Research Report

Psychological Science
2017, Vol. 28(11) 1683–1693

DOI: 10.1177/0956797617718800
www.psychologicalscience.org/PS
SAGE

Making Retrospective Confidence Judgments Improves Learners' Ability to Decide What *Not* to Study

Alison M. Robey, Michael R. Dougherty, and Daniel R. Buttaccio
Department of Psychology, University of Maryland

Abstract

Predictions about future retrieval success, known as judgments of learning (JOLs), are often viewed as important for effective control over learning. However, much less is known about how retrospective confidence judgments (RCJs), evaluations of past retrieval success, may affect control over learning. We compared participants' ability to identify items that would benefit from additional study after making either a JOL or an RCJ. Participants completed a cued-recall task in which they made a metacognitive judgment after an initial recall attempt and before making a restudy decision. Participants who made RCJs prior to their restudy decisions were more accurate at identifying items in need of being restudied, relative to participants who made JOLs. The results indicate that having participants assess their confidence in past retrieval success can nudge them toward better utilizing of valid information when deciding which items are in need of further study.

Keywords

study decisions, judgments of learning, metacognition, retrospective confidence judgments, open data, open materials, preregistered

Received 6/23/16; Revision accepted 6/12/17

Judgments of learning (JOLs) are on-line assessments of how likely it is that a studied item will be successfully recalled in the future (see Dunlosky & Metcalfe, 2009, for a review). A common assumption is that JOLs can be used to help effectively guide study behavior and improve learning. For example, if a learner can accurately discriminate between items that have been sufficiently learned versus those that have not, then presumably this information can be used to determine which items need to be restudied (e.g., Metcalfe & Kornell, 2005; Son & Metcalfe, 2005; see Metcalfe, 2009, for a review). The use of JOLs to inform restudy decisions both has face validity and is consistent with the observation that JOLs correlate with future retrieval success, especially when made after a short delay (the delayed-JOL effect; Nelson & Dunlosky, 1991). Interestingly, however, past work shows that judgments about past retrieval success, or retrospective confidence judgments (RCJs), are better predictors of memory recall than JOLs (e.g., Dougherty, Scheck, Nelson, & Narens, 2005; Hines, Touron, & Hertzog, 2009; Ryals, Rogers, Gross, Polnaszek, & Voss, 2016; Wattier & Collins, 2011). Thus, to the extent that RCJs are more predictive of recall than JOLs, RCJs should also be a better guide for deciding which items to restudy.

In the prototypical delayed-JOL paradigm, *prejudgment recall and monitoring* (PRAM; Nelson, Narens, & Dunlosky, 2004), learners study a set of paired associates and then, after a short delay, complete a prejudgment recall task in which they attempt to recall the target word when prompted with the cue. After each recall attempt, learners assess their confidence that they will be able to retrieve the target on a future memory test. We augmented this paradigm in two ways. First, we manipulated whether participants made delayed

Corresponding Author:
Alison M. Robey, University of Maryland at College Park, Department of Psychology, Biology-Psychology Building, College Park, MD 20742-5031
E-mail: arobey@umd.edu

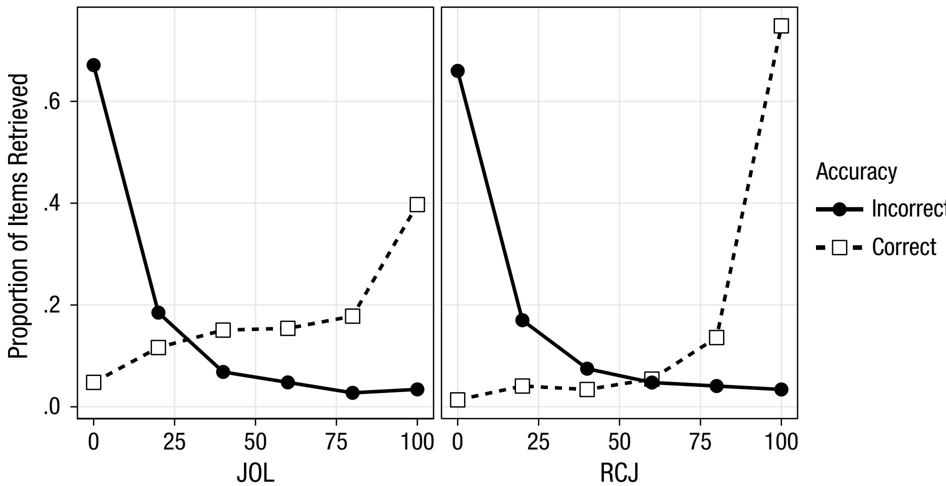

Fig. 1. Example distributions showing the proportion of items correctly and incorrectly retrieved as a function of participants' metacognitive judgment, separately for participants who made judgments of learning (JOLs) and retrospective confidence judgments (RCJs). Data were estimated from Dougherty, Scheck, Nelson, and Narens (2005, Fig. 2).

JOLs or RCJs after prejudgment recall (Dougherty et al., 2005). In contrast to the JOL condition, in the RCJ condition, participants rated their confidence in the accuracy of the target item they had just retrieved. Second, after making their confidence judgment, participants made a binary decision: If given the opportunity, would they choose to restudy the cue-target pair? Assuming that participants use their confidence judgments as the basis of their restudy decisions, participants making RCJs should be better at discriminating items that require further study, compared with participants making JOLs. To assess how RCJs and JOLs may change future study behavior, it is also necessary to compare them with pure study decisions. Therefore, we also included a no-judgment condition in which participants made only the restudy decision.

A critical question regarding the differences between RCJs and JOLs concerns why RCJs are more predictive of future recall than JOLs. One key difference between RCJs and JOLs lies in the conditional distributions of confidence judgments. Figure 1 plots typical distributions of RCJs and JOLs conditional on prejudgment recall accuracy (see Dougherty et al., 2005). When prejudgment recall is incorrect, the distribution of confidences for RCJs and JOLs are nearly identical, both showing high positive skew. However, when prejudgment recall is correct, the confidence distributions differ considerably, with RCJs showing negative skew and JOLs showing an almost uniform distribution. This pattern explains why RCJs are more predictive of recall: RCJs are more discriminant of correctly recalled items. Thus, we hypothesized that RCJs would lead to more accurate restudy decisions, but only for items for which prejudgment recall is correct. Specifically, we expected that participants in the RCJ condition would be less likely to select to restudy items that had been correctly retrieved at prejudgment recall relative to those in the JOL condition, because participants making RCJs are more likely to assign high confidence to correctly recalled items than to incorrectly recalled items. We assumed that both types of judgments would improve restudy decisions relative to the no-judgment condition.

Method

Participants

Three hundred forty-six participants were recruited from the University of Maryland subject pool and received course credit for participating. Data were collected for both an initial study and an exact replication. Sample size for both studies was determined a priori to be 50 participants per condition, a sample size comparable with that of a similarly designed experiment (Dougherty et al., 2005). For the initial study, data collection was stopped when at least 50 participants had completed the study in each of the three randomly assigned conditions. For the replication study, data were collected through the end of the semester, which resulted in slightly more than 50 participants per condition.

One hundred seventy-one undergraduates participated in the initial study. Of those who participated, 151 completed the study: 50 in the RCJ condition, 51 in the JOL condition, and 50 in the no-judgment condition. Data collection for the exact replication began

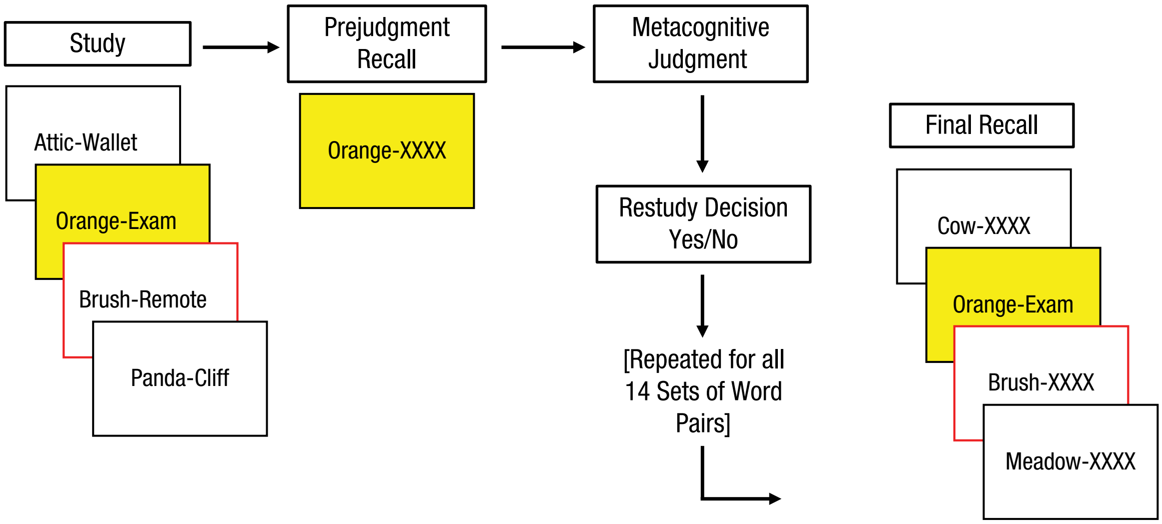

Fig. 2. A visual representation of the procedure in one block. Participants first completed a study phase, in which they saw a series of word pairs. One of these pairs was the target (highlighted here in yellow), and the rest were distractors. Then, participants completed prejudgment recall, in which they were shown the first word of the target pair and asked to supply the second. In the following phase, participants made a metacognitive judgment—either a judgment of learning or a retrospective confidence judgment—indicating how confident they were in their response (or selected a random number between 1 and 6). After this, they were asked whether or not they would restudy this word pair if given the opportunity. In the last phase, participants were asked to recall the second word from all target word pairs and from a random selection of nontarget pairs from the study phase (highlighted here with a red border).

immediately after completion of the initial study. No modifications were made between the two studies. One hundred seventy-five undergraduates participated in the replication, 160 of whom completed the study: 52 in the RCJ condition, 55 in the JOL condition, and 53 in the no-judgment condition. The participants who did not complete the studies either experienced computer errors or chose to stop of their own volition.

Materials

The MRC Psycholinguistics Database (Wilson, 1988) was used to create 450 word pairs. Words were limited to four- to eight-letter, one- or two-syllable nouns with familiarity, concreteness, and imageability ratings above 640, 410, and 410, respectively. From the words that fit this criteria, 900 were selected, with the constraint that the first three letters of all words were different. The words were randomly matched into pairs, and 56 of the pairs were randomly selected to serve as the target word pairs. The remaining word pairs went into a pool, where they could be selected for the practice and distractor trials.

Design and procedure

The study had a between-subjects design with three conditions: RCJ, JOL, and no judgment. The design of this study was based on the design used by Dougherty et al. (2005), which utilized PRAM methodology (Nelson et al., 2004). This type of design requires participants to engage in recall of target words immediately preceding their metacognitive confidence judgments, in order to eliminate the effects of covert retrieval attempts with some types of judgments. The study consisted of five phases that were repeated 14 times each in four blocks. All stimuli were programmed and presented with PsychoPy (Peirce, 2007, 2009). Before the experiment began, participants completed practice trials of the first four phases, 4 times each. The design of the study can be seen in Figure 2.

Study phase. In the study phase, four to six word pairs were presented on the screen one at a time, and participants were instructed to study the pairs so that they would be able to recall the second word when presented with the first. Participants were instructed to study all word pairs; however, they were tested only on the target word pairs appearing as one of the first three word pairs. The remaining word pairs served as distractors. Presenting a variable number of distractor pairs before the target word pair prevented participants from identifying which word pair they would be tested on. The number of pretarget distractor pairs was randomly determined for each set. Three distractors were always presented after the target word pair, which allowed for a consistent delay between encoding and retrieval of the target word pairs. Because the number of distractors before the target word pair was randomly determined for each set, the total number of studied word pairs could vary from 224 to 336. All participants, however, studied the same 56 target word pairs regardless of the number of distractors. All word pairs were presented for 5 s each. (For purposes described in Section A of the Supplemental Material available online, a dot-probe task was also included on 20% of the study trials).

Prejudgment recall. Immediately following the presentation of the last distractor item in each set, participants were presented with the first word of the target pair and instructed to type the second word. All recall was self-paced, and participants were required to respond to all prompts. The recalled word was counted as correct if the first three letters typed by the participant matched the first three letters of the target word (cf. Dougherty et al., 2005). As described in the Materials section, all words were uniquely identifiable by their first three letters.

Metacognitive judgment. Immediately following prejudgment recall, participants made a metacognitive judgment regarding their confidence in their response (1 = low confidence, 6 = high confidence). There were three metacognitive conditions. Participants in the RCJ condition responded to the following question: "How likely is it that you retrieved the correct word during the recall test?" Participants in the JOL condition made a delayed JOL, responding to the following question: "How likely would you be to retrieve the word again on a future recall test at the end of the study?" Participants in the no-judgment condition did not make a metacognitive judgment but were told to select a random number between 1 and 6 to keep the task as similar as possible to the other conditions. All participants responded by typing their responses on the keyboard.

Restudy judgment. After making their metacognitive judgment, participants made their restudy judgment. Participants were asked "If given the opportunity, would you choose to restudy this item?" Participants made a binary decision by selecting either 1 ("yes restudy") or 0 ("no restudy") on the keyboard. Although participants were asked whether they would choose to restudy a word pair during the restudy judgment, no restudy opportunities were provided.

Final recall. After participants completed the study, prejudgment recall, metacognitive judgments, and restudy judgments for all 14 sets in a block, they completed final recall for all target word pairs and 14 randomly selected distractor word pairs. As in prejudgment recall, participants completed a self-paced cued-recall task by typing in their responses. Again, the recalled word was counted as correct if the first three letters typed by the participant matched the first three letters of the target word. After completing final recall, participants completed two questionnaires measuring their strategy use and mind wandering (see Section B in the Supplemental Material for further details).

Results

Data from the initial study and replication study were analyzed both separately and combined using the default Bayes factor (BF) *t* test (Rouder, Speckman, Sun, Morey, & Iverson, 2009), with proportion data transformed using the logit transformation. The BF indexes the degree to which the data support any particular model (including the null model). Subscripts indicate whether they express evidence for the alternative hypothesis (BF_{10}) or the null hypothesis (BF_{01}). Metacognitive accuracy was assessed at the individual subject level by computing a rank order correlation between a participant's recall accuracy and his or her corresponding confidence judgments. Kendall's τ was also used as an index of the accuracy of restudy decisions. We chose Kendall's τ-b as the correlation coefficient, as opposed to the Goodman-Kruskal γ correlation on the basis of the findings of Masson and Rotello (2009) illustrating that γ is systematically biased. Nevertheless, it is worth pointing out that all conclusions based on the use of τ are consistent with those based on γ. Although we focus our exposition on the results of the combined analyses, results from the separate analyses are included in the tables.

There were no meaningful differences in the overall percentage of correct responses across the three conditions for either prejudgment recall (RCJ = 56.4%, JOL = 51.7%, no judgment = 50.4%) or final recall (RCJ = 41.2%, JOL = 41.0%, no judgment = 35.7%), with strong support for the null hypothesis (BF_{10}s = 0.16, 0.10). However, there was overwhelming evidence that RCJs were more correlated than were JOLs with prejudgment recall (BF_{10} = 6299017.0; see Table 1), a finding that is consistent with prior studies and supports the assertion that RCJs should provide a more accurate basis for restudy decisions (Dougherty et al., 2005; Hines et al., 2009; Ryals et al., 2016; Wattier & Collins, 2011). There was also modest support for the correlation between confidence judgments and final recall being higher for RCJs than for JOLs (BF_{10} = 4.42). Although both RCJs and delayed JOLs were accurate predictors of future recall, RCJs were more accurate.

Figure 3 plots the conditional confidence distributions. As in prior research, the distributions for both RCJs and JOLs for incorrect prejudgment recall were highly positively skewed. In contrast, the distribution for correctly retrieved items differed; RCJs showed a

Table 1. Comparison of the Mean Correlations (Kendall's τ-*b*s) Between Metacognitive Judgments and Cued Recall

Data set and phase	RCJ	JOL	BF_{10}	Median effect size estimated from the posterior distribution
Combined data				
Prejudgment recall	.70	.57	6,299,017.00	0.86 [0.57, 1.14]
Final recall	.55	.49	4.42	0.35 [0.09, 0.61]
Initial study				
Prejudgment recall	.69	.58	47.35	0.66 [0.26, 1.05]
Final recall	.54	.49	0.47	0.24 [–0.15, 0.62]
Replication				
Prejudgment recall	.69	.55	37,328.14	1.0 [0.60, 1.41]
Final recall	.57	.49	3.21	0.45 [0.08, 0.81]

Note: For effect sizes, 95% credible intervals are given in brackets. RCJ = retrospective confidence judgment; JOL = judgment of learning; BF_{10} = Bayes factor referring to the amount of evidence in favor of the hypothesis that there is a difference between the RCJ and JOL conditions.

negative skew, and JOLs showed a uniform distribution. This finding is important because we assume that the same information used to make confidence judgments also provided the basis for restudy decisions. However, did type of confidence judgment (RCJ vs. JOL) affect restudy decisions as predicted?

As shown in Table 2, there is convincing evidence that the correlation between restudy decisions and final recall was greater for the RCJ condition than the JOL condition (BF_{10} = 9,668.76). Restudy decisions were also better correlated with prejudgment recall in the RCJ condition (mean $\tau = -0.33$) compared with the JOL condition (mean $\tau = 0.03$). These patterns suggest that participants in the RCJ condition should be better able to discriminate between those items that would benefit from restudy and those items that would not. This was indeed the case, as shown in Table 3. Participants in the JOL condition chose to restudy more items (55.4%) that had been retrieved correctly at prejudgment recall, compared with participants in the RCJ condition (29.0%). This pattern also held for final recall: Participants in the JOL condition chose to restudy more items (61.3%) than participants in the RCJ condition (26.1%) that were ultimately correctly retrieved at final recall. For items incorrectly retrieved at prejudgment or final recall, the difference in the percentage of restudy decisions between the RCJ and JOL conditions was not meaningful. Note that RCJs were more strongly correlated with restudy decisions ($\tau = -0.36$) than were JOLs ($\tau = 0.04$), which suggests that participants in the RCJ condition

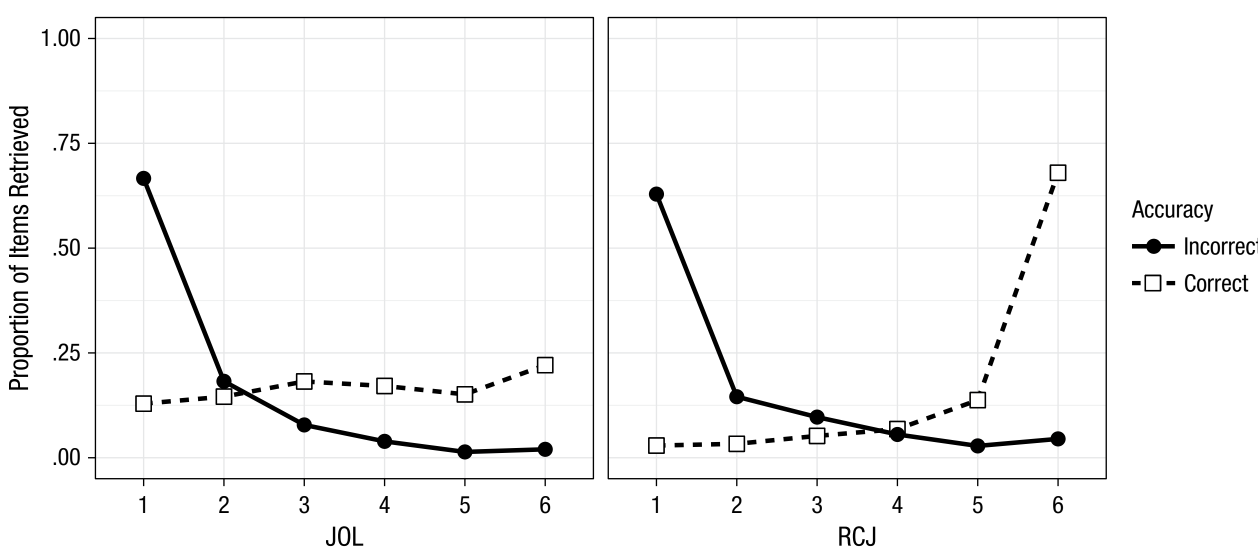

Fig. 3. Mean proportion of items correctly and incorrectly retrieved during prejudgment recall as a function of participants' metacognitive judgment, separately for participants who made judgments of learning (JOLs) and retrospective confidence judgments (RCJs).

Table 2. Comparison of the Mean Correlations (Kendall's τ-*b*s) Between Restudy Decisions and Final Recall, Prejudgment Recall, and Metacognitive Judgment

Data set and phase	RCJ	JOL	BF_{10}	Median effect size estimated from the posterior distribution
Combined data				
Final recall	–.29	.04	9,668.76	0.74 [0.44, 1.04]
Prejudgment recall	–.33	.03	5,875.38	0.70 [0.38, 1.01]
Metacognitive judgment	–.36	.04	10,763.24	0.72 [0.41, 1.03]
Initial study				
Final recall	–.27	–.01	5.45	0.53 [0.10, 0.96]
Prejudgment recall	–.29	.001	4.51	0.52 [0.09, 0.95]
Metacognitive judgment	–.33	.02	7.11	0.58 [0.16, 0.99]
Replication				
Final recall	–.31	.07	363.09	0.80 [0.35, 1.24]
Prejudgment recall	–.37	.06	324.15	0.81 [0.40, 1.21]
Metacognitive judgment	–.39	.06	352.71	0.81 [0.37, 1.24]

Note: For effect sizes, 95% credible intervals are given in brackets. RCJ = retrospective confidence judgment; JOL = judgment of learning; BF_{10} = Bayes factor referring to the amount of evidence in favor of the hypothesis that there is a difference between the RCJ and JOL conditions.

based their restudy decisions on the same information used to make confidence judgments, whereas participants in the JOL condition did not. We return to this issue in the Exploratory Analyses section.

The next question of interest is whether either of the judgment conditions differed from the no-judgment condition, which may reflect the restudy decision behaviors participants made naturally. Table 3 provides the mean percentage of items that participants chose to restudy, separately for items that were correctly recalled and not correctly recalled in all three groups. Ideally, participants should choose to restudy only those items that were incorrectly retrieved at prejudgment recall or ultimately were incorrectly retrieved at final recall. Though participants in both the RCJ and no-judgment conditions generally chose to restudy fewer of the items they correctly retrieved, compared with participants in the JOL condition, the data were inconclusive in regard to other comparisons with the no-judgment condition. Note, however, that participants in the JOL condition chose to restudy more items (M = 29.77) than participants in the RCJ (M = 22.37, BF_{10} = 11.93) and no-judgment (M = 22.63, BF_{10} = 6.59) conditions.

Table 4 provides the mean percentage of restudied and not restudied items that were incorrectly recalled during final recall. If restudy decisions accurately capture which items would benefit from restudy, then participants should have more retrieval errors for items they chose to restudy. Although there was no difference between the RCJ (77.5%) and no-judgment (71.4%) conditions, participants in both of these conditions had more difficulty recalling items they had chosen to restudy, compared with participants in the JOL condition (57.3%). On the other hand, among items that were not selected for restudy, incorrect recall was lower in the RCJ condition (49.1%) compared with both the JOL (62.1%) and no-judgment (61.2%) conditions. There was no difference in incorrect recall between the JOL and no-judgment conditions when items were not selected for restudy.

Exploratory Analyses

Given the striking differences between participants who made JOLs and those who made RCJs, we ran a number of follow-up exploratory analyses to identify potential explanations for these large discrepancies. These analyses were not part of our original analysis plan. The analysis approach involved policy-capturing methodology. This methodology is used in judgment and decision-making research to identify how people utilize cues in decision-making contexts, as conceptualized within Brunswik's (1956) lens model (Dougherty & Thomas, 2012; Hammond, McClelland, & Mumpower, 1980; Hammond, Rohrbaugh, Mumpower, & Adelman, 1977; Karelaia & Hogarth, 2008). For our purposes, we used a variant of the maximum rank correlation estimator (Han, 1987) developed by Dougherty and Thomas (2012; see also Tidwell, Dougherty, Chrabaszcz, & Thomas, 2017) called general monotone modeling (GeMM) to estimate regression parameters at the individual subject level. GeMM is a semiparametric regression procedure in which regression parameters are estimated by maximizing Kendall's τ. The approach is nonparametric both in terms of the assumed functional form and in terms of the error distribution, and it has been shown to be a general regression procedure that

Table 3. Comparison of the Mean Percentage of Restudy Decisions for Items That Were Correctly and Incorrectly Recalled

Phase and recall accuracy	RCJ (%)	JOL (%)	No judgment (%)	RCJ vs. JOL		RCJ vs. no judgment		JOL vs. no judgment	
				BF_{10}	Median effect size	BF_{10}	Median effect size	BF_{10}	Median effect size
Combined data									
Prejudgment recall									
Correct	29.0	55.4	38.1	71,551.10	–0.74 [–0.45, –1.02]	0.80	–0.24 [–0.51, 0.03]	23.62	–0.44 [–0.16, –0.72]
Incorrect	58.7	51.8	45.1	0.20	0.09 [–0.17, 0.35]	2.13	0.31 [0.03, 0.59]	0.44	0.21 [–0.07, 0.48]
Final recall									
Correct	26.6	57.5	36.4	2,568,976.00	–0.84 [–0.54, –1.13]	0.68	–0.24 [–0.52, 0.04]	334.48	–0.50 [–0.02, –0.98]
Incorrect	52.1	51.2	43.3	0.18	–0.10 [–0.35, 0.16]	0.62	0.23 [–0.03, 0.48]	1.66	0.31 [0.04, 0.58]
Initial study									
Prejudgment recall									
Correct	32.1	56.8	38.1	36.52	–0.64 [–0.27, –1.0]	0.27	–0.13 [–0.50, 0.24]	3.68	–0.47 [–0.08, – 0.85]
Incorrect	58.6	56.3	46.8	0.21	0.02 [–0.34, 0.38]	0.59	0.27 [–0.10, 0.64]	0.46	0.48 [–0.14, 0.62]
Final recall									
Correct	30.1	57.7	35.0	62.78	–0.67 [–0.24, –1.09]	0.23	–0.08 [–0.46, 0.30]	16.80	–0.61 [–0.21, –1.0]
Incorrect	53.9	55.9	46.1	0.27	–0.25 [–0.49, 0.24]	0.45	0.22 [–0.18, 0.62]	1.09	0.38 [–0.01, 0.77]
Replication study									
Prejudgment recall									
Correct	26.0	54.0	38.0	362.39	–0.76 [–0.38, –1.14]	1.06	–0.34 [–0.71, 0.03]	1.46	–0.38 [–0.01, –0.74]
Incorrect	58.8	47.7	43.4	0.29	0.18 [–0.18, 0.53]	0.91	0.32 [–0.05, 0.69]	0.28	0.15 [–0.21, 0.51]
Final recall									
Correct	23.4	57.3	37.8	11,168.63	–0.93 [–0.53, –1.33]	1.57	0.40 [0.02, 0.77]	3.92	–0.48 [–0.09, –0.86]
Incorrect	50.4	48.0	40.8	0.29	–0.06 [–0.41, 0.30]	0.36	0.22 [–0.12, 0.56]	0.43	0.24 [–0.12, 0.59]

Note: Median effect sizes were estimated from the posterior distribution; 95% credible intervals for these effect sizes are given in brackets. RCJ = retrospective confidence judgment; JOL = judgment of learning; BF_{10} = Bayes factor referring to the amount of evidence in favor of the hypothesis that there is a difference between the RCJ and JOL conditions.

Table 4. Comparison of the Mean Percentage of Restudied and Not Restudied Items That Were Incorrectly Recalled During Final Recall

Data set and restudy decision	RCJ (%)	JOL (%)	No judgment (%)	RCJ vs. JOL: BF_{10}	RCJ vs. JOL: Median effect size	RCJ vs. no judgment: BF_{10}	RCJ vs. no judgment: Median effect size	JOL vs. no judgment: BF_{10}	JOL vs. no judgment: Median effect size
Combined data									
Restudy	77.5	57.3	71.4	2,327.57	−0.63 [0.35, 0.91]	0.37	0.20 [−0.08, 0.48]	14.72	−0.43 [−0.15, −0.71]
No restudy	49.1	62.1	61.2	7.25	−0.40 [−0.11, −0.68]	7.38	−0.39 [−0.11, −0.66]	0.15	0.03 [−0.26, 0.32]
Initial study									
Restudy	76.3	59.4	74.2	4.02	0.40 [0.07, 0.87]	0.23	0.06 [−0.32, 0.44]	1.58	−0.39 [−0.78, 0.01]
No restudy	48.5	61.0	62.9	1.70	−0.41 [−0.02, −0.80]	1.81	−0.40 [−0.03, −0.77]	0.23	0.06 [−0.32, 0.43]
Replication study									
Restudy	78.7	55.4	68.8	160.95	0.74 [0.31, 1.16]	0.66	0.28 [−0.12, 0.67]	2.04	−0.43 [−0.07, −0.78]
No restudy	50.0	61.4	59.5	1.04	−0.36 [−0.74, 0.03]	0.97	−0.33 [−0.69, 0.04]	0.21	0.04 [−0.33, 0.40]

Note: Median effect sizes were estimated from the posterior distribution; 95% credible intervals for these effect sizes are given in brackets. RCJ = retrospective confidence judgment; JOL = judgment of learning; BF_{10} = Bayes factor referring to the amount of evidence in favor of the hypothesis that there is a difference between the RCJ and JOL conditions.

encompasses a variety of parametric procedures as special cases, including ordinary least squares, the Box-Cox model, and logistic regression (see Han, 1987; Tidwell et al., 2017). In our application of GeMM to policy capturing, we estimated the degree to which participants' confidence judgments and restudy decisions were predicted by prejudgment-recall accuracy and latency. Both of these variables are valid indices of future recall success, as one would expect that the probability of accuracy at final recall should increase as a function of accuracy at prejudgment recall and decrease as a function of the latency of prejudgment recall (i.e., more slowly recalled items should have a lower probability of being recalled later).

Figure 4a plots the mean standardized regression parameters and 95% confidence intervals when

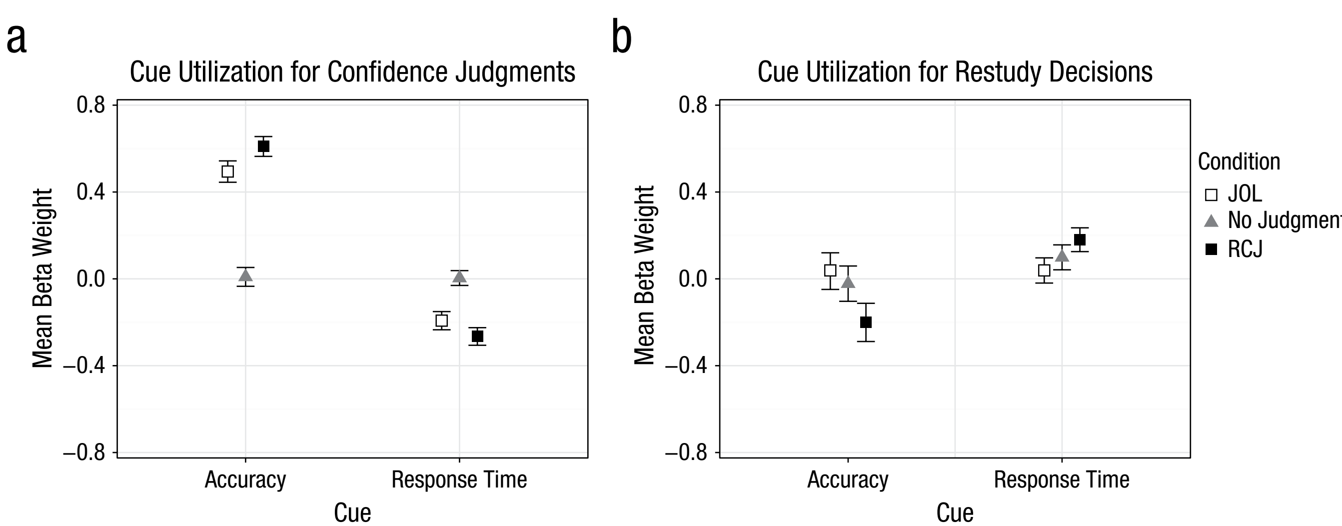

Fig. 4. Results of the exploratory analysis: mean beta weight for accuracy and response time is shown for each condition, separately for analyses predicting (a) confidence judgments and (b) restudy decisions. Depending on condition, participants made retrospective confidence judgments (RCJs), judgments of learning (JOLs), or no judgments. Error bars show 95% confidence intervals.

prejudgment-recall accuracy and response time (RT) were used to predict confidence judgments. As can be seen, the regression parameters for both the JOL and RCJ conditions differ from zero, which suggests that participants in both conditions used these two cues when making their confidence judgments. Predictably, regression parameters for participants in the no-judgment condition did not differ from zero. Figure 4b plots the mean standardized regression parameters (with 95% confidence intervals) for prejudgment-recall accuracy and RT when predicting restudy decisions. Surprisingly, participants in the JOL condition appeared to completely disregard prejudgment-recall accuracy and RT when making restudy decisions, as evidenced by the fact that the regression parameters did not differ from zero. This contrasts with the RCJ condition, in which both regression parameters differed significantly from zero: Participants in the RCJ condition were less likely to select items for restudy that were correctly recalled and more likely to select items that were recalled more slowly. Interestingly, participants in the no-judgment condition appeared to utilize prejudgment-recall RT (but not prejudgment-recall accuracy) when making their restudy decisions.

Discussion

We tested whether having participants make RCJs would improve specific item-restudy decisions relative to having participants make delayed JOLs. We found that by orienting participants to focus on past retrieval success, participants in the RCJ condition were better able to discriminate among items that would benefit from additional study and items that were already sufficiently learned to be retrieved at final test. Participants in the JOL condition decided to restudy 26% more items that had been recalled correctly at prejudgment recall, compared with participants in the RCJ condition, whereas there were no between-conditions differences in decisions related to items recalled incorrectly. This pattern was even more pronounced for final recall, in which participants in the JOL condition chose to restudy 31% more of the items that they ultimately correctly retrieved at final study. This suggests that having participants make RCJs improves their ability to identify what not to restudy, compared with when participants make JOLs. Contrary to expectations, however, neither judgment condition was consistently better than the no-judgment condition: Restudy decisions made with no metacognitive judgments were comparable with those in the RCJ condition and better than those in the JOL condition.

Analysis of the predictors of confidence judgments and restudy decisions suggests that having participants make different types of judgments (JOLs vs. RCJs) affects the information that is used when deciding which items to restudy. Although both JOLs and RCJs appeared to be based on accuracy and RT during prejudgment recall, the relation was stronger for participants in the RCJ than in the JOL condition. The use of this information when making metacognitive judgments is consistent with previous research showing that when other information is not available, memory for past test and retrieval fluency are heavily relied on when making delayed JOLs (Finn & Metcalfe, 2008; Koriat & Ma'ayan, 2005). Interestingly, it appeared as though participants in the JOL condition completely disregarded both prejudgment recall and RT when making restudy decisions. This may seem surprising given the vast amount of past research suggesting that JOLs can be used to effectively guide study behavior; however, when empirically tested, improvements in metacognitive judgments do not always lead to improvements in effective restudy decisions (Kimball, Smith, & Muntean, 2012). When retrieval accuracy was not directly assessed, as in the no-judgment condition, participants appeared to make restudy decisions on the basis of retrieval fluency of the prejudgment recall. Only participants in the RCJ condition continued to incorporate both prejudgment recall and prejudgment recall RT into their restudy decisions.

We suggest that having participants make RCJs focuses their attention on the current status of an item in memory (its current retrievability) and increases the use of predictive cues in formulating restudy decisions. Our findings challenge conventional wisdom that JOLs are well suited for enhancing metacognitive control over learning and suggest that having participants instead focus on past retrieval success (using RCJs) can nudge them toward more optimal restudy decisions. That said, much more work is needed to evaluate the generality of the findings across both different stimuli and longer-retention intervals and to assess whether RCJs impact overall learning when participants are given an opportunity to carry out their restudy decisions. Many studies have shown that JOLs lead to increased learning outcomes, and our current work does not contradict these findings, as participants in the JOL condition selected items for restudy that they retrieved incorrectly (the most likely source of improvement). However, the question becomes whether learning can be increased even more with RCJs as participants will spend less study time reviewing material that is already learned.

Action Editor

Kathleen McDermott served as action editor for this article.

Author Contributions

A. M. Robey and M. R. Dougherty developed the study concept. A. M. Robey, M. R. Dougherty, and D. R. Buttaccio designed the study. A. M. Robey and D. R. Buttaccio programmed the study. A. M. Robey and M. R. Dougherty wrote the first draft of the manuscript. D. R. Buttaccio provided critical edits to the manuscript.

Declaration of Conflicting Interests

The authors declared that they had no conflicts of interest with respect to their authorship or the publication of this article.

Funding

The research reported here is based on work supported by the National Science Foundation under Grant Nos. SES-1426831 and BCS-1522699.

Supplemental Material

Additional supporting information can be found at http://journals.sagepub.com/doi/suppl/10.1177/0956797617718800

Open Practices

All data and materials have been made publicly available via the Open Science Framework. Data can be accessed at https://osf.io/n6d9f/. The PsychoPy script used to collect the data is provided at https://osf.io/vqbzw/, along with other materials. The design and analysis plans for the replication were preregistered at https://osf.io/e4rpu. The complete Open Practices Disclosure for this article can be found at http://journals.sagepub.com/doi/suppl/10.1177/0956797617718800. This article has received badges for Open Data, Open Materials, and Preregistration. More information about the Open Practices badges can be found at http://www.psychologicalscience.org/publications/badges.

References

Brunswik, E. (1956). *Perception and the representative design of psychological experiments* (2nd ed.). Berkeley: University of California Press.

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

Dougherty, M. R., & Thomas, R. P. (2012). Robust decision making in a nonlinear world. *Psychological Review*, *119*, 321–344.

Dunlosky, J., & Metcalfe, J. (2009). *Metacognition*. Thousand Oaks, CA: Sage.

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

Hammond, K. R., McClelland, G., & Mumpower, J. (1980). *Human judgment and decision making: Theories, methods, and procedures*. New York, NY: Praeger Scientific.

Hammond, K. R., Rohrbaugh, J., Mumpower, J., & Adelman, L. (1977). Social judgment theory: Applications in policy formation. In M. F. Kaplan & S. Schwartz (Eds.), *Human judgment and decision processes in applied settings* (pp. 1–29). New York, NY: Academic Press.

Han, A. (1987). Non-parametric analysis of a generalized regression model: The maximum rank correlation estimator. *Journal of Econometrics*, *35*, 303–316. doi:10.1016/0304-4076(87)90030-3

Hines, J. C., Touron, D. R., & Hertzog, C. (2009). Metacognitive influences on study time allocation in an associative recognition task: An analysis of adult age differences. *Psychology of Aging*, *24*, 462–475.

Karelaia, N., & Hogarth, R. M. (2008). Determinants of linear judgment: A meta-analysis of lens model studies. *Psychological Bulletin*, *134*, 404–426.

Kimball, D. R., Smith, T. A., & Muntean, W. J. (2012). Does delaying judgments of learning really improve the efficacy of study decisions? Not so much. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*, 923–954.

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

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

Metcalfe, J. (2009). Metacognitive judgments and control of study. *Current Directions in Psychological Science*, *18*, 159–163.

Metcalfe, J., & Kornell, N. (2005). A Region of Proximal Learning model of study time allocation. *Journal of Memory and Language*, *52*, 463–477.

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

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

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

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

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

Ryals, A. J., Rogers, L. M., Gross, E. Z., Polnaszek, K. L., & Voss, J. L. (2016). Associative recognition memory awareness improved by theta-burst stimulation of frontopolar cortex. *Cerebral Cortex*, *26*, 1200–1210. doi:10.1093/cercor/bhu311

Son, L. K., & Metcalfe, J. (2005). Judgments of learning: Evidence for a two-stage process. *Memory & Cognition*, *3*, 1116–1129.

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

Wattier, N. W., & Collins, C. A. (2011). Metamemory for faces, names, and common nouns. *Acta Psychologica*, *138*, 143–154.

Wilson, M. (1988). MRC psycholinguistic database: Machine-usable dictionary, version 2.00. *Behavior Research Methods, Instruments, & Computers*, *20*, 6–10.