Volume 30 · Number 5 · May 2022
Memory
Editors Martin A. Conway and Mark L. Howe
Routledge

Memory

Routledge
Taylor & Francis Group

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/pmem20

Confidence ratings are better predictors of future performance than delayed judgments of learning

Adam L. Putnam, Will Deng & K. Andrew DeSoto

To cite this article: Adam L. Putnam, Will Deng & K. Andrew DeSoto (2022) Confidence ratings are better predictors of future performance than delayed judgments of learning, Memory, 30:5, 537-553, DOI: 10.1080/09658211.2022.2026973

To link to this article: https://doi.org/10.1080/09658211.2022.2026973

MEMORY
2022, VOL. 30, NO. 5, 537–553
https://doi.org/10.1080/09658211.2022.2026973

Confidence ratings are better predictors of future performance than delayed judgments of learning

Adam L. Putnam [a], Will Deng[a] and K. Andrew DeSoto [b]

[a]Department of Psychology, Furman University, Greenville, SC, USA; [b]Association for Psychological Science, Washington, DC, USA

ABSTRACT
What is the best way to predict future memory performance? The intuitive answer is through judgments of learning (JOLs), in which people estimate how likely they are to remember something in the future. Recent theory, however, suggests that a retrospective confidence rating made just after a retrieval attempt might be a better predictor in some situations. In three preregistered experiments, we compared delayed JOLs to confidence ratings. People studied paired associates (E1) or psychology vocabulary terms (E2 & E3), then took a practice cued-recall test in which they made either a JOL or confidence rating after each response. They then took a final test. In Experiment 1, confidence ratings offered higher resolution (metacognitive accuracy) of memory for paired associates than did JOLs, but in Experiments 2 and 3, the advantage of confidence ratings was much smaller. A mini meta-analysis indicated that confidence ratings have a small advantage in predicting future performance over delayed JOLs. We argue that the two judgments rely on similar cues, and that even though JOLs explicitly ask people to predict future performance, doing so does not enhance prediction accuracy. Rather, the presence of a retention interval in the JOL cue adds variability to the judgment process.

ARTICLE HISTORY
Received 30 July 2021
Accepted 31 December 2021

KEYWORDS
Metacognition; judgments of learning; retrospective confidence judgments; paired associates; self-consistency

Imagine it is 10:00 PM and a student is studying for tomorrow's psychology exam. They have organised their notes into a study guide and are taking a practice test. While studying via this practice test, the student will ask themself questions like, "do I know this?", "will I remember this tomorrow?" or "do I need to spend more time on this fact?" The answers to these metacognitive judgment questions will affect what the student chooses to study and may determine how well they do on the exam. The goal of the current paper is to explore whether explicitly thinking about performance on a future test enhances students' predictions of performance on that future test. In particular, we examined whether judgments of learning, which explicitly ask learners about future performance, enhance predictions compared to confidence ratings.

Judgments of learning

Memory researchers use *judgments of learning* (abbreviated JOLs; for reviews, see Dunlosky & Metcalfe, 2009; Rhodes, 2016) to measure people's predictions about future memory performance. In a typical experiment, subjects see a to-be-remembered item (e.g., the word pair *sailor - anchor*) and make a rating, usually on a numeric scale, to indicate their prospective confidence that they will remember the item on a future test (as queried by a question such as "How confident are you that you will remember the target in 10 minutes?") Research has suggested that when people make metamemory judgments, they are not directly accessing memory contents, but are instead *inferring* the memory strength of the item based on available cues such as the difficulty of the item, how long they studied the item, and internal experiences related to processing the item, such as remembering a prior recall attempt or the fluency of retrieving related information (see cue-utilization theory, Koriat, 1997).

One of the consequences of this inferential process is that people may use cues that do not actually improve prediction or ignore cues that do. On one hand, Rhodes and Castel (2008) showed that subjects believed that words presented in a large font would be easier to remember than words presented in a small font even though they were equally memorable (for a review, see Luna et al., 2018). On the other hand, Koriat and colleagues (2004) showed that subjects provided similar JOLs to items regardless of whether they thought the final test would be in 10 minutes or 1 week, demonstrating that subjects can also neglect cues that *are* related to memorability,

CONTACT Adam L. Putnam adam.putnam@furman.edu Department of Psychology, Furman University, 3300 Poinsett Highway, Greenville, SC 29613, USA

Supplemental data for this article can be accessed https://doi.org/10.1080/09658211.2022.2026973

such as retention interval. Results like these suggest that JOLs can be inaccurate.

However, JOLs can be made more accurate. One way to improve the accuracy of JOLs is to insert a time delay between when a subject first learns an item and when they make the JOL (Nelson & Dunlosky, 1991). These *delayed* JOLs are more accurate because 1) information about the item is not available in working memory (so the JOL better assesses long-term memory) and 2) subjects may covertly retrieve the item, which enhances future recall and brings JOLs—which are often overconfident—in line with eventual recall performance (Kimball & Metcalfe, 2003; Spellman & Bjork, 1992). Thus, changing structural aspects of how JOLs are made can enhance their predictive accuracy.

Finally, a growing body of research has examined how the framing of JOLs might affect confidence, showing that making judgments-of-forgetting or judgments-of-retention can lead to more accurate predictions than following standard JOL instructions (Finn, 2008; Tauber & Rhodes, 2012). In contrast, other work suggests that while judgments-of-forgetting do lead to differences in confidence ratings, these differences more likely arise from artifacts of how the judgment scale is used rather than a change to the subjective underlying confidence of the subjects (England et al., 2017; Serra & England, 2012). The key point from this research is that subtle changes to the framing or instructions of a JOL prompt may influence JOL accuracy, but as Rhodes (2016) suggests, more research is needed to understand exactly how such framing affects JOLs.

Despite variability in the implementation of JOLs (e.g., different instructions and different timing delays), their core principle is that subjects are explicitly asked to predict future memory performance. Intuition suggests that the best way to predict future memory performance would be to try and predict future memory performance, which is what a JOL asks subjects to do. However, a different metacognitive rating may actually provide a better prediction of future performance.

Confidence ratings

Another type of metacognitive judgment is the *retrospective confidence rating* (often referred to as "confidence rating"). Instead of asking subjects to project into the future, confidence ratings ask subjects to assess how well they did in a prior memory retrieval attempt (e.g., "How confident are you that your answer is correct?"). Confidence ratings are generally fairly accurate (at least, compared to standard JOLs; Koriat & Goldsmith, 1996). But, as with JOLs, confidence ratings are likely informed by a variety of cues. One cue in particular—retrieval fluency—appears to be a strong predictor of confidence ratings, with fast response times correlating with high confidence ratings (Benjamin & Bjork, 1996). This relationship makes sense, given that one's ability to retrieve something in the past is a good predictor of future retrieval (Estes et al., 1960; Tulving, 1964). Curiously, JOL accuracy may also be improved by past retrieval attempts (when such attempts are available; Lovelace, 1984; Metcalfe & Finn, 2008), suggesting that regardless of whether learners are making future predictions or evaluating past performance, retrieval fluency is an important diagnostic cue.

Comparing the predictive accuracy of JOLs and confidence ratings

Despite JOLs and confidence ratings asking different questions— "how *will* you do" versus "how *did* you do"—both metacognitive judgments may engage similar memorial processes and are subject to influences (some similar, some different) that may help or harm judgment accuracy (see Dunlosky & Tauber, 2014, to read about a framework that suggests most metacognitive judgments rely on common processes). For example, retrieval fluency may be an important cue for both delayed JOLs and confidence ratings, although this retrieval may be covert for JOLs and overt for confidence ratings. Given these similarities, we wondered whether confidence ratings might be better able to predict future memory performance.

Studies directly comparing the predictive accuracy of these two judgments are relatively rare, likely because the two are traditionally made at different times and for different purposes (Dougherty et al., 2005 provides a direct comparison; Busey et al., 2000 and Pierce & Smith, 2001 provide related findings). However, if delayed JOLs and confidence ratings are made immediately after a retrieval attempt, the only difference between them is that the delayed JOLs explicitly ask subjects to project into the future (yet are implicitly influenced by the past), whereas confidence ratings explicitly ask subjects to consider the past (and are indeed influenced by the past). In practice, the difference is simply a matter of instructions —whether they invite subjects to look towards the future or consider the prior retrieval attempt that just occurred. How might looking towards the future affect judgment accuracy? Does considering the future enhance metacognitive accuracy? As mentioned earlier, from an intuitive standpoint, JOLs, which explicitly look at the future, should be more accurate at predicting future memory performance because they prompt subjects to consider forgetting, something that children as young as 4 recognise (Lyon & Flavell, 1993). However, three lines of research suggest that there may be either no difference in metacognitive accuracy between these judgment types or a slight advantage for confidence ratings.

First, we turn to Koriat's (2012, 2015) *self-consistency model of subjective confidence*, which suggests that when people make a memory judgment, they evaluate prior knowledge and other cues in memory. High confidence ratings occur when many cues point to the same answer.

An aspect of this theory is that, " … subjective confidence is prospective in nature: It reflects an assessment of the reproducibility of the choice—the likelihood that the same answer will be chosen when the same question is presented again" (Koriat, 2012, p. 86). For example, if an American is asked, "What is the capital of the United States?" they are likely to answer, "Washington, D.C." and would likely provide that answer with high confidence each time the question is asked. In contrast, when asked, "What is the capital of South Carolina?" an American (not from the Southeast U.S.) might answer "Columbia" on one occasion, but "Charleston" on another occasion, as a result of different memory cues pointing toward different answers. This lack of cue consistency likewise contributes to low confidence ratings for either answer.

If subjects have good memory resolution—that is, that they assign higher confidence ratings when they are more likely to be correct—confidence ratings should also serve as an accurate indicator of future memory performance: a high confidence rating means that the individual is likely to provide the same response in the future. The key implication of the self-consistency model for the current study is that a past retrieval attempt (which triggers experiences of retrieval fluency and related knowledge) is one of the strongest and most diagnostic cues for predicting the accuracy of a retrieval attempt. Because past retrieval often predicts future retrieval, and because confidence ratings are strongly related to retrieval fluency, confidence ratings may be just as accurate (or more accurate) at predicting *future* retrieval (Benjamin & Bjork, 1996; Estes et al., 1960; Tulving, 1964).

Second, even though most people intuitively understand that forgetting occurs over time, it is less clear whether they consider this belief-based cue when making a JOL. As noted above, one study showed that subjects provided similar JOLs regardless of whether the anticipated retention interval was 10-minutes or 1-week, suggesting that subjects either ignored the anticipated retention interval or did not weight it heavily (Koriat et al., 2004, Experiment 1). A key factor in that result, however, is that the anticipated retention interval was manipulated between-subjects; when retention interval was manipulated within-subjects, people start to account for forgetting in their JOLs (Koriat et al., 2004, Experiment 3; see also Rawson et al., 2002). These studies suggest that people may not consider the retention interval at all in making a JOL if the retention interval is not made salient, which would lead to the prediction that the accuracy of JOLs would be similar to that of confidence ratings.

Third, some empirical evidence has suggested that confidence ratings may be more accurate than delayed JOLs. In a study by Dougherty and colleagues (2005), subjects studied paired associates, then 30 s later attempted to recall each pair, after which, they made a confidence rating, a JOL, or both. Ten minutes later, subjects took a final test on the pairs. Critically, confidence ratings better predicted final test performance than did JOLs. Dougherty and colleagues suggested that both JOLs and confidence ratings are heavily influenced by retrievability, but that additional cues influence JOLs (supported by an analysis showing that subjects took longer to make JOLs than confidence ratings). This pattern led the authors to conceptualise the relationship between JOLs and confidence ratings as "JOL = confidence + variation." According to Dougherty and colleagues, this variation could be random noise or more systematic variation, but in either case was likely due to subjects trying to predict future recall, which introduced additional inferential errors compared to making a confidence judgment.

Together, these lines of research imply that, in terms of predicting future performance, confidence ratings will either be equal to or more accurate than delayed JOLs, a somewhat surprising prediction given that JOLs explicitly ask people to think about the future whereas confidence ratings do not.

The current studies

The goal of the studies reported here was to test our prediction that confidence ratings would be equal to or more accurate predictors of future performance than JOLs. To do so, we started with delayed JOLs, which are already more accurate than immediate JOLs (and thus provide a strong test of our hypothesis), and compared them to confidence ratings using both laboratory and educationally-relevant materials. If it is the case that confidence ratings are more accurate than JOLs, some of the cues involved in making a JOL (projecting to the future in general or attending to a specific anticipated retention interval) may *not* be diagnostically valid cues (i.e., ones that help people to predict future performance).

Our experiments were similar to Dougherty et al. (2005), but with two key differences. First, Experiment 1 of Dougherty and colleagues had subjects make both confidence ratings and JOLs, which may exaggerate differences between the two types of judgments; we used a fully between-subjects manipulation (see McDaniel & Bugg, 2008 for a discussion of how design elements can influence memory phenomena). Second, Dougherty and colleagues' method used a continuous procedure where subjects' initial studying, first retrieval attempt, and metacognitive judgment were temporally intertwined, meaning that the initial retrieval and metacognitive judgments may have affected the encoding of subsequent items. This is problematic because taking a test may affect later encoding (i.e., test-potentiated learning, Arnold & McDermott, 2013) and because subjects may adjust their metacognitive judgments as a result of retrieval (Tauber & Rhodes, 2010). In the current study, we had subjects complete study of all items before attempting retrieval or making a metacognitive judgment, thus allowing us to better examine the accuracy of the metacognitive predictions.

We designed three preregistered experiments to test the hypothesis that retrospective confidence ratings would lead to better metacognitive accuracy than JOLs. In each study, subjects learned either cue-target word pairs (Experiment 1) or psychology terms and definitions (Experiments 2a, 2b, and 3). After a brief distractor task, subjects attempted retrieval for each item and then provided a confidence rating or JOL. After the initial test, subjects took a cued final recall test (immediately or after a delay in E3). We also report a mini meta-analysis that summarises the results of all three studies.

Experiment 1

Experiment 1 evaluated whether JOLs or confidence ratings would better predict future test performance. Subjects studied weakly related word pairs, then completed a first test where one group made JOLs and the other group made confidence ratings. Then, subjects completed the final test after five minutes. We predicted that subjects in the confidence group would show higher metacognitive accuracy (resolution) in predicting final test performance than subjects in the JOL group.

Method

We preregistered Experiment 1 on the Open Science Framework (OSF.IO/Z2H3B). The preregistration described our planned sample size, stopping rules, subject exclusion rules, main hypotheses, analysis plan, and other details. All measures that we collected in the experiment were included in the preregistration and are reported here or in the supplemental materials. The Carleton College IRB approved Experiments 1–2 and the Furman University IRB approved Experiment 3.

Subjects

We recruited subjects from Amazon's Mechanical Turk website (MTurk). We determined the sample size based on a power analysis indicating that 190 subjects would be sufficient to detect a medium-sized effect ($d = 0.50$) with 90% power. We unexpectedly ended up with data from 199 subjects—some subjects completed the study without submitting their payment code to MTurk, allowing more subjects to participate.[1] Of the 199 subjects who participated, we excluded seven from our analyses according to our preregistered exclusion rules: one subject was not a fluent English speaker, one subject used notes during the experiment, and five subjects scored less than 5% correct on the final test.

In the final sample of 192 subjects (mean age = 37.5, *SD* = 11.40, 97 female and 95 male), 95 subjects were randomly assigned to the confidence condition and 97 to the JOL condition.

Materials

Subjects studied 40 weakly related word pairs, for example, *sailor-anchor* and *hunger-pizza* (taken from Putnam et al., 2014). Each word was a noun and between 4 and 8 letters long. The pairs had both a forward (cue-to-target) and backward (target-to-cue) strength between .01 and .02.

Procedure

The experiment was administered via Qualtrics and consisted of a study phase, an initial test (where metacognitive judgments were made), and a final test. Upon starting the experiment, subjects were randomly assigned to the confidence or JOL condition.

Study phase. Subjects read instructions saying they would be reading word pairs to remember for an upcoming test where they would be cued with the first word of the pair and need to recall the second word. Subjects were asked to not take notes during the experiment. During the study phase, word pairs appeared on screen one at a time in a black font on a white background for 4 s. A blank white screen appeared between each word pair for 500 ms. The word pairs appeared in a random order for each subject. After studying all 40 pairs, subjects participated in a distractor task—spotting the difference between two pictures—for 1 min.

Initial test. Subjects then took an initial test on the studied word pairs. Although there was no time limit, subjects were told to move as quickly and accurately as possible. Subjects saw a cue word followed by question marks (e.g., *hunger - ???*) and typed the associated word from the study phase into a text box and submitted their response. After responding, a new screen appeared where subjects made their metacognitive judgment. Subjects in the *JOL* group saw, "How confident are you that in about 10 minutes you will be able to recall the second word of the pair when prompted with the first?" whereas subjects in the confidence group saw, "How confident are you that you recalled this pair correctly?" In other words, subjects in the JOL condition predicted future performance, while subjects in the confidence condition rated their confidence on the response they just submitted. In both cases, the cue word and the target were not displayed on the screen when subjects made the metacognitive judgment.

Subjects made their metacognitive judgments by using a slider that appeared below the question prompt. The slider was anchored with *not at all confident* on the left and *entirely confident* on the right; the numbers 0, 20, 40, 60, 80, and 100 appeared above the slider (although subjects could select any position on the slider), and the slider started on 0 for every trial. After submitting their metacognitive rating, subjects moved on to the next item. No feedback was provided. After finishing the practice test, subjects worked on a distractor task—listing as many US states as they could—for 5 minutes.

Final test. On the final test, subjects again saw the first word of a pair followed by question marks, and they recalled the target word by typing it into a text box. The cues appeared in a random order, and there was no time limit for responses. After finishing the final test, subjects answered demographic questions (reporting age, gender, and fluency in English) and whether they had written anything down during the experiment.

Results and discussion

Analysis plan

Data and analysis scripts for all experiments are posted on the OSF. Before conducting analyses, we corrected spelling mistakes in responses (e.g., *emrlad* became *emerald*; results throughout were the same without correcting spelling mistakes). In all of the analyses, we used a criterion of $p < .05$ for determining significance. For most comparisons, we used a Welch two-sample *t*-test, which automatically corrects for violations of homogeneity. Before conducting our comparisons, we evaluated the normality of our data by visual inspection of histograms, QQ plots, and by running a Shapiro-Wilk normality test. In cases of violations of normality, we used the non-parametric Wilcoxon Rank Sum test (which is functionally equivalent to a Mann-Whitney U test; Field et al., 2012) instead of a *t*-test. Unless noted, the nonparametric results matched the parametric results.

All of the reported analyses were preregistered, except for our comparison between groups of the metacognitive ratings (i.e., their magnitude) and the calibration curves, both of which provide valuable context. Other exploratory analyses (such as the relationship between first test recall and the metacognitive ratings) are reported in our supplemental materials.

Our main question was whether retrospective confidence ratings or JOLs were more accurate predictors of final test performance. To assess metacognitive accuracy, we examined both calibration and resolution. For calibration, we plotted calibration curves with 95% confidence intervals. For resolution, we calculated Goodman-Kruskal gamma correlations (Nelson, 1984). Gamma correlations, like Pearson correlations, range from −1.00 to 1.00, with a value of 1 indicating a perfect relationship between judgment magnitude and outcome, and are often used to measure resolution. Because some limitations of gamma have been noted (Diaz & Benjamin, 2011; Murayama et al., 2014), we also report exploratory analyses using Kendall's Tau (as suggested by Dougherty et al., 2018) and an alternative calculation for Gamma based on ROC curves, G_{trap} (Higham & Higham, 2019) in our supplemental materials.[2]

First test recall

On the first test, subjects in the confidence group had similar recall ($M = .53$, $Mdn. = .50$, $SD = .23$) as the JOL group ($M = .49$, $Mdn. = .48$, $SD = .25$), $W = 4996.5$, $p = .313$, $d = 0.14$. Given that, at this point in the experiment, the two groups were identical, this outcome was expected.

Final test recall

Subjects in the confidence group recalled the same proportion of items on the final test ($M = .52$, $Mdn. = .48$, $SD = .26$) as subjects in the JOL group ($M = .48$, $Mdn. = .48$, $SD = .26$), $W = 5036$, $p = .267$, $d = 0.15$. Thus, recall was similar regardless of the type of metacognitive judgment that subjects made. Having similar memory performance on the tests is beneficial for examining the differences in metacognitive prediction accuracy—if one group had higher test performance than the other, it would be difficult to know whether differences in metacognitive accuracy are due to metacognitive monitoring or memorial processes (Dunlosky & Metcalfe, 2009, pp. 54–55).

Metacognitive judgment distribution

As seen in Figure 1, the confidence ratings made by subjects in the confidence condition were higher than the JOL ratings made by subjects in the JOL condition, $t(189.72) = 2.05$, $p = .041$, $d = 0.30$. Because the two groups had similar recall performance on the first and final test, this difference in overall metacognitive ratings may explain the differences in resolution that we describe below.

Calibration

Figure 2 displays calibration curves for the confidence and the JOL groups. These curves plot the relationship between metacognitive ratings and final test performance. Confidence ratings were binned into five categories and plotted against average proportion recalled for those bins. The straight diagonal line represents perfect calibration, or when subjects' metacognitive ratings would hypothetically exactly match their performance on the final test. Both groups were similarly calibrated, showing underconfidence at the lowest levels of confidence but overconfidence at every other point on the scale, a pattern known as the *hard-easy effect* (Dunlosky & Metcalfe, 2009).

Resolution

The main focus of our study was examining how well subjects' metacognitive ratings predicted final test recall.[3] For the confidence group, gamma was .88 ($Mdn. = .92$, 95% CI = [.85, .90]) which was significantly higher than the JOL group, for which gamma was .72 ($Mdn. = .84$, 95% CI = [.65, .79], $W = 5982$, $p < .001$, $d = 0.61$).[4] Thus, as predicted, subjects were more accurate at predicting future test performance when they estimated their current test performance rather than projecting into the future.

Summary of experiment 1

Subjects in the JOL and the confidence groups were similarly calibrated, but the confidence group showed greater metacognitive accuracy as measured by resolution. Thus, our main hypothesis, that retrospective confidence

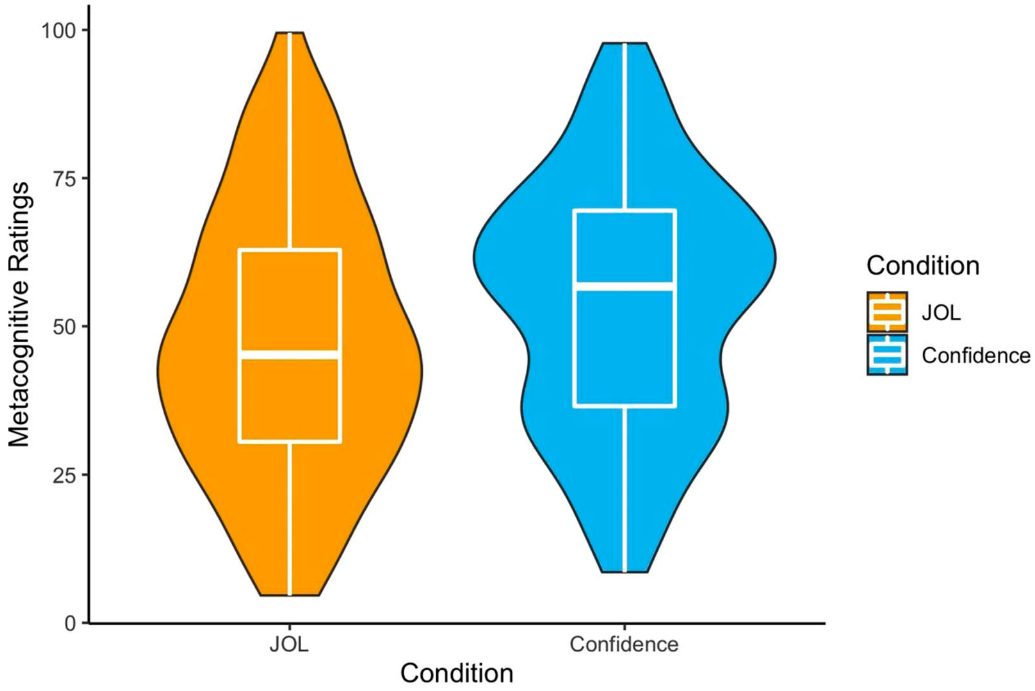

Figure 1. Violin Plots of Metacognitive Ratings in Experiment 1.

Note. The violin plot displays the probability density of responses at different values of metacognitive ratings. The central bars of the embedded box plot represent the medians, the upper and lower bounds of the box represent the lower and upper quartiles, and the whiskers represent 1.5 x the interquartile range.

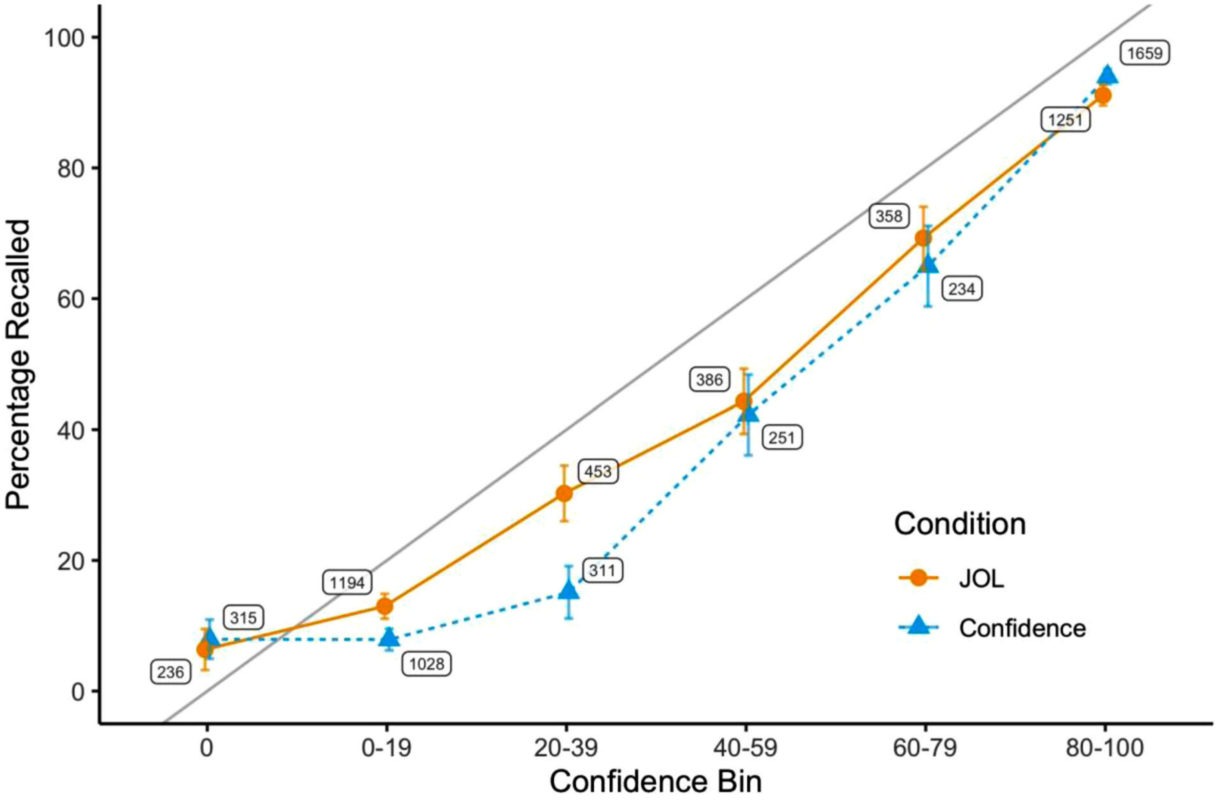

Figure 2. Calibration Curve for Experiment 1.

Note: The grey line represents perfect calibration, and the numbers in the labels refer to the number of observations each point represents. Error bars depict 95% confidence intervals.

ratings would be more accurate at predicting future performance than JOLs, was supported. This outcome is consistent with Koriat's self-consistency model (2012), in showing that confidence indexes reliability (or self-consistency). This outcome is also consistent with prior work suggesting that confidence ratings may be more accurate predictors of future recall than JOLs (Dougherty et al., 2005).

The results of Experiment 1 contribute to the understanding of Dougherty et al. (2005) in four ways. First, because final recall was the same for both conditions, we can hypothesise that the difference in resolution was due to a metacognitive process, rather than a memorial process at final test. Second, because the encoding phase was finished before subjects made any metacognitive judgments, we can also infer that the difference in resolution was due to the metacognitive judgment itself, rather than due to how making a metacognitive judgment might have affected subsequent encoding. Third, using a between-subjects manipulation instead of a within-subjects manipulation provided a stronger test of the central hypothesis. Finally, the current results support the JOLs = confidence + variation model; if considering the retention interval led to a consistent drop in metacognitive ratings, resolution would be unaffected. The lower resolution in the JOL group, however, suggests that considering the retention interval was adding noise to the judgment process.

Experiment 2

Rhodes (2016), among others, has called for researchers to use more complex materials when studying metamemory. The goal of Experiment 2 was to extend the finding of Experiment 1 to more educationally relevant materials. In Experiment 2, subjects studied psychology vocabulary terms (e.g., *prosopagnosia* and *reliability*) and their definitions. On the initial and final tests, subjects saw a definition and recalled the paired vocabulary word. As in Experiment 1, subjects provided JOLs or confidence ratings after each response on the initial test. Experiment 2 was conducted twice (referred to as 2A and 2B). The only difference was that in 2B, we also measured reaction times for the first test responses and the metacognitive judgments as a way to see whether the two judgments were relying on the same processes. We again predicted that confidence ratings would lead to higher resolution than JOLs.

Experiment 2A

Method

Experiment 2A was preregistered at OSF.IO/WHNA7.

Subjects. In Experiment 2A, we aimed to recruit 175 MTurk workers—after seeing the medium-sized effect of resolution in Experiment 1, a power analysis suggested 175 subjects would be adequate. We ended up with 179 subjects for the same reason as in Experiment 1. As noted in our preregistration, we excluded data from any subject who reported not being fluent in English (two subjects), any subject who reported writing down information before the final test (two subjects), and any subject who scored less than 5% on the final test (six subjects). Our final sample consisted of 169 people (mean age = 35.38, *SD* = 12.21, 89 female and 80 male), with 84 subjects randomly assigned to the confidence group and 85 subjects assigned to the JOL group.

Materials and procedure. The materials consisted of psychology vocabulary terms and definitions (e.g., *Prosopagnosia: The inability to recognize faces; this disorder is usually produced by lesions in the parietal lobes*). We created the materials by selecting 45 definitions from the glossary of several introductory psychology texts. We then ran a pilot test (see supplemental materials) and selected 20 medium-difficulty items.

The procedure was identical to Experiment 1, except for the new materials and a timing adjustment to accommodate those materials. Subjects studied each key term and its definition for 12 s during the study phase. In the recall tests, subjects were cued with the definition (*The inability to recognize faces; this disorder is usually produced by lesions in the parietal lobes*) and recalled the vocabulary term (*Prosopagnosia*). Otherwise, the tests were identical to E1.

Results and discussion

First test recall. On the first test, the confidence group recalled the same proportion of items (*M* = .48, *Mdn.* = .45, *SD* = .25) as did the JOL group (*M* = .52, *Mdn.* = .50, *SD* = .25), *W* = 3247, *p* = .309, *d* = 0.16.

Final test recall

The JOL group recalled a similar proportion of items on the final test (*M* = .51, *Mdn.* = .50, *SD* = .27) as did the confidence group, (*M* = .48, *Mdn.* = .43, *SD* = .26), *W* = 3363, *p* = .514, *d* = 0.12. Thus, as in Experiment 1, recall was similar regardless of the type of metacognitive judgment that subjects made.

Metacognitive judgment distribution

As seen in Figure 3 (top panel), the confidence ratings made by subjects in the confidence condition were similar to the JOLs made by subjects in the JOL condition, *W* = 3640, *p* = .827, *d* = 0.04. Thus, in contrast to Experiment 1, both groups not only had similar test performances, but also made similar metacognitive judgments.

Calibration

Figure 4 (top panel) displays calibration curves for the confidence and the JOL conditions predicting performance on the final test. Both groups showed the hard-easy effect.

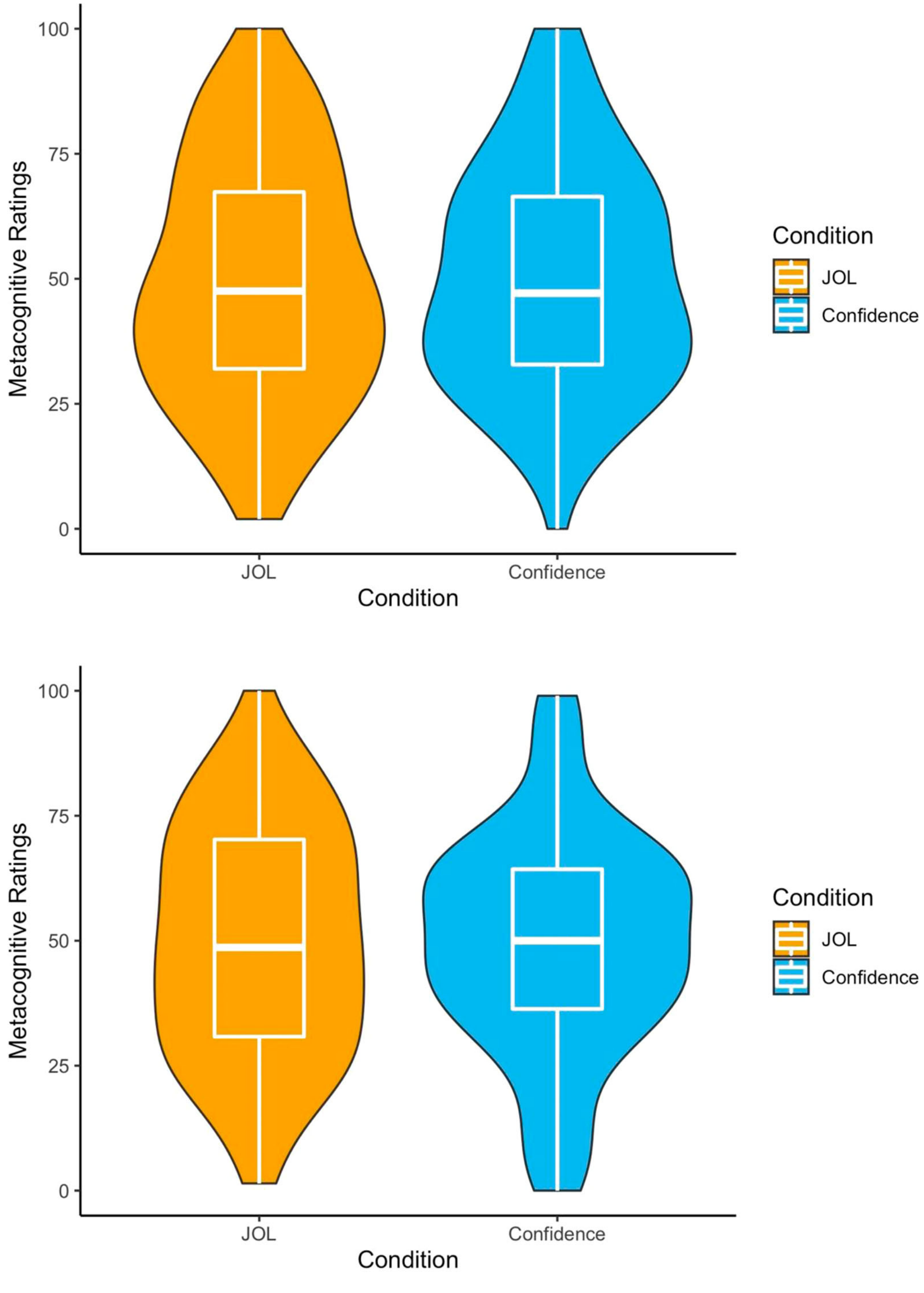

Figure 3. Violin Plots of Metacognitive Ratings in Experiment 2A (Top Panel) and 2B (Bottom Panel).

Note. The violin plot displays the probability density of responses at different values of metacognitive ratings. The central bars of the embedded box plot represent the medians, the upper and lower bounds of the box represent the lower and upper quartiles, and the whiskers represent 1.5 x the interquartile range.

Resolution

In contrast to our hypothesis and the results of Experiment 1, gamma was numerically higher for the confidence group ($M = .67$, $Mdn. = .84$, 95% CI = [.57, .77]) than the JOL group ($M = .64$, $Mdn. = .69$, 95% CI = [.56, .72]), but this difference was not statistically significant, $W = 3679$, $p = .074$, $d = 0.08$.[5] Thus, we found no evidence that subjects were more accurate in predicting future test performance regardless of whether they were making JOLs or retrospective confidence ratings.

Why did this null effect occur? One possibility is that because the subjects may have had background knowledge related to psychology, they were drawing on existing knowledge rather than their episodic experience during the study phase (Glenberg et al., 1987). We conducted an exploratory analysis where we excluded subjects who

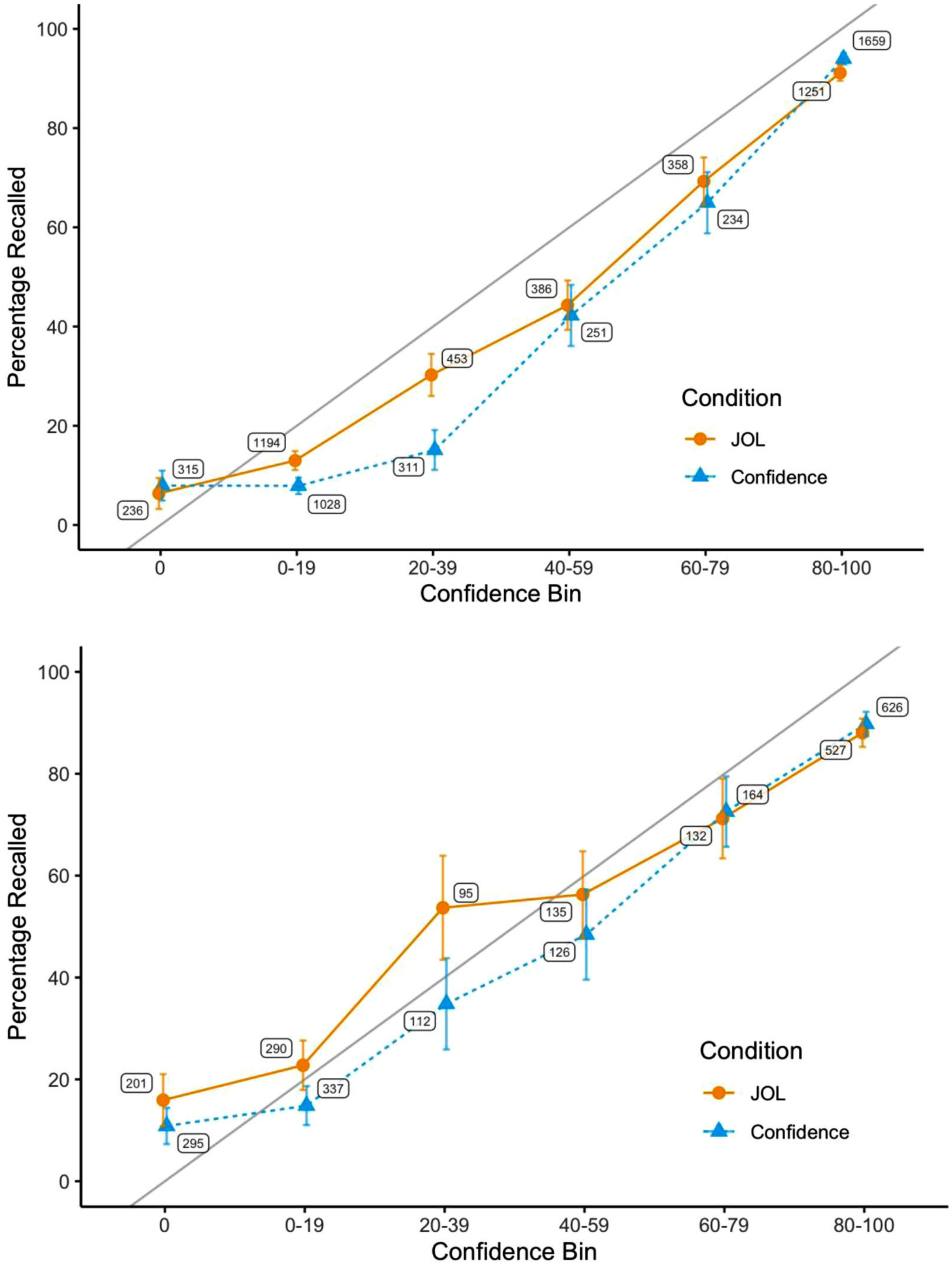

Figure 4. Calibration Curves For Experiment 2A (Top Panel) and 2B (Bottom Panel).

Note: The grey line represents perfect calibration, and the numbers in the labels refer to the number of observations each point represents. Error bars depict 95% confidence intervals.

had taken more than two psychology classes (see supplemental materials for details). When the subjects with prior psychology knowledge were excluded, the confidence group ($M = .71$, *Mdn.* = .83, 95% CI = [.62, .80]) had higher resolution than the JOL group ($M = .64$, *Mdn.* = .69, 95% CI = [.55, .72]), $W = 1451$, $p = .038$, $d = 0.20$. We conducted Experiment 2B to confirm this result.

Experiment 2B

Experiment 2B differed from 2A in two ways. First, we preregistered an analysis plan to exclude subjects who had taken psychology classes. Second, we measured reaction times (RTs) for the first test responses and the metacognitive judgments.

Method

Experiment 2B was preregistered at OSF.IO/VE6Y8.

Subjects. Our final sample had 152 subjects (mean age = 38.45, *SD* = 12.33, 80 female, 71 male, and 1 other), with 83 subjects randomly assigned to the confidence group and 69 subjects assigned to the JOL group. Our goal was to replicate Experiment 2A, which used 175 subjects. Experiment 2A indicated that twenty-five percent of subjects had taken more than two college psychology classes, so we recruited more subjects than needed so that after dropping subjects with more than two psychology classes we would have a final sample of around 175. Our initial sample consisted of 231 subjects. We dropped subjects who had missing data or did not finish the experiment (15), who reported taking two or more psychology classes (56), who reported writing down information (5), and who scored less than 5% on the final test (3), leaving a final sample size of 152. The exclusion criteria were preregistered.

Materials and procedure. The method was identical to Experiment 2A, except that we measured response time for responses on the first test and when making the metacognitive judgment. The analyses related to the reaction time measures are reported in the supplemental materials.

We also included one additional question in 2B that was not included in 2A. At the end of the experiment, we asked subjects "In the middle phase of this experiment you were asked to make a metacognitive judgment—to make a confidence rating about your memory performance. In space below, please briefly describe the internal process you used to make those metacognitive judgments. Specific examples may be helpful," as part of an exploration of whether subjects had any insights into how they made their metacognitive judgments. We describe this outcome in the supplemental materials.

Results and discussion

First test recall. On the first test, the confidence group recalled the same proportion of items ($M = .53$, *Mdn.* $= .50$, $SD = .24$) as did the JOL group ($M = .59$, *Mdn.* $= .55$, $SD = .24$), $W = 2485$, $p = .161$, $d = 0.25$.

Final test recall. The JOL group had similar recall on the final test ($M = .57$, *Mdn.* $= .60$, $SD = .25$) as did the confidence group ($M = .52$, *Mdn.* $= .55$, $SD = .25$), $W = 2570$, $p = .278$, $d = 0.19$. Thus, as in the previous experiments, recall was similar regardless of the type of metacognitive judgment that subjects made.

Metacognitive judgment distribution. As seen in Figure 3 (bottom panel), the confidence ratings made by subjects in the confidence condition were similar to the JOLs made by subjects in the JOL condition, $t(142.2) = 0.16$, $p = .870$, $d = 0.03$, with similar distributions of responses.

Calibration. Figure 4 (bottom panel) displays the calibration curves for Experiment 2B. As can be seen, both groups of subjects again displayed the hard-easy effect, replicating Experiment 2A.

Resolution. As in Experiment 2A, gamma was numerically higher for the confidence group ($M = .78$, *Mdn.* $= .90$, 95% CI = [.71, .85]) than the JOL group ($M = .70$, *Mdn.* $= .76$, 95% CI = [.62, .77]), but this difference was not statistically significant, $W = 2959$, $p = .061$, $d = 0.26$. Note that the effect size is comparable to our exploratory results in Experiment 2A ($d = .20$), and that using Kendall's Tau rather than gamma led to a statistically significant difference ($d = 0.33$).[6] Taken together, these results suggest that there may be a slight advantage of confidence ratings over JOLs in predicting future recall.

Experiment 3

Experiment 1 demonstrated that confidence ratings were more accurate predictors of future recall than JOLs, and Experiment 2 demonstrated that the two measures were equally accurate predictors. This pattern of results suggests that learners either do not attend to the anticipated retention interval when making a JOL, or that doing so can harm their predictions.

In Experiment 3, we attempted to draw subjects' attention to the anticipated retention interval in two ways. First, we extended the actual retention interval from the 10 minutes used in the prior studies to 48 hours. We expected that because people understand (in theory) that forgetting occurs over time, subjects in the JOL condition may be better able to predict their future performance compared to subjects in the confidence condition who are not consistently reminded about this delay. The longer retention interval is also a better representation of learning in classrooms.

Second, because prior work has suggested that subjects may not attend to the projected retention interval when it is manipulated between-subjects (Koriat et al., 2004), we included filler items in the JOL condition so that subjects were exposed to JOL cues that suggested the retention interval would be either 15 minutes or 48 hours (although in practice subjects completed the final test after 48 hours for all target items). Varying the anticipated retention intervals was designed to highlight the retention interval as a potential cue in the JOL condition.

To better investigate the role of retrieval fluency, we also included a condition that was similar to our JOL condition, except that subjects made a JOL based on a cue alone and *then* recalled the target, effectively reversing the order of the two tasks. This "pre-retrieval JOL" is still a delayed JOL (Nelson et al., 2004), but does not require subjects to explicitly attempt recall before making their JOL (although, of course, some subjects may still engage in covert retrieval; Putnam & Roediger, 2013; Smith et al., 2013). We anticipated that metacognitive accuracy in this condition would be lower than the other conditions because subjects were not explicitly asked to retrieve the target.

Overall, we hypothesised that subjects in the post-retrieval JOL condition would predict their final recall performance more accurately than those in the confidence condition, and that subjects in the confidence condition would be more accurate than those in pre-retrieval JOL condition.

Method

Experiment 3 was preregistered at OSF.IO/TQYKC.

Subjects

Expecting attrition from day 1 to day 2 of our study, we aimed to overrecruit our goal sample size by recruiting 650 subjects. After attrition over the two-day delay and excluding subjects based on *a priori* rules (details in the supplemental materials), our final sample consisted of 148 subjects (mean age = 39.06, *SD* = 11.28; 66 female, 78 male, and 4 nonbinary) with 48 subjects in the pre-retrieval JOL condition, 44 in the post-retrieval JOL condition, and 56 in the confidence condition.

Materials and procedure

The materials and procedure for Experiment 3 consisted of the original 20 psychology term-definition pairs used in Experiment 2, with 10 additional filler items taken from the same source. The filler items were not included on the final test.

The procedure was similar to Experiment 2 with three changes. First, during the initial study phase, all subjects studied both the 20 target items and the 10 filler items. Second, the new *pre-retrieval JOL* condition was similar to the JOL condition from Experiment 2, except that the order of recall and making a JOL was reversed: subjects saw a definition alone followed by the prompt “How confident are you that in two days you will be able to recall the term when prompted with the definition?” and then attempted recall.

Third, in the pre-retrieval JOL and post-retrieval JOL conditions, subjects saw one JOL prompt for the target items and a different JOL prompt for the filler items. This variation was intended to highlight the anticipated retention interval in a similar manner to past research (Koriat et al., 2004). For the target pairs, the prompt read, “How confident are you that in 48 hours you will be able to recall the term when prompted with the definition?” In contrast, for the 10 filler pairs, subjects saw, “How confident are you that in 15 minutes you will be able to recall the term when prompted with the definition?” Thus, subjects in the two JOL conditions considered two different anticipated retention intervals. In the confidence condition, subjects saw both the target and filler items during the first test phase, but as there was no retention interval presented as part of the confidence prompt, these items were essentially treated identically.

After completing the intial study and first test/metacognitive rating phase, subjects were dismissed (in other words, there was no test in 15 minutes for the filler items). Two days later, subjects received an email invitation to take the final test, which consisted of the twenty target vocabulary terms (the filler items were not tested).

Results and discussion

We conducted all analyses on only the 20 target items. In addition to calculating resolution with gamma, we also preregistered our decision to examine Kendall’s Tau (Dougherty et al., 2018). We also report exploratory measures of resolution with G_{trap}.

Test 1 recall

As in the prior studies, the recall rates for the first test were similar among the confidence ($M = .48$, *Mdn.* = .48, $SD = .20$), pre-retrieval JOL ($M = .46$, *Mdn.* = .45, $SD = .22$), and post-retrieval JOL ($M = .47$, *Mdn.* = .43, $SD = .25$) conditions, $F(2,145) = 0.05$, $p = .947$, $\eta^2 < .001$.

Final test recall

As in the prior studies, the recall rates were similar across conditions for the final test, with subjects recalling a similar proportion of items in the pre-retrieval JOL condition ($M = .41$, *Mdn.* = .43, $SD = .21$), the post-retrieval JOL condition ($M = .45$, *Mdn.* = .40, $SD = .24$), and the confidence condition, ($M = .41$, *Mdn.* = .43, $SD = .21$), $F(2,145) = 0.78$, $p = .459$, $\eta^2 = .01$, suggesting that, overall, recall was similar for the three groups.

Metacognitive judgment distributions

As seen in Figure 5, the metacognitive ratings differed as a function of group, $F(2,145) = 3.52$, $p = .032$, $\eta^2 = .05$. Games-Howell post-hoc tests revealed that the confidence group was more confident than the pre-retrieval group ($p = .046$), but that all other comparisons were nonsignificant. The distribution of the responses shows that the higher ratings in the confidence group were driven by a larger number of very high ratings.

Calibration

Figure 6 displays calibration curves for the three conditions predicting performance on the final test. All groups were reasonably well calibrated, and continued to show the hard-easy effect.

Resolution

In contrast to our predictions, gamma was similar for all three conditions, $H(2) = 1.78$, $p = .401$. The pre-retrieval JOL group, ($M = .63$, *Mdn.* = .72, 95% CI = [.54, .71]), was just as accurate as the post-retrieval JOL group, ($M = .69$, *Mdn.* = .77, 95% CI = [.61, .78]), and the confidence group, ($M = .67$, *Mdn.* = 72, 95% CI = [.60, .75]). The result was

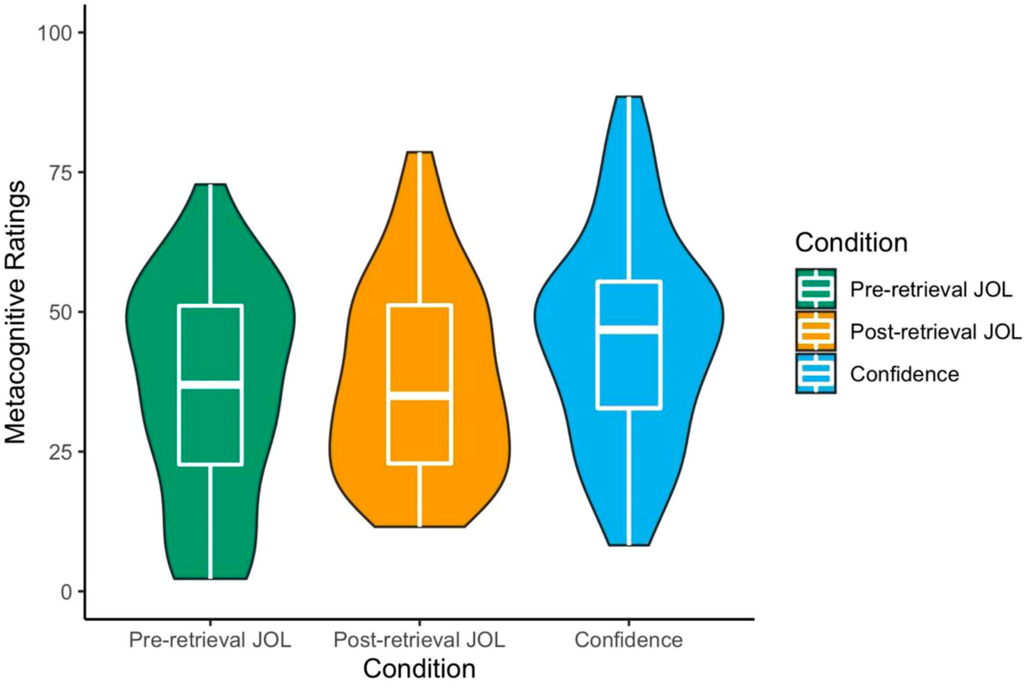

Figure 5. Violin Plots of Metacognitive Ratings in Experiment 3.

Note. The violin plot displays the probability density of responses at different values of metacognitive ratings. The central bars of the embedded box plot represent the medians, the upper and lower bounds of the box represent the lower and upper quartiles, and the whiskers represent 1.5 x the interquartile range.

similar when we used Kendall's tau instead of gamma, $H(2) = 1.78$, $p = .41$.[7]

While the lack of a difference between the post-retrieval JOL and confidence groups was expected, we were somewhat surprised to find that the pre-retrieval JOL group had similar resolution to the other groups. This is because we expected that subjects in the pre-retrieval JOL group would sometimes but not always engage in covert retrieval. However, the equal resolution performance across groups suggests one possibility: namely, that subjects in

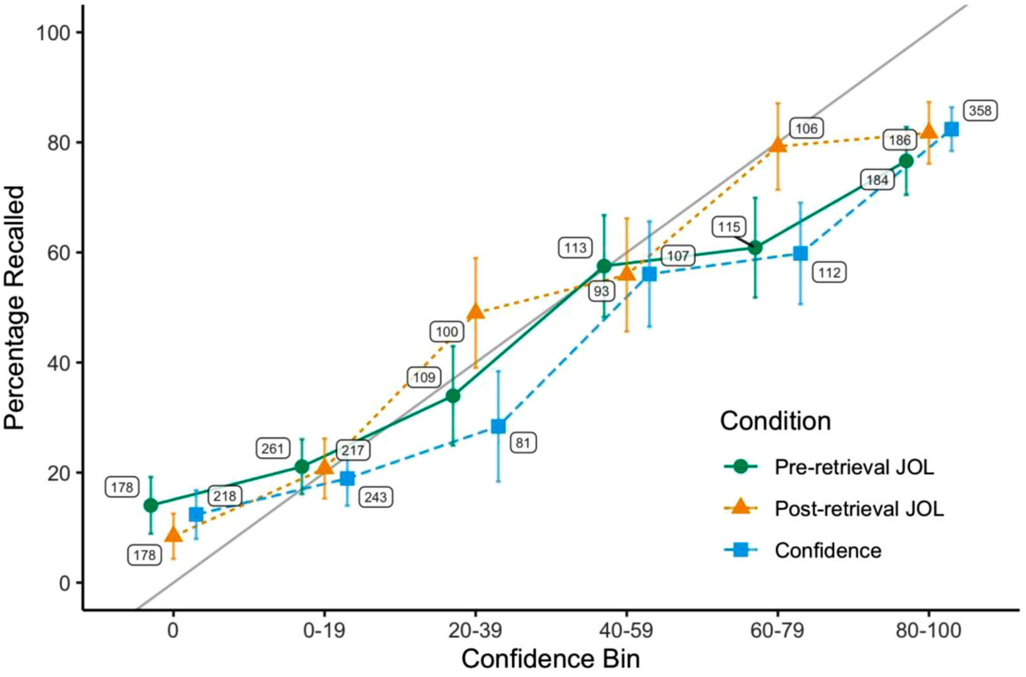

Figure 6. Calibration Curve for Experiment 3.

Note: The grey line represents perfect calibration, and the numbers in the labels refer to the number of observations each point represents. Error bars depict 95% confidence intervals.

the pre-retrieval JOL group were covertly retrieving the item responses consistently when making their JOLs. In sum, Experiment 3 provided no evidence that JOLs were more accurate than confidence ratings or vice versa.

Summary of experiment 3

Experiment 3 replicated Experiment 2 in showing that the metacognitive accuracy of the different conditions was similar regardless of whether people were directed to think about the past or the future. This lack of a difference is somewhat surprising given that we used a longer retention interval (two days) and explicitly highlighted the anticipated retention interval in the post-retrieval JOL condition by having two different delay lengths. This pattern of results suggests that providing the anticipated retention interval in a JOL prompt did not improve prediction accuracy compared to simply making a confidence rating. This may be because subjects were not attenting to the retention interval cue at all, were weighing it lightly, or did not use the cue consistently. One caveat regarding Experiment 3 is that our attrition rate from day 1 to day 2 was much higher than expected due to subjects not returning for the second session.

Mini meta-analysis of all experiments

Because of the inconsistent findings across our four studies (including small effect sizes and *p* values near .05), we used the metafor package in R (Viechtbauer, 2010) to conduct an exploratory mini meta-analysis of our three experiments examining our central question of whether confidence ratings or JOLs better predicted final test performance (Cumming, 2012). We used a random-effects meta-analysis to compare differences in gamma for the confidence ratings and JOLs, with larger numbers representing an advantage of confidence ratings over JOLs. For Experiment 3, we compared gamma for the post-retrieval condition and the confidence condition to match the prior studies. As Figure 7 suggests, the meta-analysis revealed a small advantage of confidence ratings over JOLs, $M_{\text{diff}} = 0.088$, 95% CI = [0.004, 0.172]. The confidence interval itself was reasonably small, suggesting that our studies led to a fairly precise estimate of the difference in predictive power. As part of the meta-analysis, we wondered whether material type would serve as a moderator, with larger effect sizes occurring with paired associates and smaller effects sizes occurring with the vocabulary materials. However, the model was homogeneous, revealing little variation across studies (T2 = .002).

Because of the concerns with gamma as a measure of resolution, we replicated the meta-analysis using Kendall's Tau and G_{trap} instead of gamma (details in the supplemental materials). The effect size estimates for Kendall's Tau, $M_{diff} = 0.093$, 95% CI = [0.035, 0.150], and G_{trap}, $M_{diff} = 0.085$, 95% CI = [0.021, 0.148] (See Figure S5 and Figure S6), were comparable to the analysis using gamma, but with smaller confidence intervals, again suggesting a small advantage of confidence judgments compared to JOLs.

Additionally, we also examined whether confidence ratings were consistently higher than JOLs across all three experiments. If this were the case, it would suggest that subjects were consistently lowering their metacognitive ratings to account for forgetting across the retention interval. An exploratory meta-analysis examining metacognitive ratings revealed the confidence ratings were similar to the JOL ratings, $M_{diff} = 3.72$, 95% CI = [−0.04, 7.47] (see Figure S7). This outcome suggests that subjects were not systematically lowering their judgments to account for forgetting. Rather, because resolution is lower in the JOL condition, it suggests that considering the retention interval was adding noise to the JOL judgment process.

General discussion

The goal of the current study was to understand how thinking about the future—in terms of a JOL's anticipated retention interval—affects prediction accuracy. Experiment 1 showed that retrospective confidence ratings better predicted final test performance than JOLs, whereas Experiments 2 and 3 (which used educationally relevant materials) showed that both judgments led to similarly accurate predictions. A mini meta-analysis provided evidence that confidence ratings have a small advantage over JOLs.

This set of studies have several contributions. First, it demonstrates a counterintuitive finding that confidence ratings are more accurate predictors of future performance than JOLs. Second, it demonstrates that the judgment principles in play apply to both laboratory and educationally relevant materials (although see further thoughts on this issue below). Finally, it demonstrates that the retention interval in the JOL prompt is adding noise to the judgment process, rather than a simple downwards adjustment of the metacognitive ratings to account for forgetting over the retention interval. Below, we elaborate on some of these points.

Our meta-analysis suggested that confidence ratings may have a small advantage over JOLs in terms of predicting future recall. Why might this be the case, when JOLs explicitly ask people to think about the future? Our interpretation is as follows: confidence ratings are based heavily on retrieval fluency and past retrieval attempts, both of which are valid and diagnostic cues (at least, when non-deceptive items are used; Benjamin & Bjork, 1996; Koriat & Goldsmith, 1996; Kelly & Lindsay, 1993). Since past performance predicts future performance (Estes et al., 1960; Tulving, 1964), confidence ratings are not only accurate in gauging performance on the just-completed-test, but in predicting future performance. JOLs and confidence ratings are thought to use similar processes, with the key exception being that JOLs explicitly

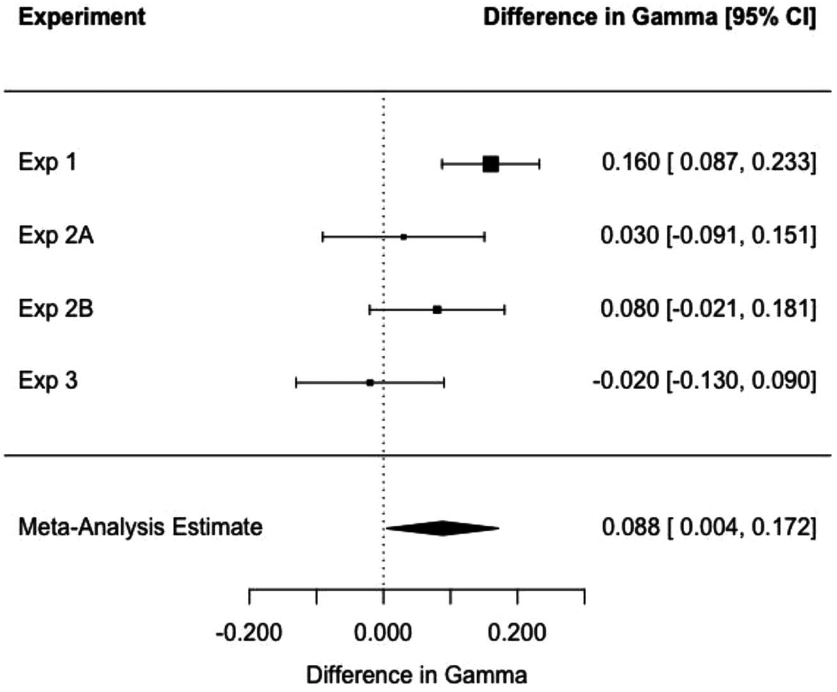

Figure 7. Forest Plot Depicting Effect Sizes and 95% confidence intervals from Experiments 1-3.

Note. The size of the points indicates the weighting of the study in the meta-analysis. A larger effect size indicates an advantage of confidence over JOL in predicting future recall as measured by Goodman-Kruskal gamma correlations.

ask people to consider the future. However, doing so results in poorer resolution. This outcome is consistent with the JOL = confidence + variation hypothesis (Dougherty et al., 2005). Notably, a natural prediction here would be that all of the JOL ratings would be lower than the confidence ratings. If this were the case, then resolution, a relative measure of accuracy, would not have been affected. What we found was a reduction in resolution without any change in the ratings themselves, suggests that considering the retention interval adds noise to the JOL judgment. One way of conceptualising this outcome is that confidence ratings are fairly accurate predictors (mostly signal) but that including the retention interval as a cue adds noise to that process.

The somewhat counterintuitive pattern we see in our experiments is consistent with predictions from Koriat's self-consistency model (2012). The self-consistency model suggests that internal consistency is a strong predictor of confidence ratings. Converging cues lead to high confidence, and a major cue here is a sense of reproducibility, or the likelihood of producing the same answer in the future. Koriat (2012) reviewed evidence demonstrating that response speed predicts both question accuracy and confidence, but that both relationships are mediated by self-consistency. Thus, confidence ratings are related to past retrieval attempts and index producing the same answer in the future.

An issue to consider is that the effect size was much stronger in Experiment 1 ($d = 0.60$) compared to the other studies that used educationally relevant materials ($d = 0.06$ to 0.26). One possibility is that, when using educationally relevant materials, subjects might be drawing on their general familiarity of a domain rather than the episodic experience of studying the materials (Glenberg et al., 1987). In other words, subjects might be using prior knowledge rather than relying on the episodic experience of having studied the items earlier. Alternatively, past studies have shown that JOL accuracy is higher for related target-response pairs compared to unrelated ones. Thus, the inherently strong association between our term-definition pairs might have masked any potential differences between the JOL and confidence rating accuracy (Susser & Mulligan, 2019). This raises question about the generalizability of other metacognitive research based on laboratory materials—depending on the circumstances, the same principles might not apply when subjects have prior knowledge or the material is highly organised. Future research should continue to extend basic findings from lab-based metacognition tasks to educationally-relevant materials (Rhodes, 2016).

Limitations

There are three limiting conditions of our studies. First, our implementation of JOLs—not only delayed, but after a retrieval attempt—are unusual compared to standard methods (Dunlosky & Metcalfe, 2009). Of course, immediate JOLs are less accurate than delayed JOLs (Nelson & Dunlosky, 1991), and explicit retrieval instructions (instead of subjects covertly retrieving answers) will likely increase both metacognitive accuracy and final test performance. Notably, Experiment 3 suggests that subjects were engaging in covert retrieval when making a delayed JOL and that making a JOL before or after an explicit retrieval attempt led to similar accuracy. Critically, our procedure allowed us to directly measure how much thinking about the future helped prediction accuracy (or rather, how much it did not help).

Second, it is certainly possible that the specific type of materials used could influence the relative accuracy of different kinds of judgments. Thus, further research should continue to explore whether the relative accuracy of confidence ratings and JOLs show consistent patterns with more complex materials.

Finally, in our studies, the actual best predictor of future test performance might be performance on the practice test. Indeed, Table S1 shows that subjects were unlikely to recall new items or forget items between the first and final test. Thus, a caveat is while the the current study focused on evaluating different kinds of metacognitive judgments, from a practical perspective, initial recall may be the most important element in predicting future recall.

Implications

The pedagogical implications of our studies are clear: if learners are trying to predict future test performance, they should not make JOLs. Even delaying JOLs, which normally enhances preditions, might not significantly increase the accuracy of predictions related to more realistic materials (Thomas et al., 2016). Instead, learners should complete a practice test and make confidence ratings on that test. Of course, taking a practice test will directly enhance memory and improve metacognitive monitoring; in many ways, the advice to take the practice test is more important than any advice related to improving metacognitive predictions (Roediger et al., 2011). However, students are often rushed for time, and confidence ratings are a quick, easy, and relatively accurate way to assess comprehension compared to looking up the answers to a practice quiz.

Concluding comments

As several psychologists – including even one who hosts daytime talk shows and TV series (Franklin, 2013) – have been purported to say, “The best predictor of future behavior is past behavior,” and the current studies are consistent with this idea. Focusing exclusively on past test performance may be a better way to predict future test performance than trying to explicitly think about the future. Returning to the example of a furiously cramming student from the introduction, our suggestion for that student is to simplify their task—instead of trying to predict the future, they should try and evaluate their past performance.

Notes

1. We conducted analyses using the data from all subjects. An analysis with just the preregistered subjects—described in the supplemental materials—yielded the same outcomes.
2. Thanks to an anonymous reviewer for suggesting this alternative computation for gamma.
3. Resolution for first test recall is reported in the supplemental materials.
4. Analyses using Kendall’s Tau, $p = .004$, $d = 0.80$, and G_{trap} (a measure of gamma based on ROC curves), $p < .001$, $d = 0.61$, yielded similar outcomes.
5. We originally preregistered a one-tailed test (which was significant) but report the more conservative two-tailed test here. Also, note that indexing resolution with Kendall’s Tau, $p = .009$, $d = 1.12$, or G_{trap}, $p = .040$, $d = 0.26$, led to statistically significant differences with two-tailed tests.
6. Using G_{trap} led to an outcome similar to the analysis reported here with gamma, with $p = .060$ $d = 0.29$; see supplemental materials for details.
7. Using G_{trap} to measure resolution yielded identical outcomes, $p = .498$. See supplemental materials for details.

Acknowledgements

We thank the Carleton College Psychology Department and Furman University for funding this research. We also thank Grace Gilmore and Peter Dehkes for their essential work on Experiments 1 and 2A and Jae Eun Lee, Lucia Ray, and Hailey Rinella for their assistance.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The author(s) reported there is no funding associated with the work featured in this article.

ORCID

Adam L. Putnam http://orcid.org/0000-0002-8708-9268
K. Andrew DeSoto http://orcid.org/0000-0001-9061-0301

References

Arnold, K. M., & McDermott, K. B. (2013). Test-potentiated learning: Distinguishing between direct and indirect effects of tests. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *39*(3), 940–945. https://doi.org/10.1037/a0029199

Benjamin, A. S., & Bjork, R. A. (1996). Retrieval fluency as a metacognitive index. In L. M. Reder (Ed.), *Implicit memory and metacognition* (pp. 309–338). Erlbaum.

Busey, T. A., Tunnicliff, J., Loftus, G. R., & Loftus, E. F. (2000). Accounts of the confidence-accuracy relation in recognition memory. *Psychonomic Bulletin & Review*, *7*(1), 26–48. https://doi.org/10.3758/BF03210724

Cumming, G. (2012). *Understanding the new statistics*. Routledge.

Diaz, M., & Benjamin, A. S. (2011). The effects of proactive interference (PI) and release from PI on judgments of learning. *Memory &*

Cognition, *39*(2), 196–203. https://doi.org/10.3758/s13421-010-0010-y

Dougherty, M. R., Robey, A. M., & Buttaccio, D. (2018). Do metacognitive judgments alter memory performance beyond the benefits of retrieval practice? A comment on and replication attempt of dougherty, scheck, Nelson, and narens (2005). *Memory & Cognition*, *46*(4), 558–565. https://doi.org/10.3758/s13421-018-0791-y

Dougherty, M. R., Scheck, P., Nelson, T. O., & Narens, L. (2005). Using the past to predict the future. *Memory and Cognition*, *33*(6), 1096–1115. http://doi.org/10.3758/BF03193216

Dunlosky, J., & Metcalfe, J. (2009). *Metacognition*. SAGE.

Dunlosky, J., & Tauber, S. K. (2014). Understanding people's metacognitive judgments: An isomechanism framework and its implications for applied and theoretical research. In T. Perfect, & D. S. Lindsay (Eds.), *Handbook of applied memory* (pp. 444–463). Sage.

England, B. D., Ortegren, F. R., & Serra, M. J. (2017). Framing affects scale usage for judgments of learning, not confidence in memory. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *43*(12), 1898–1908. https://doi.org/10.1037/xlm0000420

Estes, W. K., Hopkins, B. L., & Crothers, E. J. (1960). All-or-none and conservation effects in the learning and retention of paired associates. *Journal of Experimental Psychology*, *60*(6), 329–339. http://doi.org/10.1037/h0043400

Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. SAGE.

Finn, B. (2008). Framing effects on metacognitive monitoring and control. *Memory & Cognition*, *36*(4), 813–821. https://doi.org/10.3758/MC.36.4.813

Franklin, K. (2013, January 3) *"The Best Predictor of Future Behavior Is … Past Behavior": Does the popular maxim hold water?* Psychology Today. https://www.psychologytoday.com/us/blog/witness/201301/the-best-predictor-future-behavior-is-past-behavior

Glenberg, A. M., Sanocki, T., Epstein, W., & Morris, C. (1987). Enhancing calibration of comprehension. *Journal of Experimental Psychology. General*, *116*(2), 119–136. https://doi.org/10.1037/0096-3445.116.2.119

Higham, P. A., & Higham, D. P. (2019). New improved gamma: Enhancing the accuracy of Goodman-Kruskal's gamma using ROC curves. *Behavior Research Methods*, *51*, 108–125. https://doi.org/10.3758/s13428-018-1125-5

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. https://doi.org/10.1006/jmla.1993.1001

Kimball, D., & Metcalfe, J. (2003). Delaying judgments of learning affects memory, not metamemory. *Memory and Cognition*, *31*(6), 918–929. https://doi.org/10.3758/BF03196445

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. https://doi.org/10.1037/0096-3445.126.4.349

Koriat, A. (2012). The self-consistency model of subjective confidence. *Psychological Review*, *119*(1), 80–113. http://doi.org/10.1037/a0025648

Koriat, A. (2015). When two heads are better than one and when they can be worse: The amplification hypothesis. *Journal of Experimental Psychology: General*, *144*(5), 934–950. http://doi.org/10.1037/xge0000092

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. http://doi.org/10.1037/0096-3445.133.4.643

Koriat, A., & Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review*, *103*(3), 490–517. https://doi.org/10.1037/0033-295X.103.3.490

Lovelace, E. A. (1984). Metamemory: Monitoring future recallability during study. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(4), 756–766. https://doi.org/10.1037/0278-7393.10.4.756

Luna, K., Martín-Luengo, B., & Albuquerque, P. B. (2018). Do delayed judgements of learning reduce metamemory illusions? A meta-analysis. *Quarterly Journal of Experimental Psychology*, *71*(7), 1626–1636. https://doi.org/10.1080/17470218.2017.1343362

Lyon, T. D., & Flavell, J. H. (1993). Young children's understanding of forgetting over time. *Child Development*, *64*(3), 789–800. https://doi.org/10.2307/1131218

McDaniel, M. A., & Bugg, J. M. (2008). Instability in memory phenomena: A common puzzle and a unifying explanation. *Psychonomic Bulletin & Review*, *15*(2), 237–255. https://doi.org/10.3758/PBR.15.2.237

Metcalfe, J., & Finn, B. (2008). Evidence that judgments of learning are causally related to study choice. *Psychonomic Bulletin & Review*, *15* (1), 174–179. https://doi.org/10.3758/PBR.15.1.174

Murayama, K., Sakaki, M., Yan, V. X., & Smith, G. M. (2014). Type I error inflation in the traditional by-participant analysis to metamemory accuracy: A generalized mixed-effects model perspective. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *40*(5), 1287–1306. https://doi.org/10.1037/a0036914

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin*, *95* (1), 109–133. http://doi.org/10.1037/0033-2909.95.1.109

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. https://doi.org/10.1111/j.1467-9280.1991.tb00147.x

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. https://doi.org/10.1037/1082-989X.9.1.53

Pierce, B. H., & Smith, S. M. (2001). The postdiction superiority effect in metacomprehension of text. *Memory & Cognition*, *29*(1), 62–67. https://doi.org/10.3758/BF03195741

Putnam, A. L., Ozubko, J. D., MacLeod, C. M., & Roediger, H. L., III. (2014). The production effect in paired-associate learning: Benefits for item and associative information. *Memory & Cognition*, *42*(3), 409–420. https://doi.org/10.3758/s13421-013-0374-x

Putnam, A. L., & Roediger, H. L., III. (2013). Does response mode affect amount recalled or the magnitude of the testing effect? *Memory & Cognition*, *41*(1), 36–48. https://doi.org/10.3758/s13421-012-0245-x

Rawson, K. A., Dunlosky, J., & Mc Donald, S. L. (2002). Influences of metamemory on performance predictions for text. *The Quarterly Journal of Experimental Psychology*, *55*(2), 505–524. https://doi.org/10.1080/02724980143000352

Rhodes, M., & Castel, A. D. (2008). Memory predictions are influence by perceptual information: Evidence for metacognitive illusions. *Journal of Experimental Psychology: General*, *137*(4), 612–625. http://doi.org/10.1037/a0013684

Rhodes, M. G. (2016). Judgments of learning. In J. Dunlosky, & S. K. Tauber (Eds.), *The Oxford Handbook of metamemory* (pp. 1–32). Oxford University Press.

Roediger, H. L., III, Putnam, A. L., & Smith, M. A. (2011). Ten benefits of testing and their applications to educational practice. In J. P. Mestre & B. H. Ross (Eds.), *The psychology of learning and motivation: Cognition in education, Vol. 55*, (pp. 1–36). Elsevier Academic Press. https://doi.org/10.1016/B978-0-12-387691-1.00001-6

Serra, M. J., & England, B. D. (2012). Magnitude and accuracy differences between judgements of remembering and forgetting. *Quarterly Journal of Experimental Psychology*, *65*(11), 2231–2257. https://doi.org/10.1080/17470218.2012.685081

Smith, M. A., Roediger, H. L., & Karpicke, J. D. (2013). Covert retrieval practice benefits retention as much as overt retrieval practice.

Journal of Experimental Psychology. Learning, Memory, and Cognition, *39*(6), 1712–1725. https://doi.org/10.1037/a0033569

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. https://doi.org/10.1111/j.1467-9280.1992.tb00680.x

Susser, J. A., & Mulligan, N. W. (2019). Exploring the intrinsic-extrinsic distinction in prospective metamemory. *Journal of Memory and Language*, *104*, 43–55. https://doi.org/10.1016/j.jml.2018.09.003

Tauber, S. K., & Rhodes, M. G. (2010). Does the amount of material to be remembered influence judgements of learning (JOLS)? *Memory*, *18*(3), 351–362. https://doi.org/10.1080/09658211003662755

Tauber, S. K., & Rhodes, M. G. (2012). Measuring memory monitoring with judgements of retention (JORs). *The Quarterly Journal of Experimental Psychology*, *65*(7), 1376–1396. https://doi.org/10.1080/17470218.2012.656665

Thomas, R. C., Finn, B., & Jacoby, L. L. (2016). Prior experience shapes metacognitive judgments at the category level: The role of testing and category difficulty. *Metacognition and Learning*, *11*(3), 257–274. https://doi.org/10.1007/s11409-015-9144-4

Tulving, E. (1964). Intratrial and intertrial retention: Notes towards a theory of free recall verbal learning. *Psychological Review*, *71*(3), 219–237. http://doi.org/10.1037/h0043186

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, 36(3), 1-48. https://www.jstatsoft.org/v36/i03 https://doi.org/10.18637/jss.v036.i03