Mem Cogn (2011) 39:171–184
DOI 10.3758/s13421-010-0002-y

The sensitivity of judgment-of-learning resolution to past test performance, new learning, and forgetting

Robert Ariel · John Dunlosky

Published online: 9 November 2010

Abstract When people judge their learning of items across study–test trials, their accuracy in discriminating between learned and unlearned items improves on the second trial. We examined the source of this improvement by estimating the contribution of three factors—memory for past test performance (MPT), new learning, and forgetting—to accuracy on trial 2. In Experiment 1, during an initial trial, participants studied paired associates, made a judgment of learning (JOL) for each one, and were tested. During the second trial, we manipulated two variables: when the JOL was made (either immediately before or after studying an item) and whether participants were told the outcome of the initial recall attempt on trial 1. In Experiment 2, the same procedure was used with a 1-week retention interval between study and test on trial 2. In both experiments, JOL resolution was higher on trial 2 than on trial 1. Fine-grained analyses of JOL magnitude and decomposition of resolution supported several conclusions. First, MPT contributed the most to boosts in JOL magnitude and improvements in resolution across trials. Second, JOLs and subsequent resolution were sensitive to new learning and forgetting, but only when participants' judgments were made after study. Thus, JOLs appear to integrate information from multiple factors, and these factors jointly contribute to JOL resolution.

Keywords Judgment of learning · Resolution · Memory for past test · Multi-trial learning

Since Arbuckle and Cuddy's (1969) seminal article on judgments of learning (JOLs), the accuracy of people's JOLs has been intensely scrutinized (Dunlosky & Metcalfe, 2009). JOLs are predictions of the likelihood of recalling recently studied items, and their accuracy is typically estimated using the following method. Learners study a list of paired associates (e.g., *dog–spoon*); immediately after studying a given pair, they judge the likelihood of recalling the response when shown the cue on the upcoming test trial (i.e., *dog–*?). Accuracy is then measured by comparing JOLs with recall performance: Resolution refers to the degree to which JOLs discriminate between the recall of one item relative to another, whereas calibration refers to the degree to which the magnitude of judgments relates to the absolute level of performance. The effects of repeated study–test trials on JOL accuracy have been of major interest, partly because restudy trials can have both beneficial effects (increases in resolution; Koriat, 1997) and deleterious effects (decreases in calibration; Koriat, Sheffer, & Ma'ayan, 2002) on the accuracy of JOLs. Concerning resolution, people's JOLs typically increase in resolution with additional study–test practice, in that they better discriminate at the item level between subsequently recalled and unrecalled items. By contrast, these improvements in JOL resolution are typically accompanied by a decrease in calibration, which is characterized by a shift toward underconfidence with practice (Koriat et. al., 2002). That is, participants' average JOLs are typically higher than recall performance on an initial study–test trial, but their average JOLs on a second study–test trial become lower than recall performance.

Although some progress has been made in understanding the underconfidence that arises on a second trial (Finn & Metcalfe, 2007, 2008; Scheck & Nelson, 2005; Serra & Dunlosky, 2005), explanations for the equally important increase in resolution across repeated study–test trials have not been systematically explored. To fill this gap, we empirically evaluate the degree to which three factors—

R. Ariel · J. Dunlosky (✉)
Psychology Department, Kent State University,
Kent, OH 44242, USA
e-mail: jdunlosk@kent.edu

memory for past test performance, new learning, and forgetting—contribute to the effects of repeated study–test trials on JOL resolution. We also present data relevant to calibration for interested readers, but given that our focus is on JOL resolution, we do not discuss calibration any further. We describe the three factors next and then discuss our approach to estimating their contribution to JOL resolution.

People's JOLs are influenced by various factors (or cues), and the degree to which these factors are diagnostic of future test performance will influence JOL resolution (Koriat, 1997). During repeated study–test trials, one factor that is particularly diagnostic of future recall performance is prior test performance. People are extremely accurate at identifying which items they correctly recalled on previous retrieval attempts (Gardiner & Klee, 1976), and incorporating memory of past test performance could potentially lead to higher accuracy in discriminating between items that will versus will not be recalled in the future. Several researchers have suggested that increases in resolution observed following test trials are a result of relying on past test performance (e.g., Finn & Metcalfe, 2008; King, Zechmeister, & Shaughnessy, 1980). For instance, Finn and Metcalfe (2007) examined the influence of memory for past test (MPT) on JOLs across two study–test trials. They found that JOLs made on trial 2 were more highly associated with trial 1 recall than with trial 2 recall. This finding suggests that people rely on their memory for performance on a previous test trial to make subsequent JOLs.

No published experiments have estimated the joint contribution of MPT and other potentially influential factors to trial 2 judgment resolution. In fact, researchers have almost exclusively investigated the effects of single factors on JOLs in isolation (for recent exceptions, see Benjamin, 2005; Hertzog, Dunlosky, Robinson, & Kidder, 2003; Metcalfe & Finn, 2008). Two other factors that may contribute to improvements in resolution across trials are new learning and forgetting. That is, after an initial study–test trial, JOL resolution not only may be sensitive to MPT, but also could be sensitive to (1) new learning for pairs that had not been previously recalled or (2) forgetting of previously recalled pairs before the next test.

In the present experiments, we operationalized new learning and forgetting as a change in recall status across the two trials. Newly learned items are those that are not recalled on the initial study–test trial but are recalled on the second test trial. Forgotten items are those that are recalled during the initial test trial that are not recalled on the second test trial. The question is, are people's JOLs sensitive to these changes in the recall status across study–test trials? For new learning, consider items that a participant fails to recall on trial 1 (see Fig. 1, Box A,). Assuming that some of these items are recalled on trial 2, JOLs on trial 2 would be considered sensitive to this new learning if they are higher for the items that are correctly recalled on trial 2 ("newly learned" by our operationalization) than for those that are not. For forgetting, items that are recalled on trial 1 are relevant (Fig. 1, Box B), with the issue being whether JOLs made on trial 2 are lower for those items that are subsequently not recalled (i.e., "forgotten") than for those that are recalled.

From an inspection of Fig. 1, it may be evident why we focused on these three factors; namely, they comprise all the comparisons among items when conditionalzed on trial 1 recall status. For instance, comparisons made within nonrecalled items on trial 1 are relevant to new learning; comparisons made within recalled items on trial 1 are relevant to forgetting; the remaining comparisons (signified by C in Fig. 1) are most relevant to MPT. Note, however, that demonstrating that JOLs are sensitive to new learning or forgetting does not explain the mechanism that causes this sensitivity. That is, this sensitivity could be explained by several mechanisms, which include incorporating beliefs about learning or forgetting into JOLs or relying on some other cue experienced while items are being processed that is in itself diagnostic of new learning or forgetting. In the present experiments, we focused primarily on estimating the sensitivity of JOL resolution on trial 2 to the aforementioned factors. On the basis of empirical evidence with regard to sensitivity, we further discuss and evaluate mechanism in the General discussion.

Given that memory for past test performance is such a diagnostic cue, our questions are: Do people's JOLs also predict new learning and forgetting when they are made on trial 2? That is, do these factors contribute to JOL resolution on trial 2? Finn and Metcalfe (2008) provided some evidence that people's JOLs may be sensitive to new learning on trial 2. For items that were incorrectly recalled on trial 1, people gave higher JOLs to those items that were subsequently recalled on trial 2 than to those that were not recalled. However, the difference in JOL magnitude between these classes of items was small, which suggests that the impact of new learning on JOLs may be negligible (Finn & Metcalfe, 2008). Furthermore, people may also incorporate cues that predict forgetting into their trial 2 judgments, which also could contribute to JOL resolution. Rawson, Dunlosky, and McDonald (2002) have provided some evidence that people's predictions incorporate assessments about retention. Participants read brief texts and then made judgments either about their performance on a later test or about current comprehension of the text they had read. Judgment magnitude was lower for performance predictions than for comprehension judgments, suggesting that people incorporated information about potential forgetting into their judgments (see also Koriat, Bjork, Sheffer, & Bar, 2004). Most important, the sensitivity of JOL resolution to MPT, new learning, and forgetting has not yet been explored.

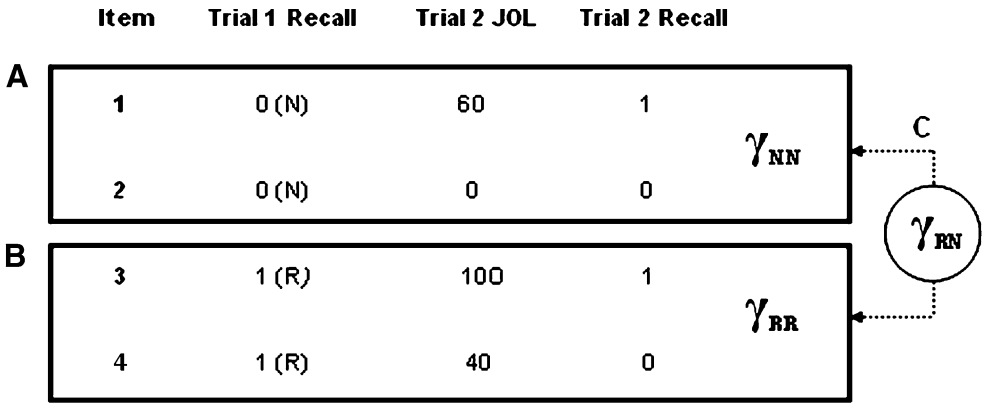

Fig. 1 Illustration of the decomposition of JOL resolution

A major goal of the present experiments was to estimate the joint contribution of these factors to JOL resolution following a study–test trial. To do so, we relied on both a recent methodological advance (prestudy JOLs; Castel, 2008) and an analytical approach to decompose resolution (Nelson, Narens, & Dunlosky, 2004). During trial 1, all participants studied paired associates (e.g., *dog–spoon*), made a JOL immediately after studying each pair, and subsequently completed a cued-recall test (*dog*–?). During a second study–test trial, participants made immediate JOLs (as on trial 1) or made JOLs immediately before studying each item (prestudy JOL). The procedure for immediate and prestudy JOLs is illustrated in Fig. 2. The key difference lies in the timing of the judgment in relation to study. For an immediate JOL (Fig. 2, left panel), participants are first presented an item for study (*dog–spoon*); the item is then removed according to the set presentation rate (e.g., 6 s), and a JOL is made in which the participant is asked, "For the pair you just studied, what is the likelihood that you will recall the second word when presented with the first word on the upcoming test." By contrast, for a prestudy JOL (Fig. 2, right panel), participants are prompted to make a JOL before the item is presented for study. In this case, a participant is asked, "For the pair you are about to study, what is the likelihood that you will recall the second word when presented with the first word on the upcoming test." After making their prestudy JOL, the item is presented for study. Thus, prestudy JOLs cannot be influenced by item-specific cues that people may use when making immediate JOLs (Castel, 2008). Put differently, the resolution of prestudy JOLs cannot be sensitive to MPT performance, new learning, or forgetting.

Most important, the use of Castel's (2008) prestudy–JOL method, along with feedback about past test performance, provided further leverage on the sensitivity of JOL resolution to MPT. Concerning feedback, when making each JOL on trial 2, some participants were told whether they had previously recalled the response to the to-be-judged item. This prompt, when combined with prestudy JOLs, allowed us to examine the sole contribution of knowledge of past test performance to trial 2 JOLs, because the only information participants can use in this circumstance is how they performed on the previous test. If knowledge of past test performance is solely responsible for JOL resolution on trial 2, participants will be just as accurate for prestudy JOLs that are prompted as they are for immediate JOLs.

To estimate the sensitivity of JOL resolution to MPT, new learning, and forgetting, we used the decomposition introduced by Nelson, Narens, and Dunlosky (2004). Resolution on trial 2 was separated into three measures so that we could estimate the contribution of each of the factors above. The general method used to conduct this decomposition is illustrated in Fig. 1. Resolution is computed for subsets of items conditionalized on trial 1 recall status. Consider the four items presented in Fig. 1, which reflect the four possible outcomes that can occur when recall status is jointly considered across trials 1 and 2. Note that according to our operationalizations above, item 1 represents new learning, and item 4 represents forgetting across the two test trials. Resolution is typically computed by making pairwise comparisons of all items, but resolution can also be computed on subsets of items in order to examine whether JOLs are sensitive to specific factors.

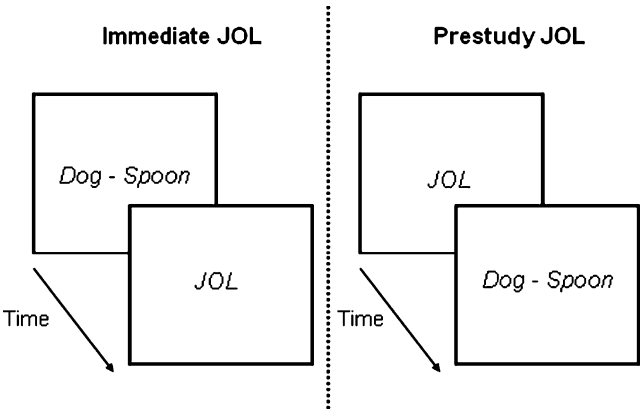

Fig. 2 Timing of JOLs on trial 2, relative to the study trial for an item in the immediate JOL group (left) and prestudy JOL group (right)

Consider items 1 and 2, which are not recalled on trial 1, which are labeled "N" for being not correctly recalled. Item 1 is subsequently recalled on trial 2, but item 2 is not recalled on trial 2. To examine whether JOL resolution on trial 2 is sensitive to new learning, a correlation is computed between trial 2 JOLs and recall performance only for items not recalled on trial 1 (Fig. 1, Box A). This correlation is designated γ_{NN} because it is based on comparing all dyads of previously unrecalled items. Assuming that participants give higher JOLs to items that are recalled on trial 2 (i.e., item 1) than they do for items not recalled later (e.g., item 2), γ_{NN} will be significantly greater than zero, indicating that JOLs are sensitive to new learning. Next, consider the items in Box B, which are labeled "R" because they are correctly recalled on trial 1. By computing resolution on trial 2 only for these items, we can examine whether resolution on trial 2 is sensitive to forgetting. This correlation is designated γ_{RR} and will be greater than 0 if JOL resolution on trial 2 is sensitive to forgetting. Finally, consider comparison C in Fig. 1. After making comparisons only between "N" items (γ_{NN}) and only between "R" items (γ_{RR}), the remaining correlation to be computed involves comparing items in Box A only with those in Box B. In this case, comparisons are made between items recalled on trial 1 (Box A) versus items that were not recalled on trial 1 (Box B), so that the resulting correlation (γ_{RN}) capitalizes on people's ability to remember past memory performance and discriminate between previously remembered items versus those that were not remembered. Given that these three estimates completely account for trial 2 resolution (i.e., the correlation across all items; Nelson, Narens, & Dunlosky, 2004), the size of the estimates indicates the relative contribution of each factor to JOL resolution.

Experiment 1

Method

Participants

One hundred twenty-five students from Kent State University participated for either course credit in introductory psychology or for $10. A 2 (JOL type: immediate or prestudy) ×2 (recall prompt: prompt about previous recall performance or no prompt, as described below) full-factorial design was used. Participants were randomly assigned to either the immediate JOL group without prompts (henceforth, *immediate JOL no-prompt* group; total $n = 31$, paid $n = 13$), the immediate JOL group with prompts (*immediate JOL prompt group*; total $n = 32$, paid $n = 12$), the prestudy JOL group without prompts (*prestudy JOL no-prompt* group; total $n = 31$, paid $n = 12$), or the prestudy JOL group with prompts (*prestudy JOL prompt group*; total $n = 31$, paid $n = 14$).

Materials and procedure

Sixty unrelated noun–noun paired associates were used in this experiment (e.g., *icebox–acrobat*). Participants completed the experiment individually on a computer. They were instructed that their goal was to study word pairs, make JOLs, and complete a paired-associate recall test. Participants completed this study–judge–test cycle twice for the same pairs. The order of presenting pairs was randomized on each trial.

Trial 1 was identical for all the participants: They studied items individually for 6 s. Immediately after an item had been studied, it was replaced with this prompt for a JOL: "For the pair you just studied, what is the likelihood that you will be able to recall the second word when later presented with the first word during the upcoming test?" Participants typed any value between 0 (*I definitely won't be able to recall this item*) and 100 (*I definitely will recall this item*). After studying all the items, participants completed a paired-associate recall test for all the pairs. Participants were not given feedback about whether their recall responses were correct or incorrect.

During trial 2, participants restudied the same 60 word pairs (6 s/pair) and again made JOLs. Participants made either JOLs immediately after studying an item (as in trial 1) or prior to studying the item (prestudy JOLs). Participants in the immediate JOL groups were given the same JOL prompt as on trial 1. Prestudy JOLs were obtained using the following prompt (Castel, 2008): "For the pair you are about to study, what is the likelihood that you will be able to recall the second word when later presented with the first word during the upcoming test?" The same scale was used for all JOLs. Also, immediately prior to making either JOL, some participants also were told whether they had correctly recalled the current item, using this *recall prompt*: "When tested, you correctly recalled (or did not correctly recall) the response to the pair that you just studied (or that you are about to study)." After participants finished studying the word pairs, a final test was administered.

Results

JOLs and recall performance

Although analysis of JOL resolution is most relevant to achieving our present goals, we begin by presenting the overall magnitude of JOLs and recall performance across items. Means across participants' mean JOLs and the percentages of items correctly recalled are presented in Table 1. JOLs did not differ among groups, $F(3, 121) = .36$, $MSE = 167.04$, $p = .78$, or between trials, $F(1, 121) = 0.84$, $MSE = 123.11$, $p = .36$. A group × trial interaction effect

Table 1 Means across individuals' mean judgments of learning (JOLs) and percentages of correct recall for trial 1 and trial 2

Group	Trial 1		Trial 2	
	JOL	Recall	JOL	Recall
Experiment 1				
iJOL–no prompt	35.3 (2.9)	20.6 (2.4)	41.3 (3.5)	51.2 (4.2)
iJOL–prompt	41.9 (2.7)	23.7 (2.5)	39.8 (2.7)	54.2 (4.2)
pJOL–no prompt	40.6 (2.7)	24.2 (2.4)	40.1 (3.5)	54.5 (3.8)
pJOL–prompt	36.3 (3.6)	20.6 (2.4)	38.6 (3.5)	53.5 (4.3)
Experiment 2				
iJOL–no prompt	39.8 (2.9)	36.5 (3.6)	33.8 (3.3)	11.8 (2.9)
iJOL–prompt	39.7 (2.9)	40.8 (3.5)	36.3 (3.5)	16.3 (2.6)
pJOL–no prompt	35.3 (2.8)	31.4 (3.1)	29.7 (4.0)	9.7 (1.4)
pJOL–prompt	40.4 (3.3)	36.3 (3.5)	31.0 (3.5)	11.4 (1.5)

Note: Values are means across individual's mean values. Standard errors of the means are in parentheses. iJOL–no prompt group = immediate JOL with no recall prompt, iJOL–prompt group = immediate JOL with recall prompt, pJOL–no-prompt group = prestudy JOL with no recall prompt, and pJOL–prompt group = prestudy JOL with recall prompt

was not significant, $F(3, 121) = 1.35$, $MSE = 198.97$, $p = .26$. Recall performance also did not differ among groups, $F(3, 121) = .27$, $MSE = 163.89$, $p = .85$. Recall did significantly improve across trials, $F(1, 121) = 641.31$, $MSE = 6.02$, $p < .001$, $\eta_p^2 = .84$, and the group×trial interaction was not significant, $F(3, 121) = .26$, $MSE = 24.19$, $p = .86$.

JOL resolution

To examine JOL resolution, we computed gamma correlations between the participants' JOLs and their recall performance across items.[1] During trial 1, all participants made immediate JOLs, and, as was expected, resolution did not differ between the immediate JOL no-prompt group ($M = .44$, $SE = .05$), the immediate JOL prompt group ($M = .54$, $SE = .04$), the prestudy JOL no-prompt group ($M = .46$, $SE = .04$), and the prestudy JOL prompt group ($M = .43$, $SE = .06$), $F(3, 124) = 1.07$, $MSE = 0.07$, $p = .36$.

Most important, consider resolution on trial 2. As is evident from inspection of Fig. 3, resolution was higher for the immediate JOL groups than for the prestudy JOL group that received the recall prompt, which indicates that general knowledge of past test performance is not the only factor contributing to trial 2 resolution. Consistent with this observation, a 2 (JOL type) ×2 (recall prompt) ANOVA revealed a main effect of JOL, $F(1, 118) = 54.89$, $MSE = 5.64$, $p < .001$, $\eta_p^2 = .32$. Resolution was also greater for the prompt groups than for the nonprompted groups, $F(1, 118) = 12.77$, $MSE = 1.31$, $p < .001$, $\eta_p^2 = .10$. The JOL type × prompt interaction approached significance, $F(1, 118) = 3.57$, $MSE = 0.37$, $p = .06$.

[1] Similar analyses were conducted using signal detection measures of discriminative accuracy, and the results yielded the same outcomes as gamma correlations. Thus, to remain consistent with the vast literature on JOLs, we report gamma correlations.

Trial 2 JOLs conditionalized on trial 1 and trial 2 recall

To provide an initial assessment about the potential contribution of the three factors to JOL resolution, we examined JOL magnitude on trial 2 for various classes of items. In particular, we conducted an analysis similar to that in Finn and Metcalfe (2007; see also Finn & Metcalfe, 2008), in which JOLs were conditionalized on trial 1 and trial 2 recall performance. JOL magnitude was computed for four subsets of items: (1) items not recalled on trial 1 that were also not recalled on trial 2 (N_1N_2 items),[2] (2) items not recalled on trial 1 that were recalled on trial 2 (N_1R_2 items), (3) items recalled on trial 1 that were recalled on trial 2 (R_1R_2 items), and (4) items recalled on trial 1 that were not recalled on trial 2 (R_1N_2 items). Mean JOLs are presented in Fig. 4.

To examine the influence of MPT, we compared N_1R_2 with R_1R_2 items. If people are relying on MPT to make trial 2 JOLs, JOLs will be higher for R_1R_2 items than for N_1R_2 items. Trial 2 recall performance is held constant in these comparisons (all items were recalled on trial 2), so the influence of other factors, such as new learning and forgetting, should not influence judgment magnitude (Finn & Metcalfe, 2007). A 2 (item status: R_1R_2 vs. N_1R_2) ×2 (JOL type) ×2 (recall prompt) ANOVA revealed an effect for JOL type, $F(1, 120) = 18.21$, $MSE = 12{,}011.53$,

[2] For this analysis conditionalized on change in recall status across trials, we designate recall status as follows: R = recalled, N = not recalled. The first value designates trial 1 performance and, hence, has a subscript of "1," whereas the second value designates trial 2 performance and has a subscript of "2." So, N_1R_2 means the subset of items not correctly recalled on trial 1 that were subsequently recalled on trial 2. Note, however, that for the decomposition of correlations, both values (N or R) pertain to recall status on trial 1; thus, γ_{RN} (see Fig. 1, comparison C) indicates the correlation computed by comparing all items correctly recalled (R) on trial 1 with all items not correctly recalled (N) on trial 1. Given that both values refer to trial 1, we did not include the subscript.

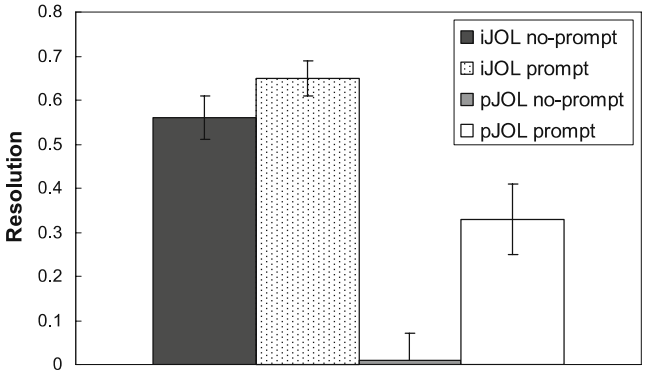

Fig. 3 Trial 2 resolution as a function of JOL group in Experiment 1. iJOL = immediate JOLs; pJOL = prestudy JOLs. Prompt/no prompt = whether participants were prompted with prior recall outcome. Error bars represent standard errors of the means

$p < .001$, $\eta_p^2 = .13$, which indicates that immediate JOLs were higher than prestudy JOLs. Participants also made higher JOLs for R_1R_2 than for N_1R_2 items, $F(1, 120) = 180.48$, $MSE = 53{,}044.409$, $p < .001$, $\eta_p^2 = .60$. The item status × JOL type, $F(1, 120) = 50.45$, $MSE = 14{,}827.35$, $p < .001$, $\eta_p^2 = .30$, and item status × prompt, $F(1, 120) = 16.37$, $MSE = 4{,}812.35$, $p < .001$, $\eta_p^2 = .12$, interactions were significant. The three-way interaction approached significance, $F(1, 120) = 3.32$, $MSE = 975.42$, $p = .07$.

Planned comparisons were conducted to examine whether differences between N_1R_2 and R_1R_2 items were greater when participants had access to item-specific cues at the time of judgments (immediate JOL groups) than when participants were just provided with knowledge of prior test performance (prestudy JOL prompt group). A 2 (item status: N_1R_2 vs. R_1R_2) ×2 (group: immediate JOL prompt vs. prestudy JOL prompt) ANOVA revealed effects for item status, $F(1, 60) = 107.03$, $MSE = 44{,}905.50$, $p < .001$, $\eta_p^2 = .64$, and for group, $F(1, 60) = 6.12$, $MSE = 3{,}687.65$, $p < .05$, $\eta_p^2 = .09$. These main effects were qualified by an item status × group interaction, $F(1, 60) = 180.48$, $MSE = 4{,}098.37$, $p < .01$, $\eta_p^2 = .14$. Separate one-way ANOVAs were conducted to examine differences in JOL magnitude for N_1R_2 and R_1R_2 items between groups. JOL magnitude for N_1R_2 items did not differ between groups, $F(3, 122) = 1.24$, $MSE = 485.51$, $p = .29$; however, JOL magnitude was significantly different for R_1R_2 items, $F(3, 121) = 17.47$, $MSE = 10{,}179.14$, $p < .001$, $\eta_p^2 = .30$. Tukey post hoc tests revealed that the immediate JOL groups gave significantly higher JOLs to R_1R_2 items than did either of the prestudy JOL groups, and the prestudy JOL prompt group gave higher JOLs to R_1R_2 items than did the prestudy JOL no-prompt groups. No other differences between groups were significant.

In summary, these planned comparisons revealed two important effects: (1) JOL differences between N_1R_2 and R_1R_2 items were greater for the immediate JOL groups than for the prestudy JOL prompt group, and (2) such differences arose because JOL magnitudes for previously *recalled* items were significantly greater for the immediate JOL groups than for the prestudy JOL prompt group. Thus, JOLs were the highest when participants not only

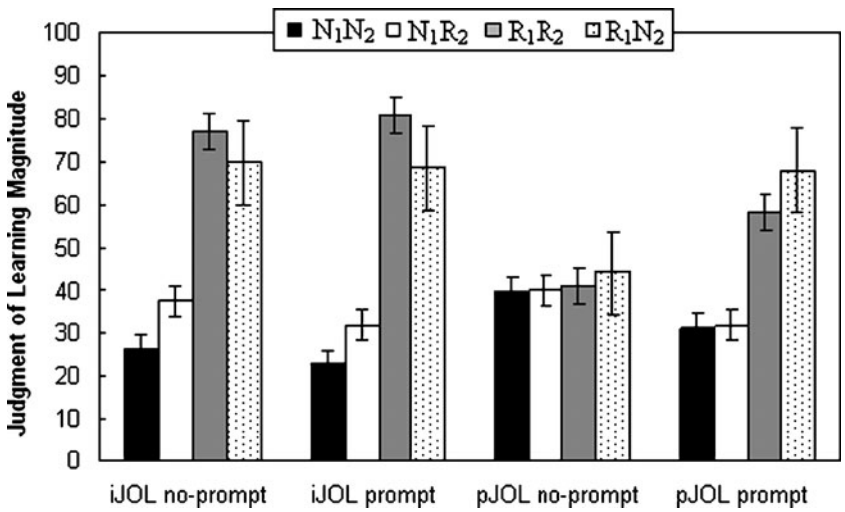

Fig. 4 Trial 2 mean judgments of learning (JOLs) in Experiment 1 conditionalized on trial 1 and trial 2 recall. R = correctly recalled, N = not correctly recalled; subscripts, 1 = recall status on trial 1 and 2 = recall status on trial 2. iJOL = immediate JOL; pJOL = prestudy JOL. Prompt/no prompt = whether participants were prompted with prior recall outcome. Error bars represent standard errors of the means. Overall JOL magnitude on trial 2 for each group cannot be obtained by averaging JOLs across subsets of items, because the number of items within each subset is not equal

knew whether a response had been recalled but also knew specifically *which* response had been recalled.

Concerning the other factors, higher JOLs for N_1R_2 than for N_1N_2 items would suggest that participants' JOLs are sensitive to new learning on trial 2. Consistent with this prediction, a 2 (item status: N_1R_2 vs. N_1N_2) ×2 (JOL type) ×2 (prompt) ANOVA revealed an effect for item status, $F(1, 122) = 32.84$, $MSE = 1{,}617.06$, $p < .001$, $\eta_p^2 = .21$, and an item status × JOL type interaction, $F(1, 122) = 28.63$, $MSE = 28.63$, $p < .001$, $\eta_p^2 = .19$, which arose because only the immediate JOL groups were sensitive to factors pertaining to new learning. JOLs also appeared to reflect sensitivity to predicting forgetting for the immediate JOL groups, because they gave lower JOLs to R_1N_2 (forgotten) than to R_1R_2 (remembered) items. However, differences between R_1R_2 and R_1N_2 items were not significant for any group.

Estimating the contributions of MPT, new learning and forgetting

On the basis of the previous analysis of JOL magnitudes (Fig. 4), MPT and new learning appear to contribute the most to JOL resolution on trial 2. To estimate their contributions, we computed partial correlations, using the decomposition described by Nelson, Narens and Dunlosky (2004). Overall resolution (Fig. 3) was decomposed into three correlations reflecting (1) discriminations between items that were recalled versus not recalled on trial 1 (or γ_{RN}, which reflects the influence of MPT and is represented by comparison C in Fig. 1), (2) discriminations between dyads of items that were not recalled on trial 1 (γ_{NN}, which reflects sensitivity to new learning, comparisons in Fig. 1, Box A), and (3) discriminations between dyads of items that were recalled on trial 1 (γ_{RR}, which reflects sensitivity to forgetting, comparisons in Fig. 1, Box B).

Overall resolution is composed of the discriminations above, as weighted in this equation (Nelson, Narens & Dunlosky, 2004):

$$\gamma = (P_{RN}{}^{*}\gamma_{RN}) + (P_{NN}{}^{*}\gamma_{NN}) + (P_{RR}{}^{*}\gamma_{RR}). \quad (1)$$

The parameters P_{RN}, P_{NN}, and P_{RR} reflect the proportion of dyads that contributed to the computation of each respective correlation. A low value for P indicates that the corresponding factor will have a limited influence on overall resolution. Mean estimates for Equation 1 parameters are presented in Table 2. We rounded to two decimal places, and all P parameters declared as 0 in Table 2 are actually very small proportions; thus, gamma values can be computed for all cells, but gammas corresponding to P parameters of (near) zero are unstable, because few observations contributed to their computation.

Consider γ_{RN}, which is relevant to the contribution of MPT to resolution on trial 2. A 2 (JOL type) ×2 (recall prompt) ANOVA examining γ_{RN} revealed an effect for JOL type, $F(1, 121) = 69.17$, $MSE = 9.57$, $p < .001$, $\eta_p^2 = .37$, which indicates that on trial 2, the immediate JOL groups were more accurate at discriminating between items previously recalled versus not recalled. Prompting participants about their past performance also boosted γ_{RN}, $F(1, 121) = 12.77$, $MSE = 3.12$, $p < .001$, $\eta_p^2 = .16$. However, this prompt had a larger influence on the prestudy JOL groups, resulting in a JOL type × prompt interaction, $F(1, 121) = 6.07$, $MSE = 0.84$, $p < .05$, $\eta_p^2 = .05$.

Next we compared γ_{NN} values to evaluate the sensitivity of JOL resolution to new learning on trial 2. Note that in the prestudy JOL groups, JOL resolution should not be sensitive to new learning, because JOLs were made prior to actually studying the items. Consistent with this prediction, the values for both prestudy JOL groups did not differ from zero, ts < 1. By contrast, γ_{NN} values for both immediate JOL groups were significantly greater than zero, ts > 4.5, which suggests that JOL resolution was sensitive to new learning on trial 2. A 2 (JOL type) ×2 (recall prompt) ANOVA revealed a significant effect of JOL type, $F(1, 121) = 23.64$, $MSE = 3.82$, $p < .001$, $\eta_p^2 = .17$. The main effect for recall prompt, $F(1, 121) = 0.11$, $MSE = 0.02$, $p = .74$, and the interaction, $F(1, 121) = .38$, $MSE = 0.06$, $p = .54$, were not significant.

Inferential statistics for γ_{RR}, which reflect the contribution of forgetting to trial 2 resolution, will not be presented, because these values are based on only a small subset of participants ($n = 31$). The reason for this outcome is evident from the low P_{RR} values, which indicate that few of the items that were recalled on trial 1 were not recalled on trial 2. Thus, even if participants were incorporating information about forgetting into their trial 2 JOLs, doing so could not influence their resolution.

Discussion

Although a great deal is now known about the heuristic nature of JOLs (Dunlosky & Metcalfe, 2009), the extent to which multiple factors jointly influence JOL resolution has received little empirical attention. The present results suggest that a multifactor approach will be required to provide a complete explanation for JOL resolution, especially when JOLs can be influenced by task experience. Two factors in particular—MPT and new learning—appeared to contribute to improvements in JOL resolution following a study–test trial.

The contribution of MPT and new learning to overall resolution on trial 2 (Fig. 3) can be estimated by examining the P and γ parameters in Table 2. Consider three key outcomes in this table. First, the P_{RN}, P_{NN}, and P_{RR}

Table 2 Decomposition of gamma correlations presented in Figs. 3 and 5

JOL Group	Proportion of Dyads			Partial Gammas		
	P_{RN}	P_{NN}	P_{RR}	γ_{RN}	γ_{NN}	γ_{RR}
Experiment 1						
iJOL no-prompt	.41 (.03)	.59 (.04)	.00 (.00)	.76 (.07)	.39 (.06)	.03 (.27)
iJOL prompt	.47 (.04)	.53 (.04)	.00 (.00)	.92 (.02)	.37 (.07)	.53 (.32)
pJOL no-prompt	.41 (.03)	.57 (.03)	.02 (.01)	.04 (.07)	-.01 (.07)	.14 (.14)
pJOL prompt	.50 (.04)	.50 (.04)	.00 (.01)	.52 (.09)	.06 (.09)	-.31 (.37)
Experiment 2						
iJOL no-prompt	.66 (.04)	.10 (.03)	.24 (.03)	.81 (.04)	.60 (.14)	.41 (.07)
iJOL prompt	.68 (.03)	.09 (.02)	.23 (.04)	.83 (.03)	.42 (.13)	.28 (.08)
pJOL no-prompt	.67 (.04)	.10 (.03)	.23 (.03)	-.00 (.09)	-.12 (.14)	.03 (.08)
pJOL prompt	.66 (.04)	.10 (.04)	.24 (.03)	.26 (.09)	.18 (.16)	.08 (.08)

Note: Values above can be used to estimate overall γ, using Equation 1 (see text). Because the values above are group means, and not individual participant's values, estimates of the overall γ using these group values will not be identical to the actual values (Figs. 3 and 4), due to rounding errors. However, when individual participant's values are imputed into the equation, the exact values reported in Figs. 3 and 4 are obtained

parameters do not differ between groups, $Fs < 2.30$. Thus, differences in resolution on trial 2 are not a function of differences between groups in the number of items previously recalled, newly learned, or forgotten across trials. Instead, they arise because of differences between groups in discriminative accuracy for each subset of items, which is represented by γ_{RN}, $\gamma_{NN,}$ and $\gamma_{RR,}$ respectively. Second, concerning the P parameters, only P_{RN} and P_{NN} differ significantly from zero, $ts > 12$, and, hence, only MPT and new learning could potentially contribute to overall resolution on trial 2. As is discussed below, we further evaluate whether JOL resolution on trial 2 is sensitive to forgetting in Experiment 2. Finally, because γ_{RN} is larger than γ_{NN} in the immediate JOL groups and the prestudy JOL prompt group, MPT resulted in higher boosts to JOL resolution on trial 2 than did new learning.

Experiment 2

In Experiment 1, we used a standard method to demonstrate and explore JOL resolution across multiple trials: Study times were experimenter-paced, the retention interval was short, and study on trial 2 occurred immediately after recall on trial 1. We suspected that manipulating these parameters would moderate the joint contribution of various factors to JOL resolution. For instance, the short retention interval did not yield forgetting, which is evident from the nonsignificant P_{RR} values in Table 2. If a longer retention interval were used on trial 2, so that P_{RR} was greater than 0 (i.e., some items recalled on trial 1 were not recalled on trial 2), JOL resolution might also be sensitive to the forgetting that occurs on trial 2. In Experiment 2, we explored this possibility by extending the retention interval between study and test on trial 2 to 1 week. Pilot data indicated that a 1-week retention interval was sufficient to increase forgetting between test 1 and test 2 and, hence, would produce P_{RR} values greater than zero. If forgetting does contribute to resolution for JOLs made on trial 2, we expected γ_{RR} values to be significantly greater than zero and, hence, contribute to overall JOL resolution.

Method

Participants

One hundred thirty-three students from Kent State University participated for course credit in introductory psychology. A 2 (JOL type: immediate or prestudy) ×2 (recall prompt: prompt about previous recall performance or no prompt, as described below) full-factorial design was used. Participants were randomly assigned to the immediate JOL no-prompt group (n = 33), the immediate JOL prompt group (n = 32), the prestudy JOL no-prompt group (n = 32), or the prestudy JOL prompt group (n = 36).

Materials and procedure

Eighty unrelated noun–noun paired associates were used in this experiment. The procedure was the same as that in Experiment 1, with the following exceptions. First, prior to the initial study–judge–test cycle, each item was presented for 4 s for study. The purpose of this study trial was to increase learning that would occur during trial 1. After this familiarity trial, participants engaged in the first study–judge–test trial. Second, during the study–judge–

test cycles for trial 1 and trial 2, items were presented for 4 s for study, instead of 6 s. Third, the retention interval between study and test on trial 2 was increased to 1 week from the time of the second study trial. The increased retention interval between study and test on trial 2 was expected to increase the likelihood of forgetting. All the participants were informed on the second trial that items would be tested 1 week later.

Results

JOLs and recall performance

Means across participant's mean JOLs and the percentages of items correctly recalled on each trial are presented in Table 1. The magnitude of JOLs did not differ across groups, $F(3, 132) = 0.63$, $MSE = 385.50$, $p = .60$. However, JOLs did decrease across trials for all groups, $F(1, 132) = 19.42$, $MSE = 2{,}506.42$, $p < .001$, $\eta_p^2 = .13$. The group × trial interaction was not significant, $F(3, 132) = 0.78$, $MSE = 3{,}102.86$, $p = .50$. Recall performance also did not differ between groups on either trial, $F(3, 133) = 1.67$, $MSE = 747.05$, $p = .18$. Recall performance decreased across trials for all groups, $F(3, 133) = 350.49$, $MSE = 39{,}348.94$, $p < .001$, $\eta_p^2 = .73$, which indicates that forgetting occurred between trial 1 and trial 2. The group × trial interaction was not significant, $F(3, 133) = .35$, $MSE = 38.92$, $p = .79$.

JOL resolution

As in Experiment 1, we first examined JOL resolution for trial 1. Because all the participants made immediate JOLs on this trial, resolution did not differ between the immediate JOL no-prompt group ($M = .54$, $SE = .03$), the immediate JOL prompt group ($M = .59$, $SE = .03$), the prestudy JOL no-prompt group ($M = .55$, $SE = .05$), and the prestudy JOL prompt group ($M = .54$, $SE = .05$), $F(3, 129) = 0.25$, $MSE = 0.01$, $p = .86$.

Most important, resolution on trial 2 is presented in Fig. 5. Resolution was higher for the immediate JOL groups than for the prestudy JOL prompt group, which replicates findings from Experiment 1 that knowledge of past test performance is not the only factor contributing to trial 2 resolution. Consistent with this observation, a 2 (JOL type) ×2 (recall prompt) ANOVA revealed a main effect of JOL type, $F(1, 126) = 84.40$, $MSE = 10.81$, $p < .001$, $\eta_p^2 = .41$. An effect for recall prompt approached significance, $F(1, 126) = 3.73$, $MSE = 0.48$, $p < .06$, $\eta_p^2 = .03$. The JOL × prompt interaction was significant, $F(1, 126) = 4.88$, $MSE = 0.63$, $p < .05$, $\eta_p^2 = .04$, which indicates that the recall prompt did improve resolution in the prestudy JOL group.

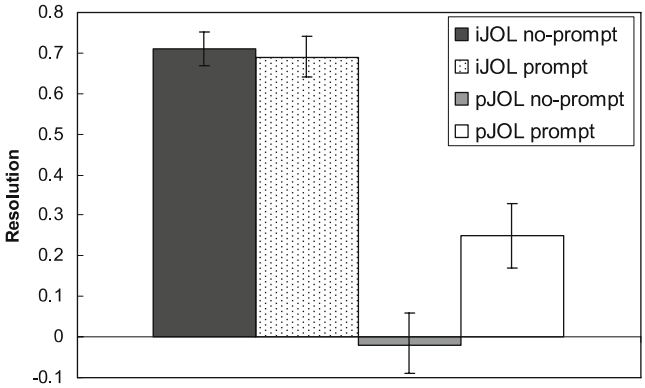

Fig. 5 Trial 2 mean resolution as a function of JOL group in Experiment 2. iJOL = immediate JOLs; pJOL = prestudy JOLs. Prompt/no prompt = whether participants were prompted with prior recall outcome. Error bars represent standard errors of the means

Trial 2 JOLs conditionalized on trial 1 and trial 2 recall

Judgments were conditionalized on the basis of trial 1 and trial 2 recall status to obtain an initial assessment of the influence of MPT, new learning, and forgetting on JOL resolution. Mean JOL magnitude for N_1N_2, N_1R_2, R_1R_2, and R_1N_2 items is presented in Fig. 6. First, we compared N_1N_2 with R_1N_2 pairs to assess the influence of MPT on JOL magnitude. Recall that in Experiment 1, this comparison was made using N_1R_2 and R_1R_2 items. Given the longer retention interval in Experiment 2, new learning between study and test on trial 2 was diminished, resulting in fewer data points for N_1R_2 items, as compared with the other three classes of items. To increase statistical power, we decided to compare N_1N_2 with R_1N_2 items to assess the influence of MPT. The logic behind comparing these two classes of items is similar to that for comparing N_1R_2 with R_1R_2 items. Differences in item status on trial 1 contribute to the impact of MPT on JOL magnitude, whereas item status remains constant on trial 2 to control for the influence of other factors (in this case, potential forgetting).

A 2 (item status: N_1N_2 vs. R_1N_2) ×2 (JOL type) ×2 (prompt) ANOVA revealed effects for JOL type, $F(1, 131) = 5.28$, $MSE = 4{,}417.72$, $p < .001$, $\eta_p^2 = .04$, indicating that JOL magnitude was higher when participants made JOLs after studying items, as compared with prior to studying them. JOL magnitude was also higher for R_1N_2 items than for N_1N_2 items, $F(1, 131) = 216.08$, $MSE = 32{,}075.52$, $p < .001$, $\eta_p^2 = .62$. These main effects were qualified by a JOL type × item status interaction, $F(1, 131) = 126.50$, $MSE = 18{,}777.67$, $p < .001$, $\eta_p^2 = .49$. An item status × prompt interaction was also significant, $F(1, 131) = 9.14$, $MSE = 1{,}356.30$, $p < .01$, $\eta_p^2 = .07$. Thus, when participants made immediate JOLs (as compared with prestudy JOLs), knowledge of past test performance resulted in a larger boost in judgment magnitude.

Concerning the sensitivity of judgment magnitude to new learning, we compared N_1N_2 with N_1R_2 items. A 2 (item

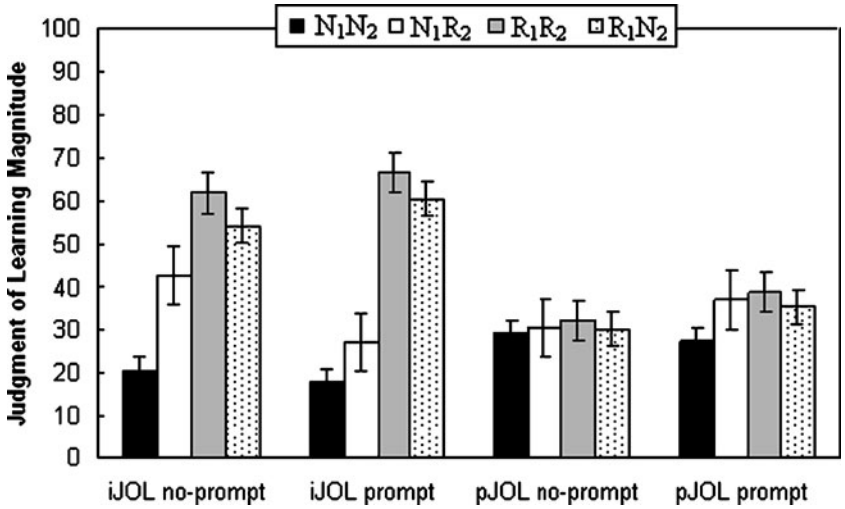

Fig. 6 Trial 2 mean judgments of learning (JOL) in Experiment 2 conditionalized on trial 1 and trial 2 recall. R = correctly recalled, N = not correctly recalled; subscripts, 1 = recall status on trial 1 and 2 = recall status on trial 2. iJOL = immediate JOL; pJOL = prestudy JOL. Prompt/no prompt = whether participants were prompted with prior recall outcome. Error bars represent standard errors of the means. Overall JOL magnitude on trial 2 for each group cannot be obtained by averaging JOLs across all subsets of items, because the number of items within each subset is not equal

status: N_1N_2 vs. N_1R_2) ×2 (JOL type) ×2 (prompt) ANOVA yielded an effect for item status, $F(1, 74) = 14.20$, $MSE = 3{,}739.61$, $p < .001$, $\eta_p^2 = .16$, which indicates that judgment magnitude was higher for N_1R_2 items. An item status × JOL type interaction was also significant, $F(1, 74) = 4.10$, $MSE = 1{,}079.32$, $p < .05$, $\eta_p^2 = .05$. This interaction arose because only immediate JOLs were sensitive to new learning.

Finally, we examined whether people's JOLs predicted forgetting on trial 2. Lower JOLs for R_1N_2 items than for R_1R_2 items would suggest that JOLs are sensitive to forgetting. JOLs were significantly lower for R_1N_2 items than for R_1R_2 items, $F(1, 123) = 27.04$, $MSE = 1{,}311.21$, $p < .001$, $\eta_p^2 = .18$. Moreover, JOLs were higher in magnitude when made immediately after study than when made before study, $F(1, 123) = 33.03$, $MSE = 42{,}523.51$, $p < .001$, $\eta_p^2 = .21$. These main effects were qualified by a item status×JOL type interaction, $F(1, 123) = 11.10$, $MSE = 538.41$, $p < .001$, $\eta_p^2 = .08$. This interaction arose because only immediate JOLs were sensitive to forgetting.

Estimating the contributions of MPT, new learning, and forgetting

We decomposed overall resolution (Fig. 5) into three correlations. First, consider γ_{RN}, which reflects the contribution of MPT (Table 2). A 2 (JOL type) ×2 (recall prompt) ANOVA examining γ_{RN} revealed an effect for JOL type, $F(1, 125) = 102.83$, $MSE = 14.86$, $p < .001$, $\eta_p^2 = .61$. This effect arose because the immediate JOL groups were better than were the prestudy JOL groups at discriminating between items that were previously recalled versus items that were not on trial 2. An effect for prompt was also significant, $F(1, 125) = 4.33$, $MSE = 0.63$, $p < .05$, $\eta_p^2 = .04$. The JOL group × prompt interaction was not significant, $F(1, 125) = 3.20$, $MSE = 0.46$, $p = .08$, $\eta_p^2 = .03$.

Now consider γ_{NN}. If participants' JOLs accurately predicted new learning on trial 2, γ_{NN} should be significantly greater than zero. Consistent with findings from Experiment 1, γ_{NN} did not differ from zero in the prestudy JOL groups, $ts < 1.49$, but was significantly different from zero in the immediate JOL groups, $ts > 2.01$. A 2 (JOL type) ×2 (recall prompt) ANOVA revealed an effect for JOL group, $F(1, 117) = 13.53$, $MSE = 2.44$, $p < .001$, $\eta_p^2 = .11$. Effects for recall prompt, $F(1, 117) = 0.31$, $MSE = 0.58$, $p = .58$, $\eta_p^2 = .003$, and the interaction, $F(1, 117) = 1.32$, $MSE = 0.25$, $p = .25$, $\eta_p^2 = .01$, were not significant.

Finally, in contrast to Experiment 1, forgetting occurred after trial 2 study (P_{RR} values were significantly greater than 0, $ts > 2$), so that we could estimate γ_{RR} values to evaluate the sensitivity of JOL resolution to forgetting. Participants in the prestudy JOL groups were not expected to be able to predict forgetting for items, because they made JOLs prior to actually studying the items on trial 2. Consistent with this prediction, values did not differ from zero for both of the prestudy JOL groups, $ts < 1.48$. By contrast, values for the immediate JOL groups were significantly greater than zero, $ts > 2$. A 2 (JOL type) ×2 (recall prompt) ANOVA revealed an effect for JOL type, $F(1, 73) = 10.96$, $MSE = 4.13$, $p < .001$, $\eta_p^2 = .14$. The effect for recall prompt, $F(1, 73) = 0.15$, $MSE = 0.06$, $p = .70$, $\eta_p^2 = .002$, and the interaction, $F(1, 73) = 2.79$, $MSE = 0.38$, $p = .10$, $\eta_p^2 = .04$, were not significant.

Discussion

When forgetting occurred after study on trial 2, people's JOLs were sensitive to interitem forgetting in a manner that significantly contributed to JOL resolution. As in Experiment 1, people's JOL resolution on trial 2 (Fig. 5) was also influenced by MPT and new learning. The relative contribution of these three factors is presented in Table 2. In the present experiment, 66.8% of the overall JOL resolution on trial 2 for each group was composed of discriminations between items that were recalled versus not recalled on trial 1 (P_{RN} values in Table 2). These high values, in combination with the high γ_{RN} values for participants who could evaluate prior retrieval success (immediate JOL groups and prestudy JOL prompt group), resulted in MPT contributing most to overall resolution. In contrast to Experiment 1, in which only MPT and new learning contributed to resolution on trial 2, forgetting also contributed. In fact, with the longer retention interval to promote forgetting, the weighted contribution of forgetting was larger than the contribution of new learning ($P_{RR} > P_{NN}$).

General discussion

Across two experiments, evidence from multiple analyses converged on the same conclusion that MPT, new learning, and forgetting contributed to JOL resolution when participants engaged in multiple study–test trials. In both experiments, MPT provided the largest contribution to differences in JOL magnitude (Figs. 4 and 6) and JOL resolution on trial 2 (relatively high P_{RN} and γ_{RN} values in Table 2).[3] These results are not entirely surprising, given previous research examining the influence of test experience on metamemory judgments (Finn & Metcalfe, 2007, 2008; King, Zechmeister & Shaughnessy, 1980). Somewhat more surprising is the finding that JOLs are also sensitive to the changes in the recall status for items that occur with additional study–test practice. That is, JOL resolution is sensitive to new learning and forgetting across trials.

Why are JOLs sensitive to new learning and forgetting?

Given that the present experiments rely heavily on correlational data, we concede that a variety of mechanisms could be driving the sensitivity of JOLs to changes in recall status across trials. For instance, people may be directly monitoring new learning and forgetting for items on trial 2. Although possible, previous research suggests that people do not have direct access to states of items in memory but that, instead, their judgments are based on heuristics and influenced by any number of cues (Serra & Metcalfe, 2009). These cues include item relatedness (Carroll, Nelson, & Kirwan, 1997; Koriat, 1997; Rabinowitz, Ackerman, Craik & Hinchley, 1982) and processing fluency (Koriat & Ma'ayan, 2005; Matvey, Dunlosky, & Guttentag, 2001), which in themselves can be predictive of subsequent performance.

One possibility is that these kinds of cue are responsible for the sensitivity of JOLs to new learning and forgetting on trial 2. Because these cues (e.g., processing fluency) are available on trial 1 as well as on trial 2, one expectation is that JOLs made on trial 1 will track the subsequent change in recall status across trials. In particular, for new learning, JOLs made on trial 1 are expected to be greater for N_1R_2 items than for N_1N_2 items. Consistent with this prediction, JOLs on trial 1 were higher for N_1R_2 items than for N_1N_2 items in Experiment 1 (N_1R_2, $M = 40$; N_1N_2, $M = 33$) and in Experiment 2 (N_1R_2, $M = 46$; N_1N_2, $M = 31$), both Fs > 25. For forgetting, JOLs made on trial 1 are expected to be greater for R_1R_2 than for R_1N_2 items. This prediction was evaluated using data from Experiment 2, in which forgetting occurred. As was predicted, JOLs on trial 1 were higher for R_1R_2 ($M = 61$) than for R_1N_2 items ($M = 49$), $F(1, 123) = 83.9$. This analysis indicates that the sensitivity of JOLs to new learning and forgetting on trial 2 is potentially due to reliance on cues that (1) are available even during the first trial, and hence, (2) the sensitivity of JOLs made on trial 2 does not necessarily reflect monitoring of memory strength that results from trial 2 encoding.

In addition to JOLs tapping item-specific cues that are diagnostic of new learning and forgetting, it is also possible that the magnitude of JOLs on trial 2 was influenced by participants' beliefs about their memory (Dunlosky & Matvey, 2001; Koriat, 1997). For example, one explanation for why JOLs were sensitive to forgetting only in Experiment 2 is that participants may have assumed that more forgetting would occur over the long retention interval between study and test on trial 2. Rawson, Dunlosky, and

[3] Technically, the MPT correlation can be influenced when JOLs are sensitive both to new learning and to forgetting within the same participant. For instance, in Fig. 1, if item 1 had a lower JOL (e.g., 40, which would still indicate sensitivity to new learning, as compared with item 2) and item 4 had a higher JOL (e.g., 60, which would still indicate sensitivity to forgetting), such sensitivities would drive down the MPT correlation (i.e., the values above would result in a discordance for comparison C in Fig. 1; for details, see Nelson et al., 2004). Such discordances are unlikely and would arise only when both new learning and forgetting occurred. Given that forgetting did not significantly occur in Experiment 1 and that new learning was relatively minimal in Experiment 2, we expected these potential sensitivities to have a minimal impact on the MPT correlations. Consistent with this expectation, when we recomputed the MPT correlations without including any comparisons that could reflect sensitivity to new learning or forgetting, the MPT correlations changed less than .03 in Experiment 1 and less than .15 in Experiment 2. In both cases, the new estimates were higher than those presented in Table 2 and, hence, still support our main conclusion that MPT contributes most to JOL resolution on trial 2.

McDonald (2002) demonstrated that metacognitive judgments are sensitive to forgetting and concluded that "individuals estimate retention when predicting performance" (p. 505). Subsequently, Koriat, Bjork, Sheffer, and Bar (2004) conducted a series of experiments that suggested that people did not incorporate beliefs about the retention interval into their judgments, because in some of their experiments, judgments did not differ as a function of differing retention intervals (e.g., immediately after study, 1 week after study, or 1 month after study). Nevertheless, when participants were informed about all other possible retention intervals prior to making their predictions (Experiment 5B) and when predictions were framed in terms of forgetting (Experiment 7), they did appear to incorporate knowledge about forgetting into their judgments.

Thus, in the present Experiment 2, it is an open question as to whether the sensitivity of JOLs was influenced by people's use of a forgetting heuristic when making JOLs. To provide preliminary evidence for this possibility, we compared JOL magnitudes on trial 2 across the experiments, because the retention interval was much longer on trial 2 in Experiment 2 than in Experiment 1 (and the retention interval was identical across experiments on trial 1, so that the major procedural difference between experiments was the trial 2 interval).[4] A prediction was that JOL magnitude on trial 2 would be lower in Experiment 2 than in Experiment 1. As was expected, the mean magnitude of JOLs on trial 1 did not differ between experiments (Experiment 1, $M = 39$, $SE = 1.5$; Experiment 2, $M = 39$, $SE = 1.5$), which suggests that the participants were well matched across experiments. More important, JOL magnitude on trial 2 was significantly lower in Experiment 2 ($M = 33$, $SE = 1.8$) than in Experiment 1 ($M = 40$, $SE = 1.6$), $t(259) = 3.00$. Moreover, JOL magnitudes in Experiment 2 also decreased across trials (Table 1). These outcomes suggest that people incorporate retention estimates into their judgments (cf. Rawson, Dunlosky & McDonald, 2002), which in turn may be contributing to the sensitivity of JOL resolution on trial 2. We leave further evaluation of this possibility for future research.

How does memory for past test performance contribute to JOL resolution?

The present approach using the prestudy JOL methodology (Castel, 2008) provides some novel insight into why MPT boosts judgment resolution. Consider the findings from the prestudy JOL prompt group in Experiments 1 and 2. This group made JOLs on trial 2 prior to actually studying items. However, they were provided with the outcome of their previous retrieval attempt for each item—that is, they knew their past test performance for the item being judged. This retrieval outcome is an extremely diagnostic predictor of future memory: The mean gamma correlation between recall on trial 1 and trial 2 was .94 and .84 in Experiments 1 and 2, respectively. Thus, participants in the prestudy JOL prompt groups could have demonstrated high levels of JOL resolution on trial 2 by using this prompt alone, such as by responding with high JOLs when informed that they recalled an item and lower JOLs when informed that they did not. However, participants in the prestudy JOL groups did not fully use this strategy, and their subsequent JOL resolution on trial 2 was relatively low. These observations raise the question of why people use knowledge of prior test performance differently in the immediate JOL groups and the prestudy JOL prompt group?

One answer lies in the differences in the subjective experiences involved in *remembering* a prior retrieval outcome versus *knowing* a prior retrieval outcome. These subjective experiences—called autonoetic versus noetic memories, respectively—are distinct (Tulving, 1985). For *autonoetic memory*, when we remember something (such as whether a study item was correctly recalled on a previous test), we engage in mental time travel that involves retrieving a specific episode of the prior testing event (Gardiner & Richardson-Klavehn, 2000). *Noetic* memory requires only semantic information, as opposed to episodic information (Tulving, 1985). As a consequence, the subjective experience of knowing is often less personal, and as Gardiner and Richardson-Klavehn noted, "There is no awareness [with knowing] of reliving any particular events or experiences" (p. 229).

In the present experiments, the prestudy JOL prompt groups could not have autoneotic memories for the to-be-judged items, because they knew only whether they correctly recalled a given item. In contrast, the immediate JOL groups could remember recalling each specific to-be-judged item on the previous test when making their JOLs on trial 2, and this autonoetic memory may have resulted in a boost of confidence for recalled items. Consistent with this possibility, confidence for memories in general is higher when people experience remembering something (autonoetic memory) versus when they just know something (noetic memory; Dunn, 2004; Tulving, 1985; but see Gardiner & Java, 1990). And, as is shown in Figs. 4 and 6, people's JOLs made on trial 2 were relatively low for previously recalled items in the prestudy JOL prompt group, as compared with the immediate JOL prompt group. This rationale and evidence suggest that one aspect of MPT that boosts confidence in future memory is the subjective experience associated with remembering a *specific* prior retrieval outcome, and not just knowing whether something was correctly recalled.

[4] We thank an anonymous reviewer for suggesting these across-experiment comparisons.

Note, however, that providing participants with knowledge about past test performance when making prestudy JOLs is not equivalent to having a noetic or know experience, because participants do not view the item when making a JOL and, hence, cannot engage in a retrieval attempt that is needed to produce a noetic experience. Thus, it is still unclear whether it is the contents of memory (e.g, autonoetic versus noetic information) that contribute to the boosts in confidence for previously recalled items or whether it is the act of engaging in retrieval itself. Both explanations are plausible, and future research needs to address which of these aspects of MPT most influences JOLs.

Summary

The present experiments used a recent methodological advance (prestudy JOLs,; Castel, 2008) and a decomposition analysis to investigate the joint contribution of MPT, new learning, and forgetting to JOL resolution. Although we used these techniques to explore the sensitivity of JOL resolution to these factors on trial 2, the techniques could be applied in numerous contexts. Concerning the decomposition, it can be used to evaluate the contribution of any factor to overall JOL resolution. For instance, Rhodes and Castel (2008) demonstrated that JOLs are sensitive to font size; JOLs were higher for items that were presented in a larger font than for those presented in a smaller font. JOL accuracy was low, overall, because font size was not diagnostic of performance. The decomposition could be used to evaluate whether this manipulation masked higher levels of resolution within each class of items (e.g., Box A in Fig. 1 could represent items in large font, Box B could represent items in a smaller font, and the analysis would be based on Equation 1). In fact, the contribution of any variable (with discrete levels) to overall JOL resolution could be explored using this decomposition technique.

In the present context, the prestudy JOL method and the decomposition revealed several important conclusions about JOL resolution on a second study–test trial. First, MPT contributed the most to boosts in JOL magnitude and improvements in resolution across trials. Second, participants' JOLs and subsequent resolution were sensitive to new learning and forgetting, but only when the participants' judgments were made after study. Post hoc analysis suggested that such sensitivity arises, in part, because people's JOLs are influenced by cues (e.g., item characteristics or processing fluency) that are present even during an initial study trial. Most important, people's JOLs integrate information from multiple cues, and these cues jointly contribute to JOL resolution after a test trial.

Acknowledgements This project was partially supported by a James S. McDonnell Foundation 21st Century Science Initiative in Bridging Brain, Mind, and Behavior Collaborative Award.

This article is based on a thesis submitted to Kent State University. Thanks to John Gunstad, William Merriman, Maria Zaragoza, and the RADlab for providing excellent comments on and critiques of this project.

References

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

Benjamin, A. S. (2005). Response speeding mediates the contribution of cue familiarity and target retrievability to metamnemonic judgments. *Psychonomic Bulletin & Review, 12*, 874–879.

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

Castel, A. D. (2008). Metacognition and learning about primacy and recency effects in free recall: The utilization of intrinsic and extrinsic cues when making judgments of learning. *Memory & Cognition, 36*, 429–437.

Dunlosky, J., & Matvey, G. (2001). Empirical analysis of the intrinsic–extrinsic distinction of judgments of learning (JOLs): Effects of relatedness and serial position on JOLs. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 27*, 1180–1191.

Dunlosky, J., & Metcalfe, J. (2009). *Metacognition*. Los Angeles: Sage Publications.

Dunn, J. C. (2004). Remember–know: A matter of confidence. *Psychological Review, 2*, 524–542.

Finn, B., & Metcalfe, J. (2007). The role of memory for past test in the underconfidence with practice effect. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 33*, 238–244.

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

Gardiner, J. M., & Java, R. I. (1990). Recollective experience in word and nonword recognition. *Memory & Cognition, 18*, 23–30.

Gardiner, J. M., & Klee, H. (1976). Memory for remembered events: An assessment of output monitoring in free recall. *Journal of Verbal Learning and Verbal Behavior, 15*, 227–233.

Gardiner, J. M., & Richardson-Klavehn, A. (2000). Remembering and knowing. In E. Tulving & F. I. M. Craik (Eds.), *The Oxford handbook of memory* (pp. 229–244). New York: Oxford University Press.

Hertzog, C., Dunlosky, J., Robinson, A. E., & Kidder, D. P. (2003). Encoding fluency is a cue used for judgments about learning. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 29*, 22–34.

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

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

Koriat, A., Bjork, R. A., Sheffer, L., & Bar, S. K. (2004). Predicting one's own forgetting: The role of experience-based and theory-based processes. *Journal of Experimental Psychology: General, 133*, 646–656.

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

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

Matvey, G., Dunlosky, J., & Guttentag, R. (2001). Fluency of retrieval at study affects judgments of learning (JOLs): An analytic or nonanalytic basis for JOLs? *Memory & Cognition, 29*, 222–232.

Metcalfe, J., & Finn, B. (2008). Familiarity and retrieval processes in delayed judgments of learning. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 34*, 1084–1097.

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.

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

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

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

Scheck, P., & Nelson, T. (2005). Lack of pervasiveness of the underconfidence-with-practice-effect: Boundary conditions and an explanation via anchoring. *Journal of Experimental Psychology: General, 134*, 124–128.

Serra, M. J., & Dunlosky, J. (2005). Does retrieval fluency contribute to the underconfidence-with-practice effect? *Journal of Experimental Psychology. Learning, Memory, and Cognition, 31*, 1258–1266.

Serra, M. J., & Metcalfe, J. (2009). Effective implementation of metacognition. In D. J. Hacker, J. Dunlosky, & A. Graesser (Eds.), *Handbook of metacognition in education* (pp. 278–298). New York: Routledge.

Tulving, E. (1985). Memory and consiousness. *Canadian Psychology, 26*, 1–12.