Memory & Cognition (2018) 46:558–565
https://doi.org/10.3758/s13421-018-0791-y

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)

Michael R. Dougherty[1] · Alison M. Robey[1] · Daniel Buttaccio[1]

Published online: 24 January 2018

Abstract

A central question in the metacognitive literature concerns whether the act of making a metacognitive judgment alters one's memory for the information about which the judgment was made. Dougherty, Scheck, Nelson, and Narens (2005, *Memory & Cognition, 33*(6), 1096–1115) attempted to address this question by having participants make either retrospective confidence judgments (RCJs; i.e., evaluations of past retrieval success), judgments of learning (JOLs; i.e., predictions of future retrieval success), or no explicit judgments. When comparing final retrieval accuracy they found that accuracy was greater for items where participants had made JOLs compared with items that received RCJs or no judgment, suggesting that simply making a JOL can improve later memory performance. The present article presents results from four separate replication attempts that fail to duplicate this finding. Combined results provide compelling evidence that making a metacognitive judgment, regardless of the type, has no impact on later memory performance above and beyond retrieval practice.

Keywords Metamemory · Recall · Judgment of learning · Reactivity · Retrospective · Confidence judgment

A central question in the metacognition literature regards the effect of metacognitive judgment on memory performance: Does making a metacognitive judgment alter the very thing it is designed to assess? Data regarding this question can be traced back to original work by Arbuckle and Cuddy (1969) who found that participants who made judgments of learning (JOLs) ultimately performed better on memory tests compared with a no-judgment control group. While a variety of studies have provided data pertinent to the question of whether judgment alters memory (e.g., Keleman & Weaver, 1997; Kimball & Metcalfe, 2003; Sommer, Heinz, Leuthold., Matt, & Schweinberger, 1995; Spellman & Bjork, 1992; Tauber & Rhodes, 2012), to our knowledge there has been no systematic attempt to answer this question unambiguously. The goal of the study presented herein represents our attempt to determine if judgment influences memory, and if so, to identify the mechanisms involved.

Dougherty, Scheck, Nelson, and Narens (2005) conducted two studies comparing the relative accuracy of JOLs and retrospective confidence judgments (RCJs) for predicting future recall. In their Study 2, participants were randomly assigned to one of three groups: A JOL group, an RCJ group, and a no-judgment group. Participants studied a list of items (paired associates) during an initial learning phase. After a short filled delay designed to clear the contents of working memory, participants were given a prejudgment recall task in which they were presented with a cue word and asked to retrieve its associate. After retrieving each item, participants were then asked to provide a confidence judgment (RCJ and JOL conditions), or they gave no judgment (no-judgment condition). Participants who made RCJs simply rated their confidence in the item they had just retrieved at prejudgment recall being correct, whereas participants in the JOL condition were asked to rate their confidence that they would be able to recall the target word when tested again later. The study, prejudgment recall, and confidence judgment tasks repeated multiple times before subjects were given a "final" cued recall test, which took place approximately 10 minutes after the judgment task. Two key findings emerged from that study. First, retrospective confidence judgments were more highly correlated with final recall than were JOLs. Second, participants who made JOLs

✉ Michael R. Dougherty
mdougher@umd.edu

[1] Department of Psychology, University of Maryland, College Park, MD 20742, USA

performed *better* on the final test than did participants in both the RCJ and no-judgment conditions. This later finding is particularly relevant for the present article because it revealed a curious result: Not only did judgment alter memory, but only a particular type of judgment—JOLs—appeared to alter memory.

The study by Dougherty and colleagues was unique in that it is one of the few studies designed in such a way to test whether judgment influenced memory independent of retrieval practice. While previous studies have shown reactive effects of JOLs on memory performance, much of this work has not isolated the unique effects of judgment above and beyond retrieval practice. Researchers have speculated that the effect of judgment on memory might operate by encouraging participants to engage in retrieval before making a judgment. Indeed, early work in the JOL literature substantiates this hypothesis: Nelson and Dunlosky (1991) reported that many of their subjects attempted to retrieve the target before making their JOL. This retrieval attempt may serve to improve the quality of the underlying memory trace, leading to better retrieval later. The study by Dougherty et al. (2005) addressed this issue by requiring participants in all three conditions to engage in retrieval before making their judgments and by varying the type of metacognitive judgment, equating the retrieval practice benefits across conditions. The key finding reported in Dougherty et al. (2005) was that JOLs led to improved retrieval compared with a no-judgment condition. In contrast, having participants make RCJs did not improve memory over and above the no-judgment condition. The fact that there were differences across conditions, even after experimentally controlling for prejudgment retrieval, suggests that there is something special about JOLs (but not RCJs) that may lead to enhanced memory beyond the benefits of retrieval practice. Dougherty et al. (2005) speculated that the nature of the JOL task, which encourages participants to think about future retrieval, may change the way participants approach the learning phase. This point, along with a call for following up on this finding, was reiterated by Rhodes (2016). While the results of the Dougherty et al. (2005) study were promising, it is important to note that the study was not designed with the intention of testing the effect of judgment on memory—the finding was serendipitous, and therefore it is appropriate to place a skeptical prior on the original finding. All skepticism aside, Rhodes (2016) explicitly cited the findings in Dougherty et al. (2005) in his call for "an agenda for future research (a) to include appropriate control conditions to assess the impact of prediction on memory, and (b) to provide a viable explanation for such reactivity (p. 76)." In this context, the paradigm used by Dougherty et al. (2005), which we replicate herein, was designed to examine the effect of judgment on memory while controlling for retrieval practice.

We calculated all pairwise Bayes factors based on the summary statistics provided in Table 3 of Dougherty et al. (2005). The Bayes factor provides a probabilistic statement of the degree to which the data support a particular model, which in our case is either the null or alternative model. For example, $\text{BF}_{10} = 4$ would represent 4 to1 odds in favor of the alternative hypothesis, and a $\text{BF}_{10} = 0.25$ would represent 4 to 1 odds in favor of the null hypothesis.[1] Though several of the comparisons yield t values that are statistically significant at the $p <$ .05 level (Table 1, bottom triangles), very few of these "effects" are convincing according to the Bayes factor (Table 1, upper triangles). For example, comparing the JOL condition to the RCJ condition on final retrieval yielded $t(122) = 2.12$, $p = .036$, and $t(122) = 2.40$, $p = .018$, when participants were given 3 seconds and 12 seconds for study, respectively. These two t values correspond to $\text{BF}_{10} = 1.43$ and $\text{BF}_{10} = 2.51$, respectively—values that typically are interpreted as noninformative because the data do not provide much evidence for either the alternative or null hypothesis. For comparison, a value of 1.0 is interpreted as equal odds (50% chance) that the null versus alternative is true. Findings in which statistically significant effects provide little or no evidence for the alternative versus the null hypothesis are surprisingly common in experimental psychology (see Wetzels, Matzke, Lee, Rouder, Iverson, & Wagenmakers, 2011) and reflect the fact that p values tend to overstate the evidence against the null hypothesis (Rouder, Speckman, Sun, Morey, & Iverson, 2009). The majority of analyses that yield strong evidence for the alternative are those that include comparing the JOL condition to a no-judgment condition; most analyses comparing JOLs to RCJs yield weak evidence for either the alternative or null hypothesis. Because the analyses are not all independent of one another (i.e., all analyses include the same subjects, and the dependent variables are correlated), one would not be justified in drawing strong conclusions from any subset of the analyses presented in Table 1.

Given the analyses presented in Table 1, it is fair to state that the data provided by Dougherty et al. (2005) are far from convincing regarding the effect of JOLs on recall accuracy and that the question of whether JOLs (but not RCJs) alter judgment above and beyond retrieval practice is not yet fully resolved. Thus, we decided to replicate the general effects of Dougherty et al. (2005) within the context of an experimental paradigm that would enable us to test a potential mechanism of the effect—if it indeed exists. The analyses presented herein include data from two new experiments with total sample sizes of $N = 151$ and $N = 138$, as well as data drawn from two previously published experiments that used a very similar experimental paradigm (see Robey, Buttaccio, & Dougherty, in press). We first present the data from one of the new

[1] Although Bayes factors are meant to be interpreted along a continuum rather than with cut points, general guidelines for interpreting Bayes factors are as follow: 1.01 to 3: anecdotal support (inconclusive); 3 to 10: moderate support; 10 to 100: strong support; >100: decisive support (cf. Jeffreys, 1961).

Table 1 Pairwise Bayes factors and *t* values for Dougherty et al. (2005)

	JOL	RCJ	No judgment		JOL	RCJ	No judgment
3 second				12 second			
Prejudgment recall accuracy							
JOL	–	0.35	2.27	JOL	–	0.47	1.44
RCJ	1.17	–	0.36	RCJ	1.41	–	0.24
No judgment	2.35^{*}	1.18	–	No judgment	2.12^{*}	0.71	–
Final recall accuracy							
JOL	–	1.43	12.09	JOL	–	2.51	45.37
RCJ	2.12^{*}	–	0.29	RCJ	2.40^{*}	–	0.23
No judgment	3.06^{**}	0.94	–	No judgment	3.53^{**}	0.6	–
Conditional recall							
JOL	–	2.51	3.82	JOL	–	344.55	143.78
RCJ	2.40^{*}	–	0.196	RCJ	4.16^{**}	–	0.20
No judgment	2.59^{*}	0.20	–	No judgment	3.90^{**}	0.24	–

Note. Upper triangle provides BFs; lower triangle provides pairwise independent-sample *t* tests. $N = 62$ (JOL), $N = 62$ (RCJ), $N = 60$ (no judgment). JOL = judgments of learning; RCJ = retrospective confidence judgments. * .05, ** .01

experiments and then follow these analyses with a mini meta-analysis that includes data from all four studies (for a total sample size of 600).

Experiment overview

The experimental paradigm closely followed the method outlined in Dougherty et al. (2005), with a small number of modifications that we felt were either justified or inconsequential. First, rather than manipulate study duration across two levels (3 seconds and 12 seconds) as a within-subjects factor, we chose to present all items for 5 seconds. Second, a dot-probe task was conducted simultaneously with the study phase. This task was included as a measure of attention. We reasoned that if making JOLs prompted participants to engage in more elaborative rehearsal during study, then they should show slowed reaction time on the dot-probe trials relative to non-dot-probe trials. Third, we made a slight alteration to the metacognitive monitoring instructions. In the JOL condition of Dougherty et al. (2005), participants were presented with the following instructions: “How confident are you that in about 10 minutes you will be able to recall the second word of the item when prompted with the first?” whereas, in the new study, JOL participants were instructed “How likely would you be to retrieve Word 2 again on a future recall test at the end of the study?”. Fourth, rather than have participants make metacognitive judgments on a 6-point scale ranging from 0% to 100%, all judgments were made on a 6-point scale ranging from 1 to 6.

Method

Participants

One hundred fifty-eight participants were recruited from a university undergraduate subject pool and received course credit for participating. Seven participants were excluded due to either not completing the full experiment (three participants) or being given incorrect task instruction (four participants), resulting in 151 being included in all analyses. The sample size was chosen a priori because it is comparable to that used by Dougherty et al. (2005). Participants were randomly assigned to one of three conditions: RCJ ($n = 48$), JOL ($n = 52$), or no judgment ($n = 51$).

Materials

Four hundred fifty word pairs were created using the MRC Psycholinguistics Database (Wilson, 1988). Criteria were set so all items consisted of four-letter to eight-letter nouns with one or two syllables. The words were generated from the MRC database with ratings of familiarity >640, concreteness >410, and imageability >410. Of these word pairs, 56 were randomly selected to serve at the target word pairs for all participants.

Design and procedure

The design of this study was based on the design used by Dougherty et al. (2005). All participants completed the four stages of the study (study phase, prejudgment recall, metacognitive judgment, and final recall) 14 times within four blocks. All stimuli were presented with PsychoPy (Peirce, 2007, 2009). Before beginning the real experiment, all participants completed practice trials of the first three phases.

Study phase

Participants viewed sets of four to six word pairs, one pair at a time, for 5 seconds each, with the instructions to study the word pairs so that they would be able to complete cued recall later in the study. Although participants were instructed to study all word pairs, they were only tested on a subset of the word pairs referred to as the target word pairs. The target word pair always appeared as one of the first three pairs in a four-to-six-item set, with the remaining word pairs serving as distractors. The target appeared in varied positions within the first three pairs to keep participants from guessing the word pair on which they would be tested. Three word pairs were always presented after the target word pair to allow for a consistent delay between the initial study of the pair and prejudgment recall. All participants had the same 56 word pairs serve as the target word pairs. However, the distractor pairs were randomly selected for each participant.

Additionally, participants completed a dot-probe task during a random 20% of the word pair study trials. During the study phase of the task, four outlined square boxes surrounded the to-be-learned word pair, with the word pair centrally located among the four squares. On trials where a dot-probe was present, an asterisk (the probe) faded into one of the boxes 0–0.5 seconds after the word pair appeared. Participants were instructed to respond by pressing the space bar as soon as they detected the probe. The dot-probe task was included as a measure of the participant's attention while learning the word pairs. The specific hypothesis was that participants making JOLs might pay more attention to learning the word pair compared with the RCJ or no-judgment conditions and therefore perform worse or respond slower on the dot-probe task. This is a fairly straightforward application of dual-task methodology. Based on the assumption that if the three judgment conditions varied in terms of how much attention was allocated to encoding or rehearsal, then this should lead to a cost in performance on the dot-probe task for some conditions. On the other hand, if there were no differences in how much attention was allocated to encoding or rehearsal, then there should be no differences in dot-probe performance. Reaction time and accuracy for detecting the probes were recorded for each participant. The dot-probe task was included in the experiment to test a potential mechanism for why JOLs may enhance memory performance, assuming the effect replicates.

Prejudgment recall

After the presentation of the last distractor word pair, participants completed cued recall of the target word pair. This led to a roughly 15-second delay between encoding and prejudgment recall. Participants were presented with the first word of the pair (i.e., the cue) and asked to retrieve the second. Retrieval was self-paced, with participants typing their responses using the keyboard and hitting "Enter" to document they were done.

Metacognitive judgment

Immediately following prejudgment recall, participants made a metacognitive judgment on the word pair they just retrieved. Participants in the RCJ condition answered the following question: "How likely is it that you retrieved the correct word during the recall test?" Participants in the JOL condition answered, "How likely would you be to retrieve Word 2 again on a future recall test at the end of the study?" All JOLs and RCJs were made on a 6-point scale, with 1 = *very unlikely* and 6 = *very likely.* Participants in the no judgment condition were instructed to "select a random number between 1 and 6," in order to keep the condition as similar as possible to the two judgment conditions.

Final recall

After participants had completed the first three stages 14 times, they began final recall for that block. Participants again completed cued recall on all 14 target items and 14 randomly selected distractor items. Cued recall was again self-paced, with participants typing their responses and pressing "Enter" to indicate when they were done.

Results and discussion

Data and analysis code for this project are available at https://osf.io/mubzn/. Default Bayesian *t* tests and ANOVAs were conducted using the BayesFactor package (Morey & Rouder, 2015) in R (R Core Team, 2016). All analyses used the default prior on the effect size (rscale = 0.707 for *t* tests; rscale = 0.707^2 for ANOVA). Sensitivity analyses using different specifications of the prior distribution did not alter our conclusions in any way. These analyses are provided online as well. For our purposes, we are particularly interested in assessing the degree to which the data provide support for the alternative hypothesis that JOLs lead to better retrieval accuracy compared with the null hypothesis of no difference. To allow for the possibility that JOLs can be detrimental to retrieval, we used a two-sided test. Proportion data were transformed using the logit transformation before analysis. Of note, we used Kendall's tau as our measure of judgment accuracy. The use of tau departs somewhat from standard practice within the metacognitive literature, which has tended to use Goodman–Kruskal's gamma correlation coefficient to index metacognitive accuracy. Our choice of tau is based on two factors. First, though gamma and tau are identical when there are no

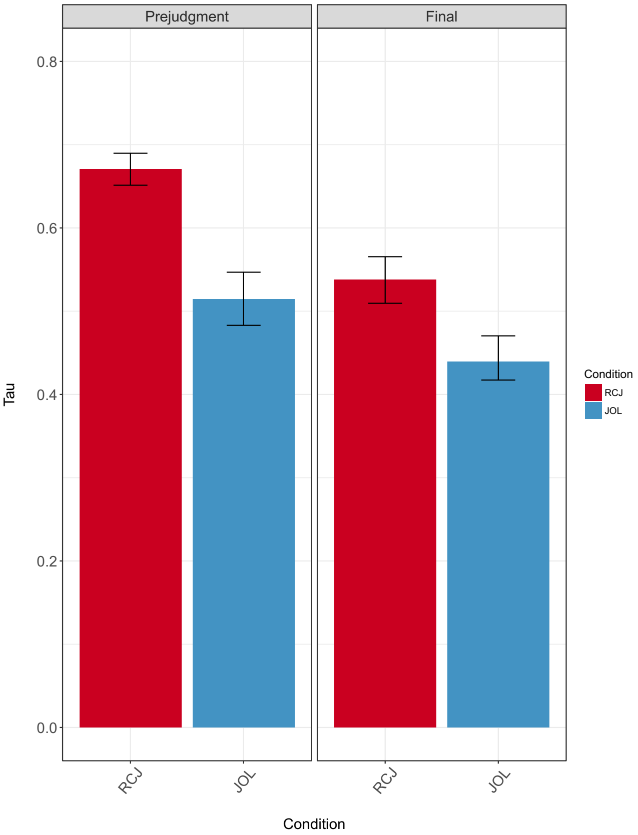

Fig. 1 Kendall’s tau between metacognitive judgment and recall accuracy for both prejudgment (left) and final recall (right)

ties, tau (but not gamma) has natural connections to both the area under the curve (AUC) for the receiver operator curve (ROC) and Pearson’s *r*, which can be estimated from tau via the transformation *r* = sin(pi * tau/2) (see Dougherty & Thomas, 2012; Tidwell, Dougherty, Thomas & Chrabaszcz, 2017 further generalizations). Second, Masson and Rotello (2009) showed that gamma was biased across a variety of conditions and therefore recommended against using gamma. Masson and Rotello instead suggested using estimates of the AUC. Because tau is closely related to the AUC, it is more appropriate than gamma.

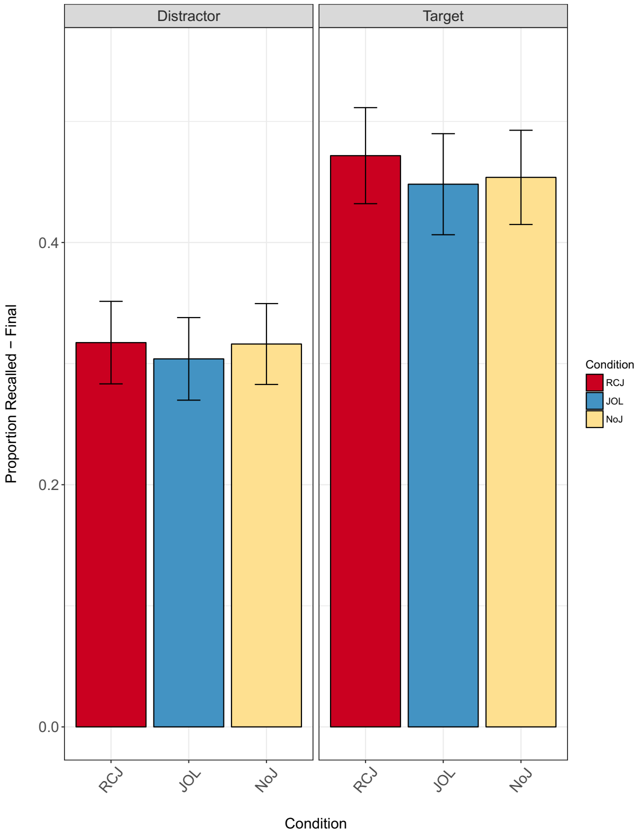

Fig. 3 Final recall accuracy for all conditions

Metacognitive accuracy

While the main purpose of this study was to examine the effect of judgment on later recall performance, we successfully replicated the central finding from Dougherty et al. (2005), which

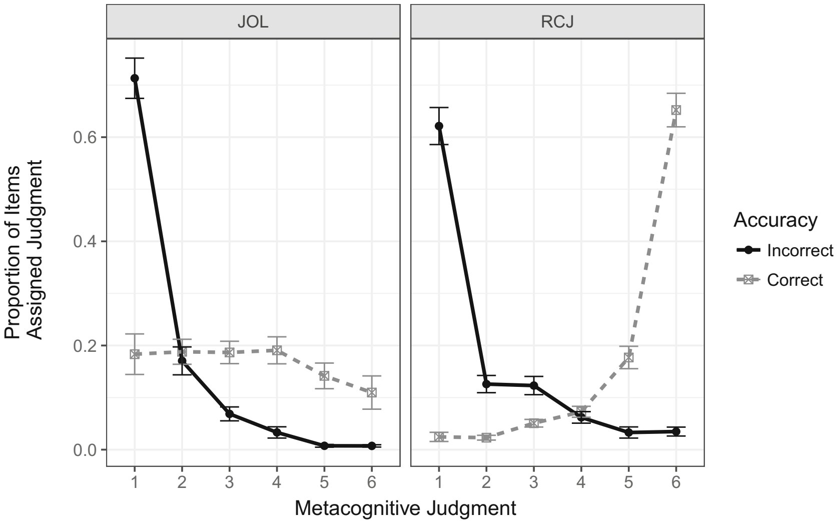

Fig. 2 Proportion of confidence distributions for each group conditionalized on correct or incorrect prejudgment recall

Table 2 Mean (SD) performance on the dot-probe task

	d'	Hit rate	Reaction time
RCJ	5.00 (0.10)	0.90 (0.02)	1081.61 (39.22)
JOL	4.71 (0.14)	0.85 (0.03)	1061.73 (31.48)
No judgment	4.94 (0.10)	0.90 (0.02)	1068.90 (51.62)

Note. RCJ = retrospective confidence judgments; JOL = judgments of learning

was that RCJ's were more predictive of future recall than were JOLs. Figure 1 plots the mean correlation between judgment and prejudgment recall accuracy (left) and judgment and final recall accuracy (right). In both cases, RCJs were more highly correlated than JOLs (BFs = 2247.44, 4.58). Dougherty et al. (2005) also noted that the conditional recall curves for JOLs differed substantially from those for participants in the RCJ condition. These curves were replicated as well (see Fig. 2 and Dougherty et al.'s Fig. 2). Additionally, the average correlations between the randomly selected number for the no-judgment condition and prejudgment and final recall were tau = 0.03 and 0.004, respectively, indicting these numbers were not meaningfully related to either prejudgment retrieval or projected retrieval success.

Memory-recall accuracy

The primary question in the present study was whether the act of making a metacognitive judgment, specifically a JOL, influenced final memory performance relative to making an RCJ or no judgment. To reiterate, Dougherty et al. (2005) reported that JOLs led to significantly better recall accuracy compared with a no-judgment condition, but that RCJs did not. Moreover, this result held for both final recall and conditional final recall, p(final recall | correct prejudgment recall). There was no significant effect for prejudgment recall, though the JOL condition performed numerically better than both RCJs and JOLs.

Figure 3 provides the mean final recall rates for the three conditions from our experiment, separated based on whether the items received prejudgment recall and judgment (targets) or no prejudgment recall and judgment (distractors). There were no differences between metacognitive judgment conditions concerning final recall accuracy, with the Bayes factor providing strong evidence for the null (BF_{10} = 0.05). A BF = 0.05 can be interpreted as a 20:1 odds (probability = 20/21= 0.95) in favor of the null hypothesis. There was, however, a main effect of trial type (target vs. distractors: BF_{10} = 13561.29), with targets being more likely to be retrieved than distractors. Importantly, there was no evidence for the Condition × Word Type interaction (BF_{10} = 0.07).

Additionally, there were no differences between conditions in prejudgment recall accuracy (BF_{10} = 0.12) and no differences in conditional retrieval (final recall accuracy given accurate prejudgment recall; BF_{10} = 0.08). Taken together these data indicate that there is little to no evidence to corroborate the hypothesis that JOLs lead to enhanced learning and later recall as suggested by Dougherty et al. (2005). Indeed, there was even a nominal retrieval advantage for the RCJ condition having higher recall, which again is contrary to Dougherty et al. (2005) and lends more credence to the conclusion that delayed JOLs do not promote better memory performance above and beyond retrieval practice.

Dot-probe performance

To ensure that the lack of findings in recall was not due to a tradeoff with secondary task performance, we analyzed the dot-probe data using d', hit-rate accuracy, and response reaction time (RT) as dependent variables (see Table 2). Response RT analyses were limited to correct trials only. Analyses of the dot-probe data using the AnovaBF function in R yielded Bayes factors that were either inconclusive or consistent with the null hypothesis: d' (BF_{10} = 0.33), hit rate (BF_{10} = 0.24), and response RT (BF_{10} = 0.07). Additionally, a comparison of dot-probe versus nonprobe trials illustrated that the presence of the dot-probe during learning had no impact on later memory performance (BF_{10} = 0.21).

Meta-analysis of four studies (*N* = 600)

Although Experiment 1 provides relatively strong evidence for the null hypothesis, it is worth noting that the sample sizes are relatively small. To address this issue, we reanalyzed the above data in combination with three additional studies. Two of these studies were designed to replicate the above study and examine restudy decisions simultaneously, and have been published elsewhere (Robey et al., in press) in the context of evaluating restudy decisions. The two experiments reported in Robey et al. (in press) used an experimental design that was identical to Experiment 1, with one difference: After making each metacognitive judgment, participants had to indicate whether they would choose to restudy

Table 3 Mean proportion (*SD*) of items retrieved correctly for all conditions aggregated across all four experiments

	Prejudgment recall	Final recall	Conditional final recall
RCJ (*N* = 223)	.57 (.24)	0.42 (0.26)	.67 (.24)
JOL (*N* = 223)	.54 (.26)	0.41 (0.30)	.68 (.25)
NoJ (*N* = 154)	.53 (.25)	0.39 (0.27)	.65 (.26)

Note. Includes data from Robey et al. (in press). RCJ = retrospective confidence judgments; JOL = judgments of learning; NoJ = no judgment the cue–target pair if given a chance (participants were not given the opportunity to restudy the items). Other than this simple modification, everything else was identical. The third study was identical to the experiments described in Robey et al. (in press) in that it included a restudy decision, but it differed from those studies in two respects: First, the experiment included only the JOL ($N = 65$) and RCJ ($N = 73$) conditions. Second, the instructions were slightly altered to mirror the instructions used in Dougherty et al. (2005). Specifically, participants in the JOL condition were asked, "How confident are you that in about 10 minutes you will be able to recall Word 2 when prompted with Word 1?" and participants in the RCJ condition were asked, "How confident are you that the reply you gave for this item is correct?" A summary of the restudy decision data for the third study are available at https://osf.io/q8eun/. This third study has not been published elsewhere.

The total sample size of the four studies was 600 ($N_{\mathrm{RCJ}} = 223$, $N_{\mathrm{JOL}} = 223$, $N_{\text{no-judge}} = 154$). An analysis comparing all three groups on final recall, prejudgment recall, and conditional recall (final recall accuracy conditional on prejudgment recall accuracy) yielded $\mathrm{BF}_{10} = 0.0184$, 0.115, and 0.2347, respectively. Limiting the analysis to just a comparison between RCJs and JOLs revealed similar results, with $\mathrm{BF}_{10} = 0.076$, 0.398, and 0.124 for final recall, prejudgment recall, and conditional recall, respectively (for descriptive statistics regarding the proportion retrieved correctly for all three conditions, see Table 3). Based on these analyses, we conclude that there is substantial evidence for the null hypothesis: Having participants make JOLs does not lead to enhancements in memory performance. Indeed, there was even a slight (though not meaningful) retrieval advantage for the RCJ condition in terms of prejudgment and final recall.

Conclusions

The data and analysis provided herein indicate that there is little evidence for the claim that metacognitive judgments enhance later retrieval, as proposed by Dougherty et al. (2005). It should be noted, however, that this conclusion is specific to the paradigm used in this study that requires all participants to engage in a recall attempt before making their metacognitive judgment. On the one hand, this methodology allows us to standardize the task requirements across the various conditions so that they can be compared. On the other hand, it does not allow us to address the very real possibility that metacognitive judgments may indeed affect learning in the absence of this requirement.

Many studies have shown reactive effects of immediate JOLs (JOLs that take place immediately following learning) on memory performance (Mitchum, Kelley, & Fox, 2016; Soderstrom, Clark, Halamish, & Bjork, 2015). In some of these studies, the act of making an immediate JOL during learning is found to improve later memory performance (Soderstrom, et al. 2015), whereas other studies show that making JOLs leads to poorer memory performance relative to a no-judgment control (Mitchum et al., 2016). Retrieval attempts are not expected to take place prior to immediate JOLs as there is no delay between learning and judgment, suggesting that benefits, when they occur, are likely due to the judgment itself, perhaps by enticing participants to engage in a form of elaboration during learning. Additionally, there is an abundance of literature illustrating that recall practice leads to measurable improvements in later memory performance (the testing effect; Karpicke & Roediger, 2007). If having participants make a metacognitive judgment increases the likelihood that they engage in recall to assess their memory, then it is also likely to lead to improvements in learning of that item (see Tauber et al., in press).

Our study was designed to assess if different metacognitive judgments improved learning over and above prejudgment recall, as suggested by Dougherty et al. (2005). The simple answer to this question is no. More broadly, this study highlights the value of replication for verifying (or not) behavioral phenomena. The original finding observed in Dougherty et al. (2005) persisted seemingly unchallenged for over 12 years, yet in our view it appears to be nothing more than a Type I error.

Author note Michael R. Dougherty, Alison M. Robey, and Daniel Buttaccio, Department of Psychology, University of Maryland; College Park, MD 20742. This study was supported in part by NSF Grant 146831 awarded to M.R.D. Data and analysis code for this article are provided at https://osf.io/mubzn/

References

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

Dougherty, M. R., Scheck, P., Nelson, T. O., & Narens, L. (2005). Using the past to predict the future. *Memory & Cognition, 33*(6), 1096–1115.

Dougherty, M. R., & Thomas, R. P. (2012). Robust decision making in a nonlinear world. *Psychological Review*, *119*(2), 321–344. doi: https://doi.org/10.1037/a0027039

Jeffreys, H. (1961). Theory of probability. Oxford, UK: Oxford University Press.

Karpicke, J. D., & Roediger, H. L. (2007). Expanding retrieval practice promotes short-term retention, but equally spaced retrieval enhances long-term retention. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 33,* 704–719.

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

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

Masson, M. E. J., & Rotello, C.R. (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, 509–527.

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

Morey, R. D., & Rouder, J. N. (2015). BayesFactor: Computation of Bayes factors for common designs (R Package Version 0.9.12-2) [Computer software]. Retrieved from https://CRAN.R-project.org/package=BayesFactor

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

Peirce, J. W. (2007). PsychoPy—Psychophysics software in Python. *Journal of Neuroscience Methods*, *162*(1), 8–13.

Peirce, J. W. (2009). Generating stimuli for neuroscience using PsychoPy. *Frontiers in Neuroinformatics,* 2–10. doi:https://doi.org/10.3389/neuro.11.010.2008

R Core Team. (2016). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieve from https://www.R-project.org/

Rhodes, M. G. (2016). Judgments of learning: Methods, data, and theory. In J. Dunlosky & S. K. Tauber (Eds.), The Oxford handbook of metamemory *(*pp. 63–80). New York, NY: Oxford University Press.

Robey, A., Buttaccio, D., & Dougherty, M. (in press). Making retrospective confidence judgments improves learners' ability to decide what 'not' to study. *Psychological Science*.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian *t* test for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16,* 225–237. doi: https://doi.org/10.3758/PBR.16.2.225.

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

Sommer, W., Heinz, A., Leuthold, H., Matt, J., & Schweinberger, S. R. (1995). Metamemory, distinctiveness, and event related potentials in recognition memory for faces. *Memory & Cognition, 23,* 1–11.

Spellman, B. A., & Bjork, R. A. (1992). People's judgments of learning are extremely accurate at predicting subsequent recall when retrieval practice mediates both tasks. *Psychological Science, 3*, 315–316.

Tauber, S. K., & Rhodes, M. G. (2012). Measuring memory monitoring with judgments of retention interval (JOR). *Quarterly Journal of Experimental Psychology, 65,* 1376–1396.

Tauber, S. K., Witherby, A. E., Dunlosky, J., Rawson, K. A., Putman, A. L., & Roediger, H. L., III (in press). Does covert retrieval benefit learning of key-term definitions? *Journal of Applied Research in Memory & Cognition.*

Tidwell, J. W., Dougherty, M. R., Chrabaszcz, J., & Thomas, R. P. (2017). Order constrained linear optimization. *British Journal of Mathematical and Statistical Psychology.* Advance online publication. doi:https://doi.org/10.1111/bmsp.12090

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E. J. (2011). Statistical evidence in experimental psychology. *Perspectives on Psychological Science, 6*(3), 291–298.

Wilson, M. D. (1988). The MRC psycholinguistic database: Machine readable dictionary, Version 2. *Behavioural Research Methods, Instruments, and Computers, 20*, 6–11.