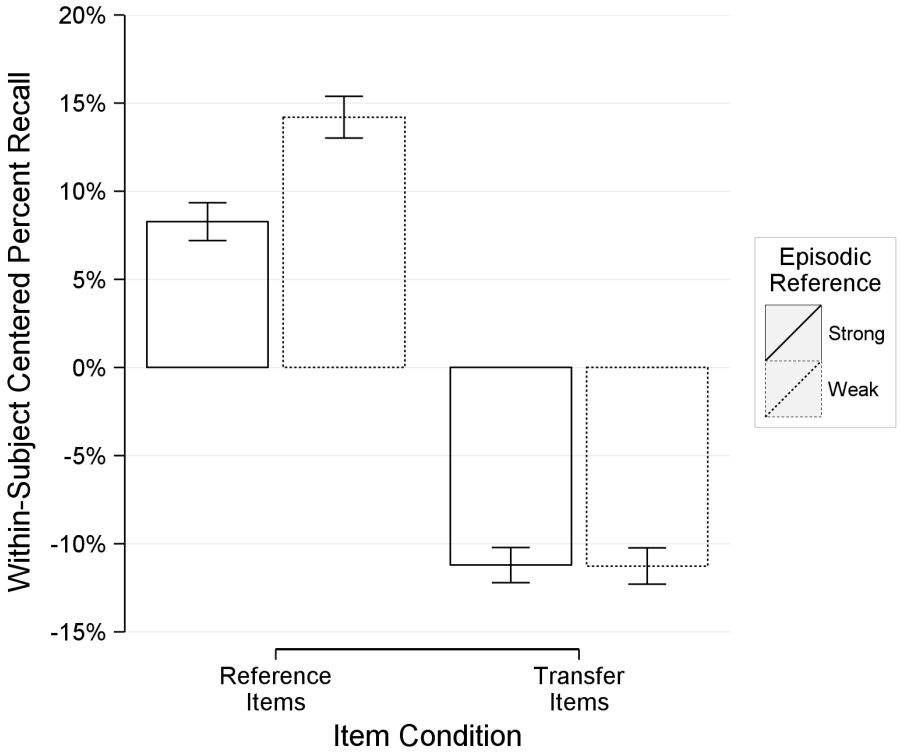

Figure 22. Experiment 2: Within-subject centered percent recall as a function of item condition (reference, transfer) and episodic retrieval (likely retrieval of the reference item, unlikely retrieval of the reference item). The values represent deflections from a subject's mean percent recall and better represent within-subject differences (i.e., item type). The error bars represent standard errors of the within-subject means.

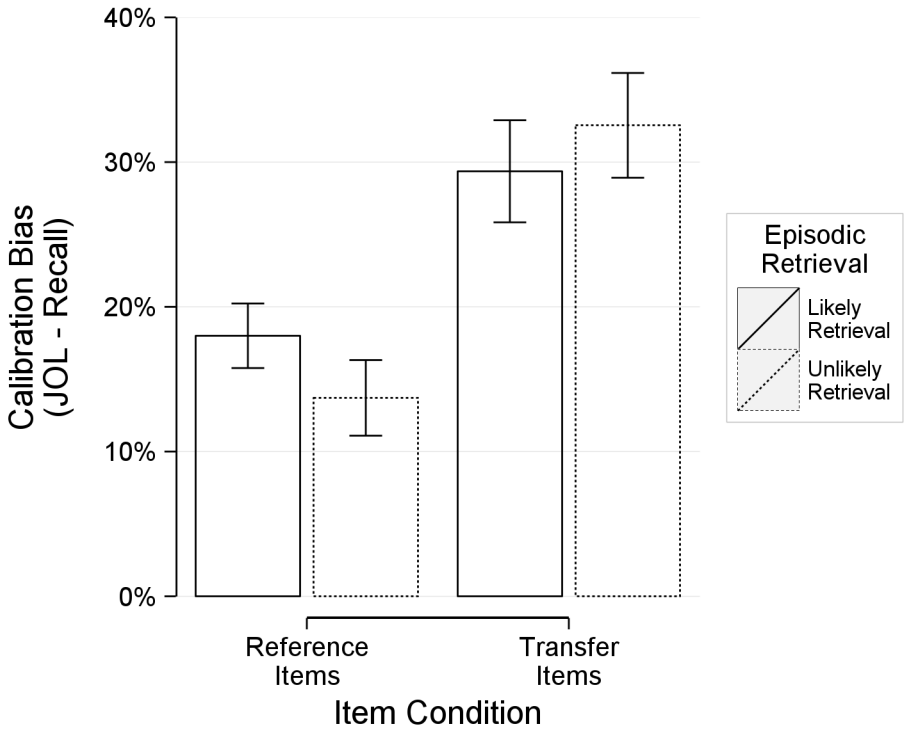

Figure 23. Experiment 2: Mean calibration bias as a function of item condition (reference, transfer) and episodic retrieval (likely retrieval of the reference item, unlikely retrieval of the reference item). Higher values indicate overconfidence. The error bars represent standard errors of the means.

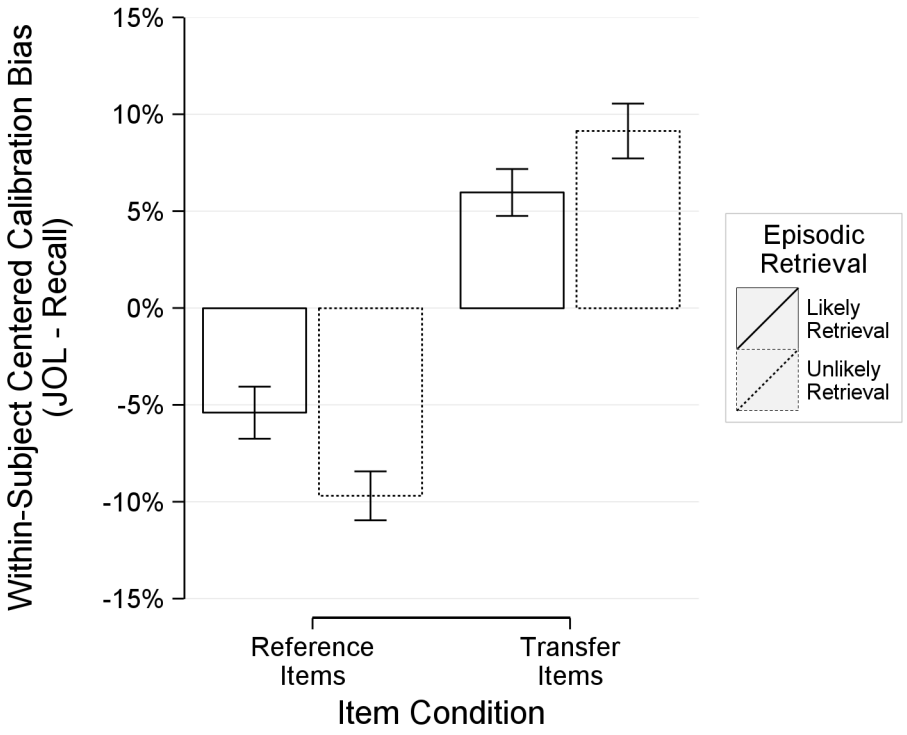

Figure 24. Experiment 2: Within-subject centered calibration bias as a function of item condition (reference, transfer) and episodic retrieval (likely retrieval of the reference item, unlikely retrieval of the reference item). The values represent deflections from a subject's mean calibration bias and better represent within-subject differences (i.e., item type). The error bars represent standard errors of the within-subject means.

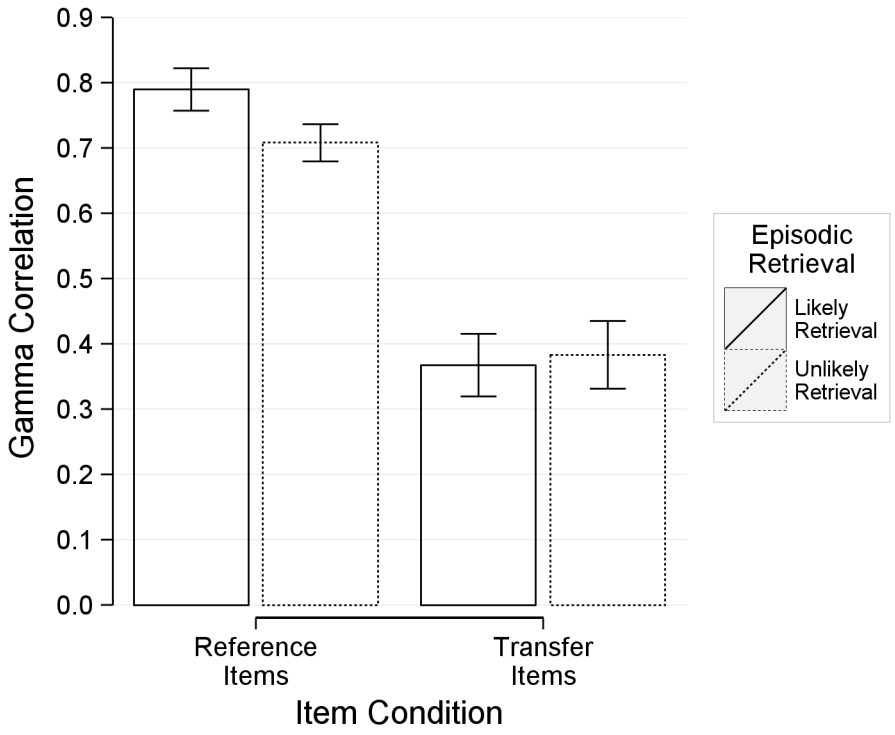

Figure 25. Experiment 2: Mean gamma correlation as a function of item condition (reference, transfer) and episodic retrieval (likely retrieval of the reference item, unlikely retrieval of the reference item). Higher values indicate better resolution. The error bars represent standard errors of the means.

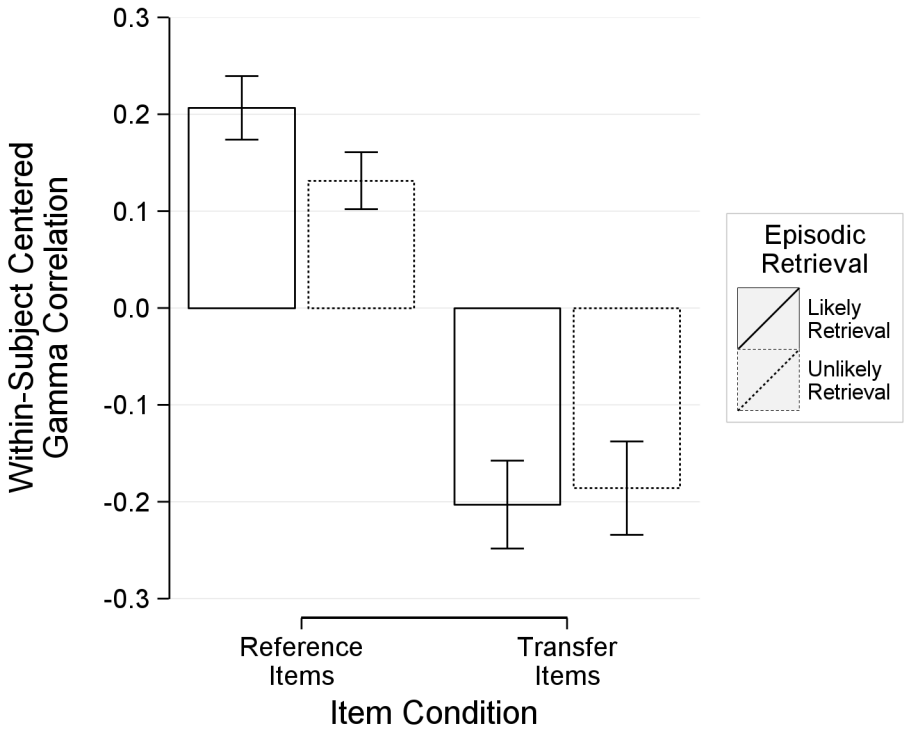

Figure 26. Experiment 2: Within-subject centered gamma correlation as a function of item condition (reference, transfer) and episodic retrieval (likely retrieval of the reference item, unlikely retrieval of the reference item). The values represent deflections from a subject's mean gamma correlation and better represent within-subject differences (i.e., item type). The error bars represent standard errors of the within-subject means.

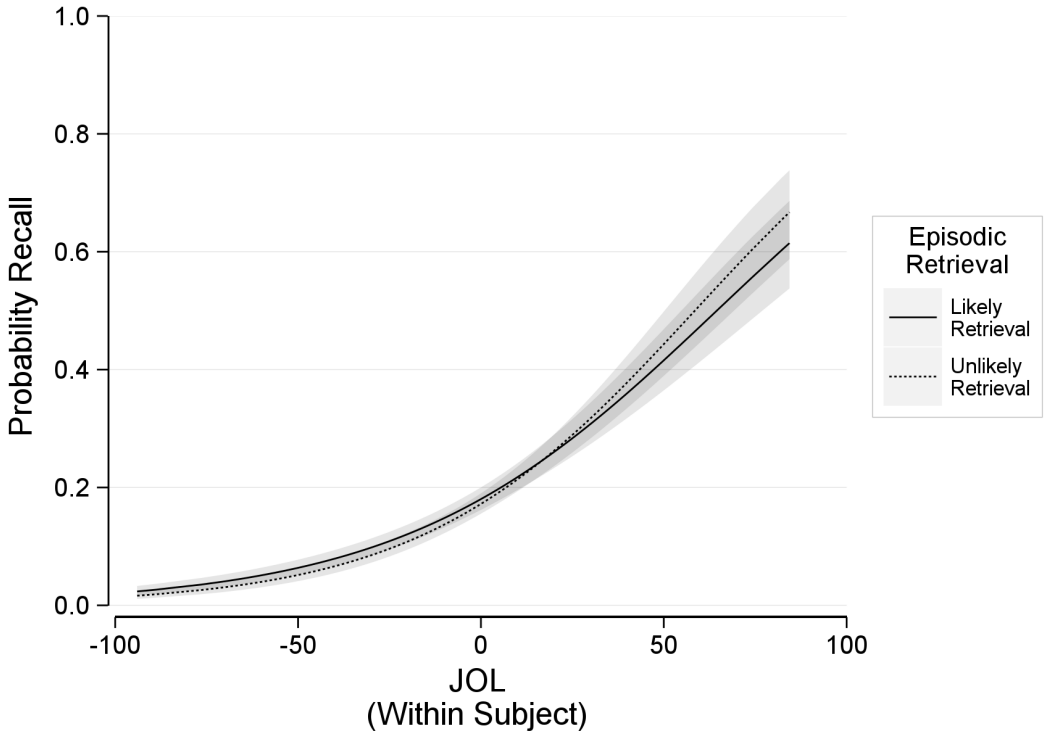

Figure 27. Experiment 2: The probability of recalling an item as a function of within-subject judgments of learning and episodic retrieval (likely retrieval of the reference item, unlikely retrieval of the reference item) at the sample mean judgments of learning. The error ribbon represents the combined standard errors of the fixed effects.

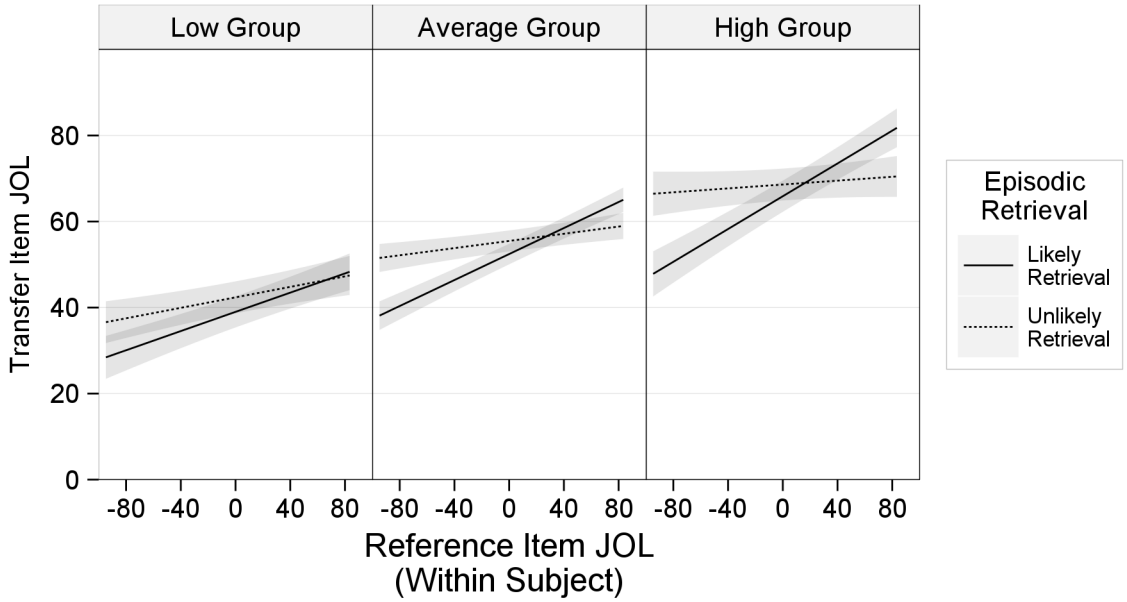

Figure 28. Experiment 2: Transfer item judgments of learning as a function of within-subject reference item judgments of learning and episodic retrieval (likely retrieval of the reference item, unlikely retrieval of the reference item). The three panels represent display the relationship at different levels of between-subject reference item judgments of learning. The low group is one standard deviation below the mean, the average group is at the mean, and the high group is one standard deviation above the mean. The error ribbon represents the combined standard errors of the fixed effects.

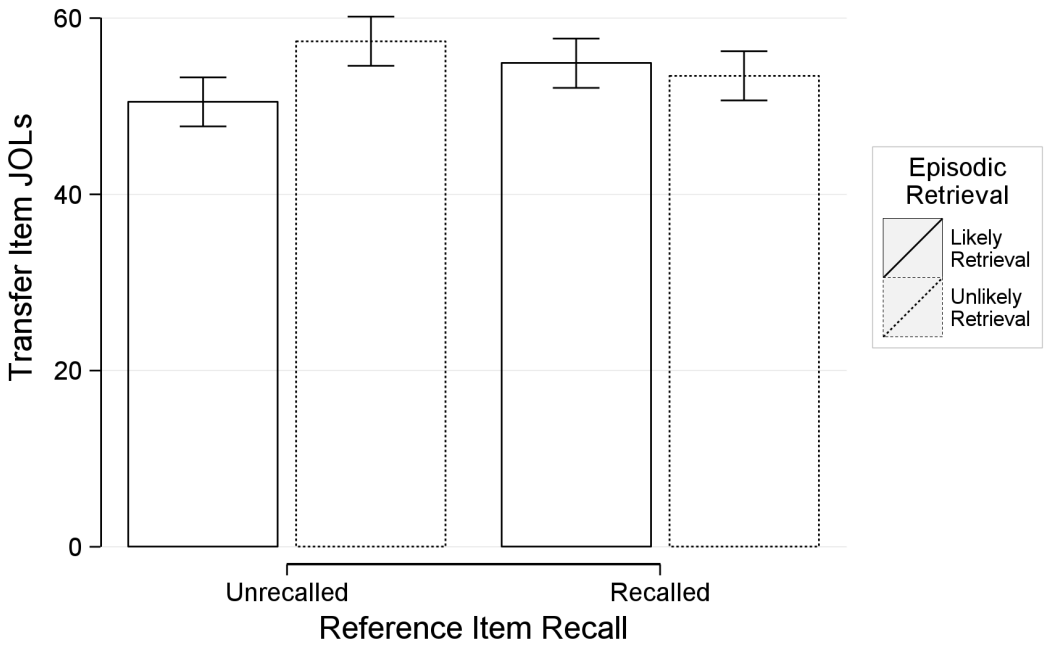

Figure 29. Experiment 2: Transfer item judgments of learning as a function of reference item final recall and episodic retrieval (likely retrieval of the reference item, unlikely retrieval of the reference item). The error bars represent the combined standard errors of the fixed effects.

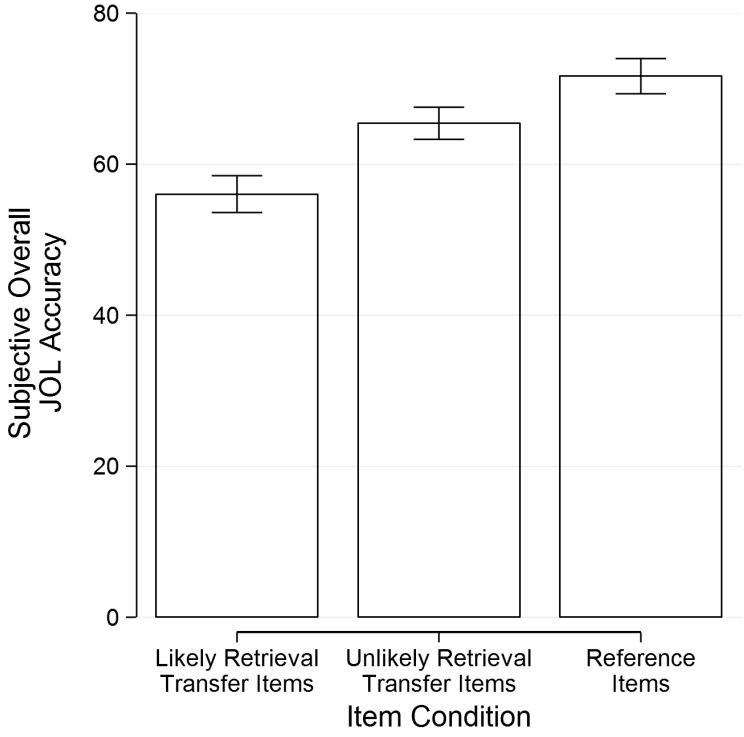

Figure 30. Experiment 2: The mean subjective accuracy of participants' judgments of learning as a function of item type. The error bars represent the standard error of the means.

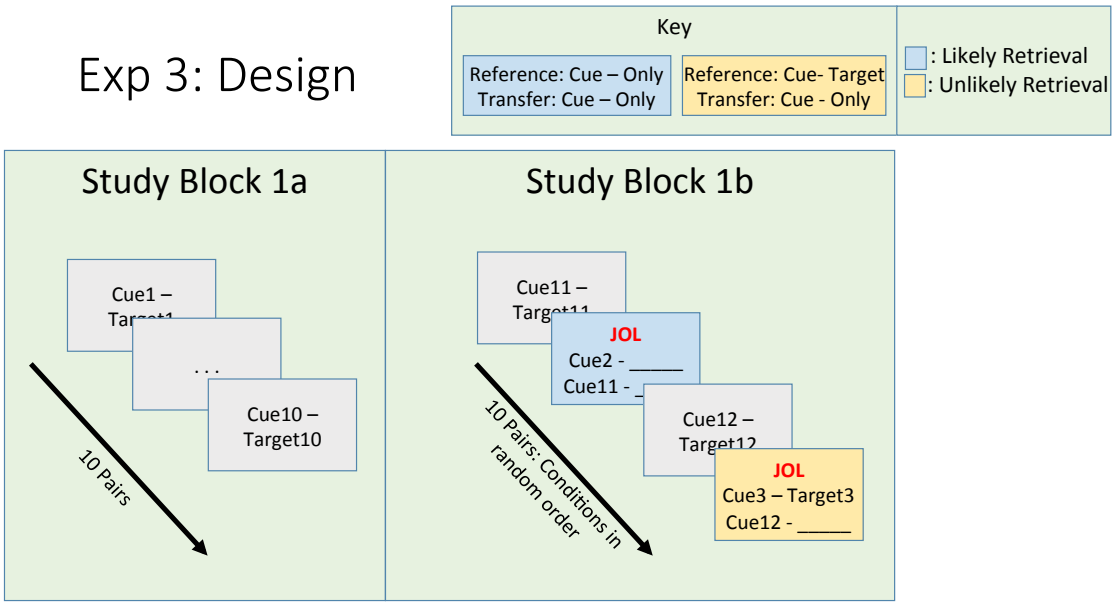

Figure 31. The design of Experiment 3. The first study block contains the reference items. They were re-presented in the second block after the study presentation of the transfer item. At that point, both the reference item and the transfer item received judgments of learning. The reference item always appeared above the transfer item and in either cue-only (blue box; likely to induce covert retrieval) or cue-target (yellow box; unlikely to induce covert retrieval). The study bocks were repeated twice within each study list.

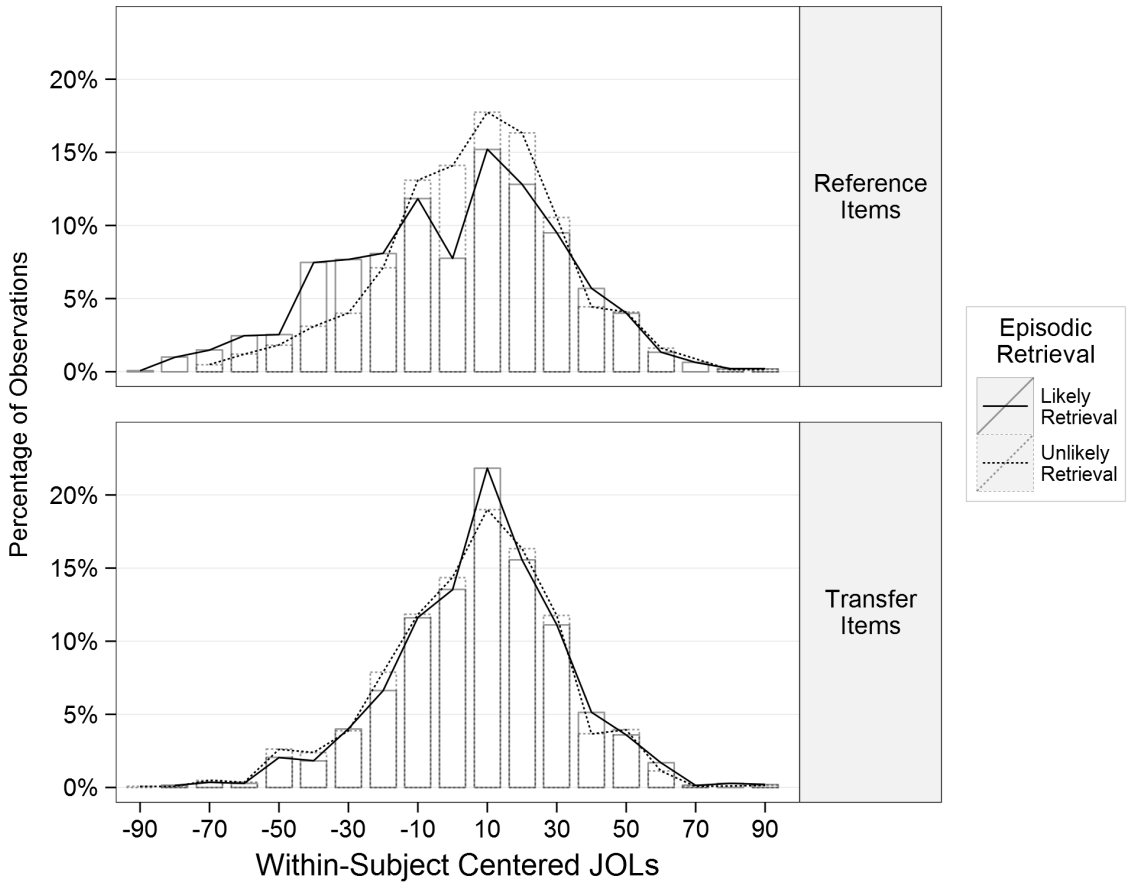

Figure 32. Experiment 3: A histogram displaying the percentage of observations within an item type (top: reference item; bottom: transfer item) given a particular within-subject centered judgment of learning. The superimposed lines indicate the between-subject episodic retrieval conditions (likely retrieval of the reference item, unlikely retrieval of the reference item).

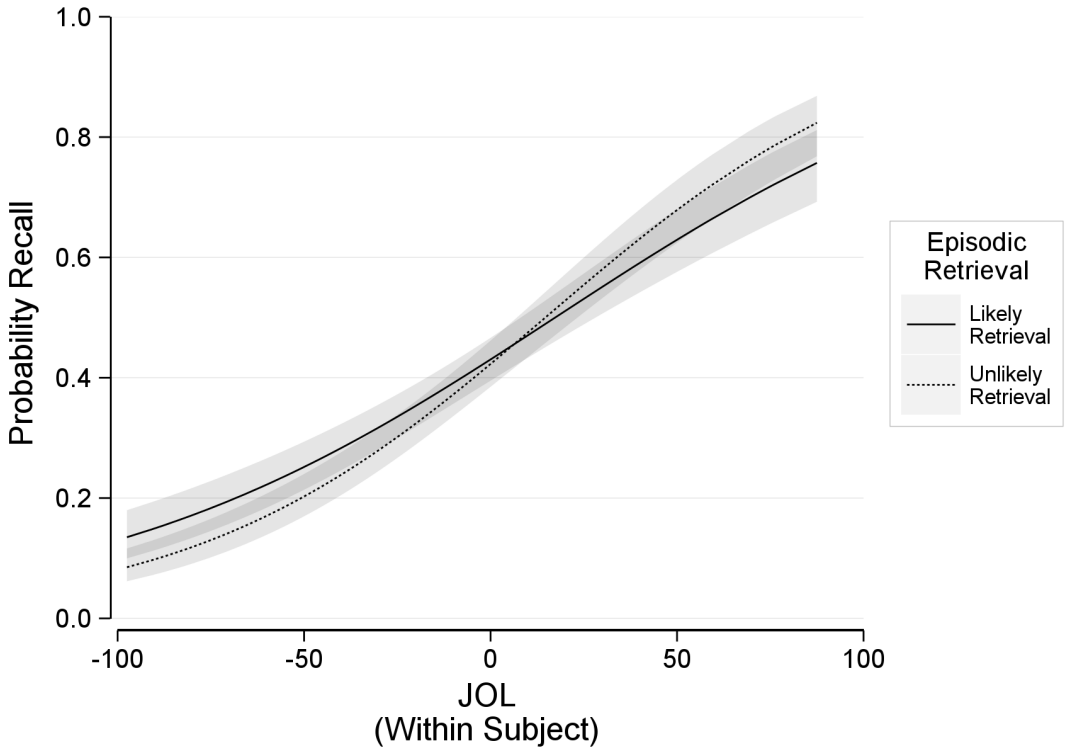

Figure 33. Experiment 3: The probability of recalling an item as a function of within-subject judgments of learning and episodic retrieval (likely retrieval of the reference item, unlikely retrieval of the reference item) at the sample mean judgments of learning. The error ribbon represents the combined standard errors of the fixed effects.

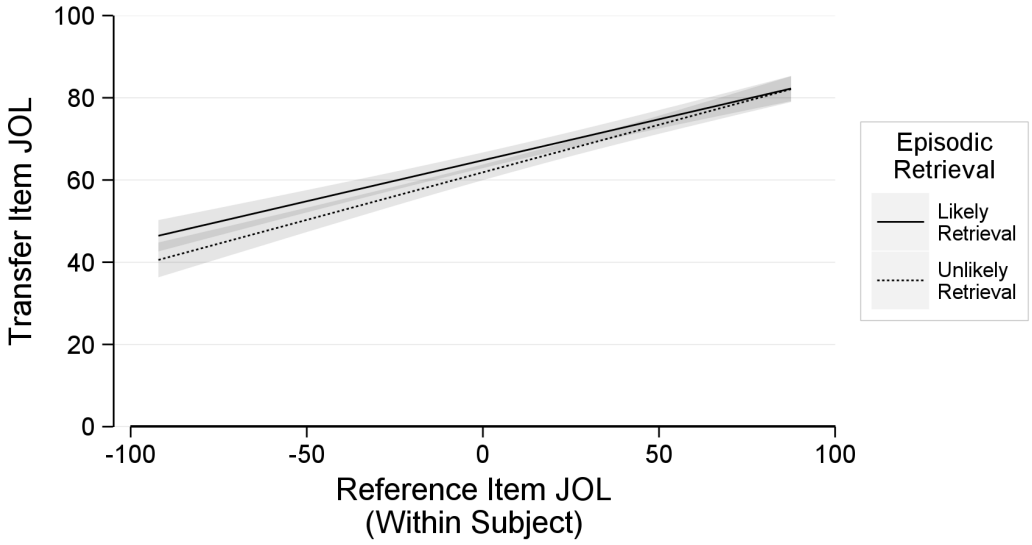

Figure 34. Experiment 3: Transfer item judgments of learning as a function of within-subject reference item judgments of learning and episodic retrieval (likely retrieval of the reference item, unlikely retrieval of the reference item). The error ribbon represents the combined standard errors of the fixed effects.

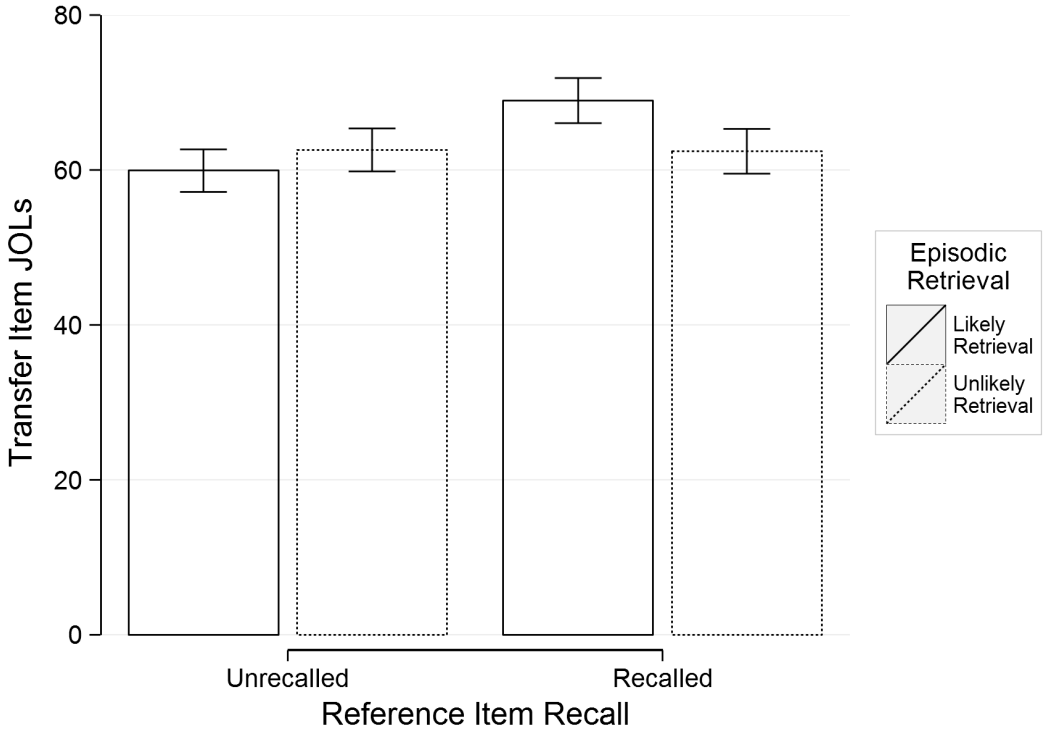

Figure 35. Experiment 3: Transfer item judgments of learning as a function of reference item final recall and episodic retrieval (likely retrieval of the reference item, unlikely retrieval of the reference item). The error bars represent the combined standard errors of the fixed effects.