Mem Cogn (2016) 44:1102–1113
DOI 10.3758/s13421-016-0623-x

The influence of feedback on predictions of future memory performance

Danielle M. Sitzman[1] · **Matthew G. Rhodes**[2] · **Nate Kornell**[3]

Published online: 31 May 2016

Abstract Kornell and Rhodes (Journal of Experimental Psychology: Applied, 19, 1–13, 2013) reported that correct answer feedback impairs the accuracy of prospective memory judgments. The current experiments explored the boundaries of this effect. In Experiment 1, participants studied Lithuanian-English word pairs, took an initial test, and were either given correct answer feedback or no feedback at all. They then made a judgment of learning (JOL) regarding the likelihood of correctly recalling the English translation on a later test. Presenting the correct answer as feedback increased average JOLs but impaired relative accuracy on a final test. Therefore, Experiments 2–4 aimed to specifically ameliorate impairments in relative accuracy following feedback. Participants in Experiment 2 were exposed to right/wrong feedback, no feedback, and correct answer feedback while making JOLs. Using such a within-subjects design did not improve relative accuracy following correct answer feedback. Experiment 3 showed that previous exposure to a test-feedback-test cycle did not improve relative accuracy. In Experiment 4, feedback was scaffolded such that the correct answer was progressively revealed. Participants corrected more errors if they could generate the correct response with fewer letter cues. However, relative accuracy did not improve in comparison to the previous experiments. Accordingly, the current experiments suggest that participants may understand that feedback is beneficial, but receiving feedback diminishes prediction accuracy for specific items and participants do not appreciate the magnitude of the benefits of feedback.

Keywords Memory · Metamemory · Feedback

Feedback is highly beneficial for correcting errors in memory. Although much work has examined the efficacy of different forms of feedback (Pashler, Cepeda, Wixted, & Rohrer, 2005), few studies have examined whether people are aware of the benefits of feedback (but see Kornell & Rhodes, 2013; Rhodes & Tauber, 2011). That is, little is known about how information from feedback influences evaluations of one's own learning. Prior research suggests that people are frequently inaccurate when assessing their own future knowledge (see Rhodes, 2016, for a review). For example, learners often overestimate how much they will remember in the future and also frequently fail to appreciate factors (e.g., additional study opportunities) that improve memory (Kornell & Bjork, 2009). Such findings reflect a stability bias in prospective memory judgments, with people regarding memory performance as stable over time while failing to consider future forgetting or future improvements (Koriat, Bjork, Sheffer, & Bar, 2004; Kornell & Bjork, 2009; Kornell, Rhodes, Castel, & Tauber, 2011).

The current experiments explored another potential case of stability bias: Predictions of future memory performance following feedback. Kornell and Rhodes (2013) reported that participants more accurately predicted future test performance following a test *without* feedback than when feedback was provided. Therefore, the goal of the current experiments was to replicate and extend Kornell and Rhodes' findings and explore methods that may help participants better incorporate the benefits of feedback into their predictions of future performance.

✉ Danielle M. Sitzman
dsitzman@ewu.edu

[1] Department of Psychology, Eastern Washington University, Cheney, WA 99004, USA

[2] Department of Psychology, Colorado State University, Fort Collins, CO, USA

[3] Department of Psychology, Williams College, Williamstown, MA, USA

Predicting future test performance

Research in metacognition often investigates those factors that influence monitoring (assessing one's learning), often by asking learners to make a judgment of learning (JOL) anticipating the likelihood that information will be remembered in the future (see Rhodes, 2016, for a review). One notable finding is that individuals frequently predict future memory performance based on their current memory state (Koriat et al., 2004; Nelson & Dunlosky, 1991). This strategy may beget highly accurate predictions (Nelson & Dunlosky, 1991; see Rhodes & Tauber, 2011, for a review) but will falter when the current state of memory is not diagnostic of future memory performance. For example, participants become underconfident in their memory predictions across multiple trials with the same list (i.e., the *underconfidence with practice* effect or UWP; Koriat, Sheffer, & Ma'ayan, 2002). Studies of UWP generally have participants engage in 2-5 study-test cycles in which they study word pairs, make JOLs regarding the likelihood of recalling the word pair on an upcoming test, and then take a test. On the first trial, participants are typically overconfident with JOLs exceeding actual memory performance. Although memory performance increases across trials (due to more study opportunities), the magnitude of participants' predictions either remains stable or increases slightly (leading to underconfidence).

Finn and Metcalfe (2008) proposed that JOLs reflected the recall outcome for that item on the previous test. Items recalled during the trial 1 test were given a high JOL during Trial 2 and items that were not recalled were given a low JOL during Trial 2. Accordingly, Finn and Metcalfe (2007, 2008) suggested that participants rely on *Memory for Past Test* (MPT), judging future memorability based on whether information was recalled during a previous test (see also Serra & Ariel, 2014; for other influences, see Ariel & Dunlosky 2011; Tauber & Rhodes, 2012).

Feedback and memory predictions

Decades of research show that feedback is highly beneficial for memory (Butler, Karpicke, & Roediger, 2008; Kulhavy & Anderson, 1972; Sitzman, Rhodes, & Tauber, 2014; Sitzman, Rhodes, Tauber, & Liceralde, 2015; Skinner, 1954). For example, feedback provides another study opportunity, enhances retention of correct information, identifies information needing additional study, and facilitates error correction. Indeed, in contrast to continued study-test cycles (as in studies of the UWP effect), providing feedback allows the learner to rapidly correct errors, suggesting that feedback offers a salient, diagnostic cue for guiding future behaviors. Accordingly, optimal learning may entail a metacognitive understanding that feedback is beneficial and corrects errors.

Given the myriad benefits of feedback, it surprising how little work has examined the extent to which feedback influences memory predictions. Kornell and Rhodes (2013) report a notable exception (see also Rhodes & Tauber, 2011, Experiment 3[1]). In their first experiment, participants studied 36 weakly related word pairs and either restudied the pairs or received a test with or without feedback, whereby they were prompted to provide the target when given the cue. Participants receiving feedback were shown the correct cue-target pair after providing a response. Immediately following each trial, participants made a JOL about the future likelihood of remembering the word pair and later completed a final test on all word pairs. The influence of feedback on JOLs was explored via both absolute accuracy (the overall correspondence between JOLs and memory performance) and relative accuracy (the degree to which participants' JOLs distinguished between information that would or would not be remembered).

Consistent with previous research, recall on a final test was superior when participants received feedback (82.4 % correct) than an initial test without feedback (56.9 %). However, average JOLs did not differ between the groups (60.1 % and 60.2 %, respectively), indicating that feedback impaired absolute accuracy. Relative accuracy (measured via gamma correlations) was also hindered by feedback. Overall, participants in the test-only condition (i.e., without feedback) showed stronger gamma correlations between JOLs and final test performance ($G = .85$) than participants in the feedback condition ($G = .55$). Experiments 2 and 3 replicated these data, providing further evidence that, although memory was best when feedback was provided, feedback also hindered absolute and relative accuracy.

One potential explanation for these findings is participants' judgments reflect the most recent test performance and largely neglect information gained from feedback. Therefore, although feedback helps learners correct errors, JOLs reflect the current state of that item (i.e., it was or was not remembered) and are insensitive to improvements in performance from feedback. We tested this account in each of the experiments reported by examining JOLs as a function of prior retrieval success. If a stability bias guided by prior retrieval success is present, then JOLs should be strongly and positively related to the probability that a target item was retrieved on the initial test. Likewise, enhancing metacognitive accuracy may involve encouraging participants to incorporate additional cues (besides memory for past test) into judgments.

[1] We note that other studies have explored the potential benefit of feedback for JOLs (e.g., Carpenter & Olson, 2012; Logan, Castel, Haber, & Viehman, 2012). However, these studies do not systemically compare feedback to a condition without feedback and thus cannot directly address the influence of feedback on JOLs.

The current experiments

In the experiments reported, we investigated (a) whether JOLs following feedback rely on memory for a past test (MPT) and (b) methods that may improve both absolute and relative metacognitive accuracy following feedback. Experiment 1 used a between-subjects design in an attempt to replicate Kornell and Rhodes' (2013) results. Similar to Kornell and Rhodes, it was expected that item-by-item judgments would be less accurate following correct answer feedback than no feedback. To anticipate, Experiment 1 revealed that the magnitude of participants' JOLs was sensitive to feedback, while item-by-item discrimination (i.e., relative accuracy) remained impaired by feedback relative to no feedback. Thus, subsequent experiments examined whether relative accuracy could be improved via a within-subjects design (Experiment 2), by prior experience (Experiment 3), or by making more diagnostic cues available (Experiment 4). In each experiment, we also tested an MPT account by examining the correlation between performance on the initial test and the subsequent JOL provided.

Experiment 1

Experiment 1 primarily sought to replicate Kornell and Rhodes' findings (2013). Participants studied Lithuanian-English word pairs and were asked to provide the English translation when shown the Lithuanian word on a later test. Some participants were provided correct answer feedback on the initial test whereas others were not provided feedback regarding their performance. Following this, participants rated the likelihood of remembering the correct English translation on a later test.

Methods

Participants

Eighty-four Colorado State University undergraduate students participated for partial course credit.

Materials

Materials consisted of 34 easy to moderate Lithuanian- English Word pairs taken from normative data reported by Grimaldi, Pyc, and Rawson (2010). Word pairs used in the current experiment were recalled by 19–56 % ($M = 30.87$, $SD = 11.62$) of participants during an initial trial in Grimaldi et al.'s study.

Procedure

Participants studied a list of 34 word pairs, with the first two and last two pairs serving as primacy and recency buffers that were not included in analyses. Each word pair was presented for 4 s. After a 500-ms interstimulus interval, the next word pair appeared. Participants studied the entire list twice. Following a 5-min math distractor task, participants took an initial test on the materials. They were shown the Lithuanian word for 10 s and asked to recall the English translation aloud to an experimenter who coded the response. Half of the participants received correct answer feedback for each item, with both the Lithuanian word and English translation appearing on the screen for 5 s. For the other half of participants, no feedback was provided after each item. In order to equate time across items, "please wait for the next screen" was displayed on the computer screen for 5 s. The order of presentation for items was uniquely randomized for each participant. After receiving feedback or waiting, participants made a JOL indicating the likelihood that, when shown the Lithuanian word, they could recall the English translation on a final test. Judgments were made on a scale of 0–100 %, with 0 % indicating no likelihood of recalling the English translation and 100 % indicating an absolute likelihood of recalling the English translation. Participants were encouraged to use the entire range of the scale.

Following another 5-min math distractor task, participants were administered a final test. They were once again shown the Lithuanian word and asked to recall the English translation. Participants were given as much time as needed to provide a response. Next, they were asked to judge their confidence in the accuracy of their response on a scale of 0–100 %, 0 % being not at all confident they are correct with 100 % being completely confident they are correct. No feedback was provided on the final test. Because confidence judgments were tangential to the focus of our experiments, they are not reported.[2]

Results

In the following analyses, we first examine performance on the initial and final tests. We then report our focal analyses examining the relationship between JOLs and performance on the final test. The alpha level was set at .05 for all analyses reported.

Test performance

On test 1, correct recall did not differ between participants in the feedback and no-feedback conditions, $t<1$ (see Table 1). However, participants in the feedback condition correctly

[2] Gamma correlations indicated that accuracy on test 2 was strongly related to confidence for both the feedback condition ($G = .95$, $SE = .02$), $t(40) = 49.58$, $p < .001$ and the no-feedback condition ($G = .98$, $SE = .04$), $t(39) = 146.41$, $p < .001$; this correlation did not differ between conditions, $t(79) = 1.10$, $p = .28$. The same pattern holds true across all experiments (full analyses can be obtained by contacting the first author).

Table 1 Percentage of items correctly recalled across experiments

	No feedback				Correct answer feedback				Right/wrong feedback			
	Test 1	Test 2	Retained	Corrected	Test 1	Test 2	Retained	Corrected	Test 1	Test 2	Retained	Corrected
Experiment 1	25.56 (2.29)	27.46 (2.43)	88.24 (2.82)	6.43 (1.03)	25.08 (3.01)	49.37 (3.77)	92.69 (1.55)	38.72 (3.51)				
Experiment 2	32.40 (3.10)	33.57 (3.06)	95.24 (1.80)	4.99 (1.45)	31.90 (3.00)	63.10 (3.24)	92.86 (3.60)	50.21 (2.10)	33.10 (3.30)	35.71 (3.10)	92.73 (3.30)	6.38 (1.50)
Experiment 3												
List 1	31.75 (2.58)	33.33 (2.75)	89.87 (2.21)	6.93 (1.68)	32.86 (3.21)	67.14 (3.43)	91.67 (3.25)	59.55 (3.74)				
List 2	44.13 (3.66)	46.83 (3.78)	95.02 (1.48)	9.67 (2.46)	47.14 (3.81)	74.92 (3.22)	96.48 (1.51)	61.34 (3.89)				
Experiment 4					33.45 (2.65)	55.12 (2.91)	92.98 (1.63)	40.14 (2.53)				

Numbers in parentheses represent the standard error of the mean

more words on test 2 compared with participants in the no-feedback condition, $t(82) = 4.88$, $p < .001$, $d = 1.07$.

Performance on test 2 was also conditionalized based on accuracy during test 1 to examine the percentage of correct responses retained from test 1 to test 2 and the percentage of errors corrected from test 1 to test 2. There was no reliable difference in the percentage of correct responses retained from test 1 to test 2 for the feedback relative to the no-feedback condition, $t(81) = 1.34$, $p = .17$, $d = .29$. However, participants in the feedback condition corrected more errors from test 1 on test 2 than participants in the no-feedback condition, $t(82) = 8.82$, $p < .001$, $d = 1.92$.

Memory predictions

Absolute accuracy Overall, participants receiving feedback provided higher JOLs ($M = 37.97$, $SE = 3.44$) than participants who did not receive feedback ($M = 21.52$, $SE = 2.32$), $t(82) = 3.96$, $p < .001$, $d = .86$. JOLs were also conditionalized based on test 1 accuracy. For correct items, participants in the feedback condition provided higher average JOLs ($M = 81.32$, $SE = 3.22$) than participants in the no-feedback condition ($M = 62.47$, $SE = 47$), $t(81) = 3.47$, $p = .001$, $d = .76$. Similarly, participants in the feedback condition provided higher average JOLs after errors ($M = 25.33$, $SE = 3.19$) than participants in the no-feedback condition ($M = 7.58$, $SE = 1.46$), $t(82) = 5.06$, $p < .001$, $d = 1.10$.

Relative accuracy For each feedback condition, gamma correlations were computed between JOLs during test 1 and recall on test 2 (see Table 2). Positive correlations indicate that participants gave higher JOLs to items that were correct on the final test.

Correlations reliably differed from zero for participants in the feedback, $t(40) = 23.10$, $p < .001$, and no feedback $t(39) = 51.52$, $p < .001$, conditions. Overall, the relationship between JOLs and final test accuracy was stronger for participants in the no-feedback condition than participants in the feedback condition, $t(79) = -5.24$, $p < .001$, $d = 1.16$. There was also a moderate, positive correlation between JOLs for items answered incorrectly on test 1 and accuracy on test 2 for the feedback condition, ($G = .40$, $SE = .08$), $t(38) = 5.28$, $p < .001$. Thus, participants differentiated to some degree between errors that would be corrected following feedback and more persistent errors.

We also explored whether the MPT heuristic played a role in JOLs on the initial test by calculating gamma correlations between accuracy of an item on test 1 and the JOLs provided on test 1. In both the feedback condition ($G = .98$, $SE = .02$), $t(40) = 63.03$, $p < .001$, and no-feedback condition ($G = .97$, $SE = .01$), $t(40) = 134.96$, $p < .001$, there was a strong correlation that did not reliably differ between feedback conditions, $t < 1$. Therefore, regardless of whether feedback was provided, participants' JOLs aligned almost perfectly with the outcome of the initial test.

Discussion

In contrast to Kornell and Rhodes (2013), JOLs in the feedback condition in Experiment 1 were higher than JOLs in the no-feedback condition for both correct and incorrect responses. After receiving feedback, participants predicted, on average, that they were more likely to retain correct responses on a later

Table 2 Mean gamma correlations between test 1 judgments of learning (JOLs) and test 2 accuracy across experiments

	Gamma		
	No feedback	Correct answer feedback	Right/wrong feedback
Experiment 1	.90 (.02)	.71 (.03)	
Experiment 2	.95 (.02)	.57 (.08)	.92 (.06)
Experiment 3			
List 1	.87 (.04)	.65 (.06)	
List 2	.94 (.05)	.69 (.05)	
Experiment 4		.61 (.03)	

Parentheses represent standard error of the mean

test and more likely to correct errors. However, their predictions of error correction (25 %) underestimated the percentage of errors corrected on test 2 (39 %), $t(41) = 3.95$, $p < .001$. $d = .61$. It is unclear why these results differ from Kornell and Rhodes. In addition to differences in materials and the mode of administration (Kornell and Rhodes primarily tested participants via MTurk), participants in our experiment exhibited much lower levels of performance. Specifically, whereas participants in Kornell and Rhodes' initial experiment correctly recalled approximately 57 % of studied items on test 1, our participants only recalled approximately 26 % of the correct English translations following two study opportunities. This increased difficulty may have influenced the basis for JOLs, an issue we address in the General Discussion.

Although the magnitude of JOLs differed from Kornell and Rhodes (2013), our results for relative accuracy comport with their findings. Participants in the no-feedback condition were highly accurate when differentiating between items they would answer correctly on test 2 and items they would answer incorrectly. However, participants in the feedback condition had more difficulty differentiating between errors that would be corrected on a later test and errors that would persist on a later test. For both feedback conditions, participants' predictions of future test performance were strongly related to past test performance.

Experiment 2

One possible reason why participants did not fully anticipate the impact of feedback in Experiment 1 is that feedback was manipulated between-subjects and thus not salient. Therefore, Experiment 2 employed a within-subjects design. Previous research has demonstrated that within-subjects manipulations can draw participants' attention to the variable being manipulated and increase sensitivity to that variable (e.g., Koriat et al., 2004; Magreehan, Serra, Schwartz, & Narciss, 2015). For example, participants who made judgments about their future memory were insensitive to the retention interval in a between-subjects design but became sensitive to it (and predicted more forgetting for a longer retention interval) when retention interval was manipulated within-subjects (Koriat et al., 2004; but see Kornell & Bjork, 2009). However, there is a distinct lack of research examining the effect of making a variable salient on *relative* judgments, and one goal of Experiment 2 was to fill this void. Accordingly, we predicted that when feedback was salient, participants might better focus on the impact of feedback especially following errors, and be more aware that they would likely benefit from the feedback, potentially increasing the relative accuracy of their judgments.

In addition, Kornell and Rhodes (2013) compared correct answer feedback to a no-feedback and a restudy condition. However, these are just two among many varieties of feedback. For example, right/wrong feedback (e.g., Pashler et al., 2005) indicates whether an answer is correct, but does not provide the information necessary to update memory. Previous work by Rhodes and Tauber (2011, Experiment 3) demonstrated that participants given right/wrong feedback prior to a JOL were highly accurate at discriminating between items they would or would not later remember compared to when feedback was not provided. This may suggest that not all forms of feedback impair metacognition. Thus, in Experiment 2, participants received correct answer feedback, right/wrong feedback, and no feedback.

Methods

Participants

Forty-two Colorado State University undergraduate students participated for partial course credit.

Materials and procedures

Experiment 2 was identical to Experiment 1, with two exceptions. In addition to the correct answer feedback and no-feedback conditions, we also included a right/wrong feedback condition. Second, unlike Experiment 1, feedback conditions were manipulated within-subjects rather than between-subjects. Thus, for one-third of the items, participants received correct answer feedback; both the Lithuanian word and the English translation appeared on the screen for 5 s. For another third of the items, participants received right/wrong feedback; the Lithuanian word appeared on the screen with either "Correct" or "Incorrect" displayed below for 5 s to indicate whether or not the participant answered correctly. For the remaining third of the items, no feedback was presented using the procedure implemented in Experiment 1. The feedback condition for an individual item was counterbalanced across participants and the order of item presentation was randomized anew for each participant.

Results

Test performance

The percentage of words correctly recalled on test 1 did not differ between feedback conditions, $F < 1$ (see Table 1). However, on test 2 (see Fig. 1), there was a difference among feedback conditions, $F(2, 82) = 45.37$, $p < .001$, $\eta^2_p = .53$. Items that received correct answer feedback were more likely to be remembered on test 2 than items receiving right/wrong feedback, $t(41) = 7.56$, $p < .001$, $d = 1.33$, or items not receiving feedback, $t(41) = 8.54$, $p < .001$, $d = 1.44$. There was no difference in test 2 performance between items receiving right/wrong feedback and items not receiving feedback, $t < 1$.

Performance on test 2 was also conditionalized based on accuracy during test 1 to examine the percentage of correct

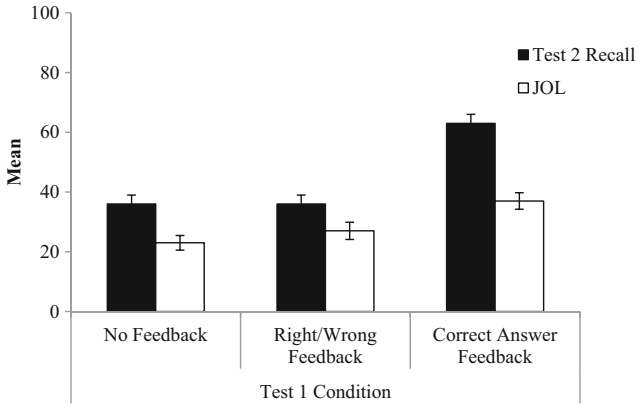

Fig. 1 Correct recall on test 2 and average judgments of learning (JOL) provided during test 1 for Experiment 2. Error bars represent one standard error of the mean

responses retained from test 1 to test 2 and the percentage of errors corrected from test 1 to test 2. A one-way ANOVA demonstrated that, across feedback conditions, there were no differences in the percentage of correct responses retained from test 1 to test 2, F<1 (see Table 1). In contrast, error correction differed across feedback conditions, $F(2,82)$ = 137.44, p< .001, η^2_p= .77. Participants corrected a greater percentage of errors when items were given correct answer feedback compared with items given right/wrong feedback, $t(41)$ = 12.78, p< .001, d= 2.58, and no feedback, $t(41)$ = 11.86, p< .001, d = 2.76. Error correction did not differ between the right/wrong feedback and no-feedback conditions, t<1.

Memory predictions

Absolute accuracy Average JOLs (see Fig. 1) given to items on test 1 differed by feedback type, $F(2, 82) = 14.30$, p<.001, η^2_p=.26. Specifically, participants' average JOLs were higher for items given correct answer feedback compared with items given right/wrong feedback, $t(41)$ =3.47, p< .001, d = .56, or no feedback, $t(41)$ = 6.24, p > .001, d = .82. JOLs did not differ between items in the right/wrong and no-feedback condition, $t(41)$ = 1.31, p = .20, d = .22.

Further analyses showed that, for correct responses, average JOLs differed based on feedback type, $F(2,66)$ = 15.64, p< .001, η^2_p = .32. Average JOLs given to items in the no-feedback condition (M = 59.60, SE = 3.84) were reliably lower than average JOLs for items in the correct answer feedback condition (M = 73.34, SE = 3.93), $t(36)$ = 4.43, p< .001, d = .57, and items in the right/wrong feedback condition (M = 72.90, SE = 4.05), $t(34)$ = 4.21, p< .001, d = .58. JOLs did not differ for items in the correct answer feedback or right/wrong feedback conditions, t< 1.

Average JOLs also differed as a function of feedback for errors on test 1, $F(2,82)$ = 48.55, p < .001, η^2_p = .54. Participants predicted that they were more likely to correct errors on a later test for items receiving correct answer feedback (M = 20.75, SE = 2.40) compared with items receiving right/wrong feedback (M = 3.56, SE = .93), $t(41)$ = 7.36, p <.001, d = 1.38, and no feedback (M = 4.89, SE = 1.11), $t(41)$ = 7.54, p < .001, d = 1.18. However, JOLs did not differ between items in the right/wrong feedback condition and no-feedback condition, $t(41)$ = 1.12, p = .27, d = .20.

Relative accuracy All mean gamma correlations were reliably greater than zero, ts ≥ 7.55, p< .001 (see Table 2) but differed by feedback type, $F(2,68) = 18.29$, p < .001, η^2_p = .35. The correlation between JOLs and final test accuracy was reliably lower for items in the correct answer feedback condition compared with items in the right/wrong feedback condition, $t(34)$ = 5.18, p< .001 d = .85, and the no-feedback condition, $t(34)$ = 4.87, p< .001, d = 1.15. Relative accuracy did not differ between the right/wrong and no-feedback conditions, t< 1. There was also no reliable correlation between JOLs for items in the correct answer feedback condition that were answered incorrectly on test 1 and accuracy on test 2, (G = .15, SE = 12), $t(35)$ = 1.33, p = .19.

As in Experiment 1, we explored an MPT account by calculating gamma correlations between accuracy on test 1 and JOLs on test 1 for items in the no feedback (G = .97, SE = .01), $t(33)$ = 109.54, p < .001, correct answer feedback (G = .93, SE = .03), $t(33)$ = 37.08, p < .001, and right/wrong feedback (G = .99, SE = .01), $t(33)$ = 123.54, p < .001, conditions. A repeated-measures ANOVA revealed a reliable difference among the feedback conditions, $F(2,66)$ = 4.21, p = .02, η^2_p = .11. While there was no difference between the no-feedback condition and correct answer feedback condition, $t(33)$ = 1.60, p = .12, d = .39, or between the no-feedback condition and the right/wrong feedback condition, $t(33)$ = -1.54, p =.13, d = .39, the gamma correlation was greater in the right/wrong feedback condition than the correct answer feedback condition, $t(33)$ = -2.62, p = .01, d = .62.[3] Given that gamma correlations were near ceiling, we suggest that this difference should be interpreted cautiously. Indeed, the most compelling element of these data is how consistently the outcome of test 1 predicts JOLs for test 2, regardless of the particular form or presence of feedback.

Discussion

Experiment 2 replicated Experiment 1 using a within-subjects design. As expected, correct answer feedback enhanced performance on test 2 compared with right/wrong feedback and no feedback. Participants anticipated this pattern, with items in the correct answer feedback condition given higher JOLs than items in the right/wrong and no-feedback conditions.

[3] Participants with invariant judgments or memory outcomes were not included in the analysis of gamma correlations, reflected in the degrees of freedom reported.

However, participants' predictions regarding the percentage of errors they would correct (20 %) reliably underestimated the percentage of errors actually corrected on test 2 (50 %), $t(41) = 7.40$, $p < .001$, $d = 1.57$. Relative accuracy was consistent with Experiment 1. That is, item-by-item judgments were less accurate in the correct answer feedback condition compared with the no feedback and right/wrong conditions, with gamma correlations in the right/wrong and no-feedback conditions near unity (+1.0).

Experiment 3

Experiments 1 and 2 demonstrated that participants' JOLs were highest following correct answer feedback. They still underestimated the power of feedback, however, and for this reason, receiving feedback made them less accurate when predicting which specific items would be remembered or forgotten. Previous work has demonstrated that task experience can enhance absolute metacognitive accuracy on a list of new items (e.g., Rhodes & Castel, 2008; Tauber & Rhodes, 2010). However, it is unclear whether knowledge gained through practice will also improve relative accuracy.

In Experiment 3, participants studied and were tested on two different lists. If participants do not appreciate the magnitude of improvement that feedback provides because they lack experience, then providing practice during a first trial may allow them to adjust JOLs during a subsequent trial with a second list. We note that this design differs from previous experiments on UWP (e.g., Koriat et al., 2002; Tauber & Rhodes, 2012) because participants were given a new list on the second trial rather than studying and testing on the same list. Thus, the potential for updating knowledge of feedback can be distinguished from prior study and test experiences with a specific item or multiple test experiences with an item (Carpenter & Olson, 2012).

Accordingly, in Experiment 3, we examined whether prior experience with a test-feedback-test cycle would allow participants to incorporate this information on a second test-feedback-test cycle with new information. Such an improvement with practice would be evident in higher JOLs for errors on the initial test and enhanced relative accuracy on the final test for the second cycle.

Methods

Participants

Forty-two Colorado State University undergraduate students completed the experiment for partial course credit.

Materials and procedures

The materials and procedures were similar to Experiment 2; however, the right/wrong feedback condition was removed. During List 1, participants studied 30 Lithuanian-English word pairs, with order randomized anew for each participant. On an initial test they were shown the Lithuanian word and asked to provide the English translation. Following this, they received correct answer feedback for half of the items or waited for the next screen before making their JOL for the remaining items. After a short 5-min distractor task, participants completed a final test. During List 2, participants completed the same process as List 1 with a new list of 30 Lithuanian-English word pairs. Items were assigned to list 1 and list 2 and the presentation order of lists was counterbalanced across participants.

Results

Test performance

Memory performance for List 1 and List 2 was examined via a 2 (List: List 1, List 2) × 2 (Test: Test 1, Test 2) × 2 (Feedback type: Feedback, No Feedback) repeated-measures ANOVA. Overall, participants correctly recalled a greater percentage of English translations during List 2 ($M = 53.30$, $SE = 3.30$) than List 1 ($M = 41.30$, $SE = 2.60$), $F(1,41) = 22.78$, $p < .001$, $\eta^2_p = .36$. Participants also correctly recalled a greater percentage of English translations on test 2 ($M = 55.60$, $SE = 2.70$) compared with test 1 ($M = 39.00$, $SE = 2.70$), $F(1,41) = 330.61$, $p < .001$, $\eta^2_p = .89$. Lastly, recall was superior for the feedback condition ($M = 55.50$, $SE = 3.00$) compared to the no-feedback condition ($M = 39.00$, $SE = 2.70$), $F(1,41) = 72.33$, $p < .001$, $\eta^2_p = .64$.

The interaction between Test and Feedback type was reliable, $F(1, 41) = 287.85$, $p < .001$, $\eta^2_p = .88$. Specifically, on test 1, recall did not differ between the feedback ($M = 40.00$, $SE = 3.13$) and no feedback ($M = 37.94$, $SE = 2.68$) conditions, $t < 1$. On test 2, items in the feedback condition were more likely to be recalled ($M = 71.03$, $SE = 3.06$) than items in the no-feedback condition ($M = 40.10$, $SE = 2.81$), $t(41) = 14.56$, $p < .001$, $d = 1.62$. Lastly, there was a 3-way interaction among List, Test, and Feedback type, $F(1,41) = 6.30$, $p = .02$, $\eta^2_p = .13$.[4]

[4] In the interest of brevity, we do not report a full breakdown of this three-way interaction (those analyses are available by contacting the first author). In general, the pattern of results is the same across lists although the magnitude of effects changes across lists, accounting for the interaction. For example, for both List 1, ($F(1, 41) = 251.94$, $p < .001$, $\eta^2_p = .86$, and List 2, $F(1,41) = 166.38$, $p < .001$, $\eta^2_p = .80$, the percentage of correct responses increased from test 1 to test 2, although the effect was slightly larger for List 1. As well, correct recall was greater following feedback than no feedback for List 1, $F(1, 41) = 51.49$, $p < .001$, $\eta^2_p = .56$, and List 2, $F(1,41) = 33.50$, $p < .001$, $\eta^2_p = .45$, but the effect was greater for List 1.

Memory predictions: absolute accuracy

List 1 and list 2 Did JOLs (Fig. 2) change as a function of practice? A 2 (List: List 1, List 2) × 2 (Feedback Type: Feedback, No Feedback) × 2 (Accuracy: Correct, Incorrect) repeated-measures ANOVA indicated that average JOLs increased from list 1 (M = 52.00, SE = 2.23) to list 2 (M = 54.94, SE = 1.99), $F(1,38) = 4.79$, $p = .04$, $\eta^2_p = .11$. Average JOLs were also greater for items in the feedback condition (M = 61.70, SE = 2.38) compared with items in the no-feedback condition (M = 45.25, SE = 2.05), $F(1,38) = 74.07$, $p < .001$, $\eta^2_p = .66$. Lastly, items answered correctly garnered higher average JOLs (M = 78.91, SE = 2.29) than items answered incorrectly (M = 28.04, SE = 2.09), $F(1,38) = 834.44$, $p < .001$, $\eta^2_p = .96$.

A reliable List × Accuracy interaction was present, $F(1,38) = 6.24$, $p = .02$, $\eta^2_p = .14$. Specifically, whereas average JOLs were similar for incorrect responses for list 1 (M = 27.97, SE = 2.31) and list 2 (M = 28.10, SE = 2.10), $t < 1$, average JOLs for correct responses increased from list 1 (M = 76.04, SE = 2.56) to list 2 (M = 81.79, SE = 2.44), $t(38) = -2.835$, $p < .01$, $d = .37$. Feedback Type also reliably interacted with Accuracy. For correct items, participants' JOLs were greater in the feedback condition (M = 84.15, SE = 2.30) relative to the no-feedback condition (M= 73.67, SE = 2.59), $t(38) = 6.05$, $p < .001$, d = .68. JOLs were also reliably greater for incorrect items in the feedback condition (M = 39.25, SE = 2.77) compared with the no-feedback condition (M= 16.83, SE = 2.25), $t(38) = 7.91$, $p < .001$, d = 1.41. No other interactions were reliable. Thus, across both lists, participants provided higher JOLs for items in the feedback condition compared with the no-feedback condition. However, practice did not change JOLs for incorrect items in the feedback condition.

Memory predictions: relative accuracy

Gamma correlations between JOLs during test 1 and accuracy on test 2 were compared in a 2 (List: List 1, List 2) × 2 (Feedback Type: Feedback, No Feedback) repeated-measures ANOVA. Overall, gamma did not reliably differ between list 1 (M = .76, SE = .04) and list 2 (M = .82, SE = .03), $F(1, 30) = 2.08$, $p = .16$, $\eta^2_p = .07$, and List did not interact with Feedback Type ($F < 1$), suggesting no benefit of practice. As in the previous experiments, gamma correlations were stronger for items in the no feedback (M = .91, SE = .02) compared with the feedback (M = .67, SE = .04) condition, $F(1,30) = 27.94$, $p < .001$, $\eta^2_p = .48$. For items in the feedback condition, gamma correlations were computed between JOLs for errors on test 1 and accuracy on test 2 for list 1 and list 2. On list 1, (G = .49, SE = .07), $t(36) = 6.78$, $p < .001$, and list 2 (G = .22, SE = .10), $t(33) = 2.21$, $p = .03$, there was reliable correlation between JOLs for errors on test 1 and accuracy on test 2; however, this correlation did not differ between the two lists, $t < 1$.

For list 1, there were strong correlations between accuracy on test 1 and JOLs on test 1 for items in the feedback (G = .92, SE = .02), $t(38) = 51.48$, $p < .001$, and no-feedback conditions (G = .95, SE = .03), $t(38) = 32.04$, $p < .001$. The same pattern was evident on list 2 for the feedback (G = .94, SE = .02), $t(38) = 58.64$, $p < .001$, and no-feedback conditions (G = .98, SE = .01), $t(38) = 121.43$, $p < .001$. Thus, participants' JOLs were predicted by test 1 outcomes across feedback conditions and lists.

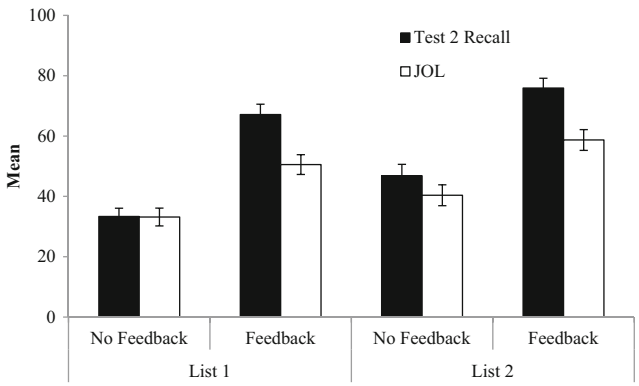

Fig. 2 Recall on test 2 compared with average judgments of learning (JOL) provided during List 1 and List 2 for Experiment 3. Error bars represent one standard error of the mean

Discussion

Experiment 3 replicated Experiments 1 and 2. Participants predicted that, on a final test, they were more likely to correctly remember items accompanied by feedback than items not receiving feedback. However, even following practice with a different list, participants' JOLs (37 % and 39 %, list 1 and list 2, respectively) underestimated the likelihood of improvement for correcting errors when given feedback (60 %, $t(41)$ 5.12, p < .001, d = .99, and 61 %, $t(41) = 5.35$, $p < .001$, d = 1.03, list 1 and list 2, respectively). Practice also had no impact on relative accuracy, as the magnitude of gamma correlations did not improve from List 1 to List 2 (cf. Koriat & Bjork, 2006). Thus, prior experience with feedback did not increase JOLs for incorrect items or overall enhance relative accuracy.

Experiment 4

As the prior experiments show, correct answer feedback enhances memory performance but hinders relative accuracy. The source of poor relative accuracy appears to be participants' predictions following errors. Across experiments, participants have generally predicted a low likelihood that errors will be corrected on a later test (i.e., JOLs have averaged about 30 % for feedback items and about 10 % for no feedback items). Instead, JOLs appear strongly tied to the outcome of initial test performance, regardless of whether feedback is provided. Although using past test performance is highly predictive of later test performance in the no-feedback condition,

it is not entirely diagnostic of error correction that will result from feedback. Indeed, in Experiments 1–3, approximately 50 % of errors on test 1 were corrected on test 2 when participants were given feedback. Thus, to appreciate the value of feedback, participants may need cues that disclose the likelihood that an error will be corrected on a later test.

In a series of experiments, Finn and Metcalfe (2010) had participants answer general knowledge questions and, after an error, receive one of four types of feedback. Of central interest to Experiment 4, some of their participants were given scaffolded feedback. Specifically, participants were initially shown the first letter of the correct response and then asked to generate the correct answer. If the correct answer was not generated, participants were shown the second letter and again attempted to generate the correct response. This procedure continued until participants either generated the correct response or the entire correct answer was displayed.

Finn and Metcalfe (2010) reported that scaffolded feedback and correct answer feedback led to comparable performance on an immediate test, but scaffolded feedback enhanced performance on delayed tests. In addition, participants were more likely to remember the correct response on a final test if they could generate the correct response before the entire word was displayed. Specifically, if participants required few cues (one or two letters) before generating a correct response, they were more likely to remember that correct response on a final test than if they needed the majority of the cues to be presented to generate the answer. This may reflect the benefits of effortful retrieval practice (e.g., Carpenter & DeLosh, 2006) or different levels of initial learning for the item. Regardless, these data suggest that the number of letters needed to generate a target is inversely related to error correction.

The goal of Experiment 4 was to determine whether the number of letters needed to generate the correct response would provide a salient cue for participants to determine which errors they would or would not correct, thus enhancing relative accuracy compared with previous experiments. If participants were sensitive to the diagnosticity of this cue, then they should provide higher JOLs to items requiring one- to two-letter cues to generate the correct response versus items requiring more cues or the entire word. By extension, their JOLs should be less likely to rely on their memory for the past test.

Methods

Participants

Sixty Colorado State University undergraduate students completed the experiment for partial course credit.

Materials and procedures

The materials and procedures were identical to Experiment 2 except for the feedback procedure during the initial test. Specifically, feedback was presented in a scaffolded manner similar to Finn and Metcalfe (2010). After a correct response, participants were shown the correct English translation for 5 s before advancing to the next item. After an incorrect response, participants were shown the first letter of the English translation and asked to recall the correct response. If they recalled correctly, the correct English translation was displayed for 5 s. If they could not recall the correct English translation, the second letter of the answer was displayed. Again, they were asked to recall the correct response. This cycle continued until the participant either recalled the correct response or the entire English word was displayed. Unlike previous experiments, the no-feedback condition was removed and feedback was scaffolded for all incorrect items during the initial test. After the initial test, the experiment continued in the same way as Experiments 1 and 2; participants completed a distractor task and then the final test.

Results

Test performance

Overall, participants correctly recalled more English translations on test 2 compared with test 1, $t(59) = 18.75$, $p < .001$, $d = .98$ (see Table 1).

Final test performance as a function of number of feedback cues needed A one-way ANOVA indicated that the percentage of errors corrected (Fig. 3) differed based on the number of letter cues needed to generate the correct response, $F(2,118) = 6.60$, $p < .01$, $\eta^2_p = .10$. When participants generated the correct response after one-letter cue, they corrected a greater percentage of errors ($M = 50.77$, $SE = 4.16$) than if they required two-letter cues, ($M = 38.68$, $SE = 3.85$), $t(59) = 2.67$, $p = .01$, $d = .39$, or three- or more letter cues ($M = 37.02$, $SE = 2.77$), $t(59) = 3.41$, $p < .001$, $d = .48$, to generate the correct response. The percentage of errors corrected did not differ if participants required two-letter cues versus three- or more letter cues to generate the correct response, $t<1$.

[0] Gamma correlations were also calculated between the number of letters cues required to generate the correct response and JOLs for items answered incorrectly on test 1. There was a strong negative correlation ($G = -.65$, $SE = .03$), $t(59) = -21.40$, $p < .001$. Thus, participants provided higher JOLs following answers generated after fewer letter cues. For errors on test 1, a gamma correlation was also calculated between the number of letter cues required to generate the correct response and accuracy on test 2. There was a small, but reliable, negative correlation ($G = -.16$, $SE = .05$), $t(58) = -3.28$, $p = .002$, indicating that participants were more likely to correct errors on test 2 if they could generate the correct response following fewer letter cues on test 1.

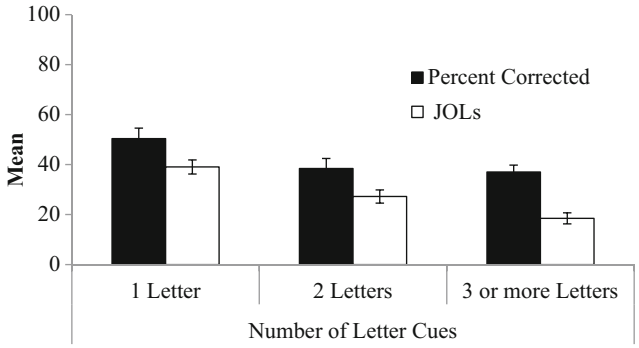

Fig. 3 Average judgments of learning (JOLs) and the percentage of errors corrected based on the number of letter cues needed during feedback for Experiment 4. Error bars represent the standard error of the mean

Memory predictions

A one-way ANOVA indicated that JOLs (Fig. 3) differed based on the number of letter cues needed to generate the correct answer, $F(2,116) = 96.88$, $p < .001$, $\eta^2_p = .63$. When participants could generate the correct response after one-letter cue ($M = 38.99$, $SE = 2.70$), they provided reliably greater JOLs than when they generated the correct response after two-letter cues ($M = 27.55$, $SE = 2.58$), $t(58) = 8.35$, $p < .001$, $d = .56$, and after three- or more letter cues ($M = 18.48$, $SE = 2.20$), $t(58) = 12.05$, $p < .001$, $d = 1.09$. On average, JOLs provided after two-letter cues were reliably greater than JOLs after three- or more letter cues, $t(58) = 6.83$, $p < .001$, $d = .48$.

Relative accuracy Overall, there was a reliable, positive correlation between participants' JOLs during the initial test and the accuracy of items on the final test ($G = .61$, $SE = .03$), $t(58) = 19.76$, $p < .001$.[5] The gamma correlation for Experiment 4 was also compared to the correlations from Experiments 1–3. If scaffolded feedback allows participants to better discern which items will be corrected on a later test, then the gamma correlation for Experiment 4 should be reliably greater than the previous experiments. However, gamma correlations between JOLs on the initial test and accuracy on the final test did not reliably differ from the correct answer feedback conditions from the previous experiments, $F(4, 189) = 1.62$, $p = .17$, $\eta^2_p = .03$ (see Table 2). For errors on test 1, a gamma correlation was also calculated between JOLs on test 1 and accuracy on test 2. Overall, there was a reliable correlation, ($G = .27$, $SE = .05$), t(57) -5.44, $p < .001$. This gamma correlation for errors on test 1 did not differ from the previous experiments, $F < 1$. Similar to previous experiments, there was a strong correlation between test 1 accuracy and JOLs ($G = .96$, $SE = .01$), $t(57) = 72.08$, $p < .001$. Thus, even with additional cues available, JOLs remained strongly related to the outcome of the initial retrieval attempt.

[5] Gamma correlations were also calculated between the number of letters cues required to generate the correct response and JOLs for items answered incorrectly on test 1. There was a strong negative correlation ($G = -.65$, $SE = .03$), $t(59) = -21.40$, $p < .001$. Thus, participants provided higher JOLs following answers generated after fewer letter cues. For errors on test 1, a gamma correlation was also calculated between the number of letter cues required to generate the correct response and accuracy on test 2. There was a small, but reliable, negative correlation ($G = -.16$, $SE = .05$), $t(58) = -3.28$, $p = .002$, indicating that participants were more likely to correct errors on test 2 if they could generate the correct response following fewer letter cues on test 1.

Discussion

Consistent with Finn and Metcalfe (2010), participants were more likely to correct errors when they were able to generate the correct response after one cue compared with items that required multiple letter cues shown before the correct response was generated. Average JOLs aligned with this pattern: Participants provided greater JOLs for items generated after few letter cues and lower JOLs for items that needed multiple cues before the correct answer was generated. In fact, participants were hypersensitive to the number of cues needed to generate the answer, which appears to have affected the magnitude of JOLs more than it affected learning. However, this high level of sensitivity did not lead to improvements in relative accuracy. Previous experiments have shown gamma correlations between JOLs and later test performance to be near perfect (close to + 1.0) when feedback was not provided after a test. The gamma correlation in Experiment 4 was still diminished (+ .61) and did not differ from the gammas reported in the previous three experiments following correct answer feedback. In addition, participants were just as likely to rely on the final product of an initial retrieval attempt as in the prior experiments.

General discussion

Previous research (Kornell & Rhodes, 2013) suggests that people may not entirely appreciate the benefits of feedback for memory performance. Providing participants with feedback about the accuracy of their answer on a test vastly improves learning, compared to withholding feedback, but also hinders participants' ability to differentiate between information they will and will not later remember. The current experiments examined methods that may enhance metacognitive accuracy following feedback and tested an account of metacognitive inaccuracy in JOLs accompanied by feedback, focusing on memory for past test performance.

In Experiments 1–3, correct answer feedback was a boon to memory relative to no feedback ($d = 1.42$) and JOLs reflected this benefit ($d = 0.85$). Did participants' conscious knowledge about the benefits of feedback lead to higher JOLs following feedback? Although the current experiments did not address this directly, an experiment similar to Experiment 2 (without the right/wrong feedback condition) was conducted in which

participants were asked to make Pre-Study JOLs (Castel, 2008). Specifically, participants were informed whether an upcoming item would receive feedback and made a JOL before seeing the item. Therefore, the only information available when making a JOL was whether or not the item would receive feedback. Consistent with Experiments 1–3, participants provided greater JOLs for items scheduled to receive feedback ($M = 35.99$, $SE = 3.41$) compared with items not scheduled to receive feedback ($M = 18.71$, $SE = 2.16$), $t(41) = 5.66$, $p < .01$. Thus, participants appear to believe that feedback will enhance performance on a later test.

It remains unclear why participants in Experiment 1 displayed differences in absolute accuracy between the feedback and no-feedback conditions whereas Kornell and Rhodes (2013) reported no difference in judgments. Aside from discrepancies in materials and participant demographics, participants in the current experiments exhibited much lower levels of performance on the initial test despite having two study opportunities. Participants in Experiment 1 may have deemed the Lithuanian-English word pairs difficult to process and thus anchored their JOLs lower on the scale. In turn, this increased task difficulty may have increased participants' sensitivity to factors that improve memory (e.g., feedback). Although the current experiments do not explore this issue, future work may benefit from examining how item difficulty may interact with feedback to influence JOLs.

Despite differences in absolute accuracy, similar to Kornell and Rhodes (2013), the current experiments found deficits in gamma correlations following feedback compared with no feedback (d = -1.06 for Experiments 1–3). Participants provided higher JOLs when they received feedback than when they did not receive feedback but were unable to effectively differentiate between errors they would correct on a later test and those that would not be corrected. These findings held even when participants had a prior experience with the effects of feedback (Experiment 3) or were exposed to conditions where feedback was or was not provided (Experiment 2). Further, item-by-item judgments following feedback remained less accurate than judgments without feedback, even when participants had access to diagnostic cues such as the number of letters necessary to recall the target (Experiment 4).

Why were participants unable to fully account for the memorial benefits, particularly related to error correction, produced by correct answer feedback? Participants likely relied on current test performance (i.e., Finn & Metcalfe, 2007) and made predictions under the assumption that performance would remain stable on a future test (e.g., Kornell & Bjork, 2009). Consistent with this possibility, gamma correlations between accuracy on test 1 and JOLs were strong ($G \geq +.92$) across all experiments. Thus, performance on a previous test was closely associated with subsequent JOLs regardless of whether feedback was provided, suggesting that participants receiving feedback largely relied on their current test performance to predict later test performance. However, on an item-by-item basis, feedback does not provide sufficient information to fully distinguish those items that will or will not be corrected.

Conclusion

The current experiments demonstrated that people understand that feedback is beneficial for memory. Specifically, average JOLs were higher for items in the feedback condition compared with items in the no-feedback condition. However, at an item-by-item level, participants had difficulty predicting which errors would be corrected by feedback. Future work may benefit by further exploring how feedback influences participants' self-regulated learning following a test.

References

Ariel, R., & Dunlosky, J. (2011). The sensitivity of judgment-of-learning resolution to past test performance, new learning, and forgetting. *Memory & Cognition, 39*(1), 171–184.

Butler, A. C., Karpicke, J. D., & Roediger, H. L. (2008). Correcting a metacognitive error: Feedback increases retention of low-confidence correct responses. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34,* 918–928.

Carpenter, S. K., & DeLosh, E. L. (2006). Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. *Memory & Cognition, 34*(2), 268–276.

Carpenter, S. K., & Olson, K. M. (2012). Are pictures good for learning new vocabulary in a foreign language? Only if you think they are not. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38,* 92–101.

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*(2), 429–437.

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*(1), 238.

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

Finn, B., & Metcalfe, J. (2010). Scaffolding feedback to maximize long-term error correction. *Memory & Cognition, 38*(7), 951–961.

Grimaldi, P. J., Pyc, M. A., & Rawson, K. A. (2010). Normative multitrial recall performance, metacognitive judgments, and retrieval latencies for Lithuanian-English paired associates. *Behavior Research Methods, 42*(3), 634–642.

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

Koriat, A., Bjork, R. A., Sheffer, L., & Bar, S. K. (2004). Predicting one's own forgetting: The role of experience-based and theory-based processes. *Journal of Experimental Psychology: General, 133*(4), 643–656.

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

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

Kornell, N., & Rhodes, M. G. (2013). Feedback reduces the metacognitive benefit of tests. *Journal of Experimental Psychology: Applied, 19,* 1–13.

Kornell, N., Rhodes, M. G., Castel, A. D., & Tauber, S. K. (2011). The ease-of-processing heuristic and the stability bias dissociating memory, memory beliefs, and memory judgments. *Psychological Science, 22*(6), 787–794.

Kulhavy, R. W., & Anderson, R. C. (1972). Delayed-retention effect with multiple-choice test. *Journal of Educational Psychology, 63,* 505–512.

Logan, J. M., Castel, A. D., Haber, S., & Viehman, E. J. (2012). Metacognition and the spacing effect: The role of repetition, feedback, and instruction on judgments of learning for massed and spaced rehearsal. *Metacognition and Learning, 7,* 175–195.

Magreehan, D. A., Serra, M. J., Schwartz, N. H., & Narciss, S. (2015). Further boundary conditions for the effects of perceptual disfluency on judgments of learning. *Metacognition and Learning*, 1-22.

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

Pashler, H., Cepeda, N. J., Wixted, J. T., & Rohrer, D. (2005). When does feedback facilitate learning of words? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(1), 3.

Rhodes, M. G. (2016). Judgments of learning. In J. Dunlosky & S. K. Tauber (Eds.), *The Oxford handbook of metamemory*. New York: Oxford UP.

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.

Rhodes, M. G., & Tauber, S. K. (2011). Eliminating the delayed JOL effect: The influence of the veracity of retrieved information on metacognitive accuracy. *Memory, 19,* 853–870.

Serra, M. J., & Ariel, R. (2014). People use the memory for past-test heuristic as an explicit cue for judgments of learning. *Memory & Cognition, 42,* 1260–1272.

Sitzman, D. M., Rhodes, M. G., Tauber, S. K., & Liceralde, V. R. T. (2015). The role of prior knowledge in error correction for younger and older adults. *Aging, Neuropsychology, and Cognition, 22,* 502–516.

Sitzman, D. M., Rhodes, M. G., & Tauber, S. K. (2014). Prior knowledge is more predictive of error correction than subjective confidence. *Memory & Cognition, 42,* 84–96.

Skinner, B. F. (1954). The science of learning and the art of teaching. *Harvard Educational Review, 24,* 86–97.

Tauber, S. K., & Rhodes, M. G. (2010). Metacognitive errors contribute to the difficulty in remembering proper names. *Memory, 18,* 522–532.

Tauber, S. K., & Rhodes, M. G. (2012). Multiple bases for young and older adults' judgments of learning in multitrial learning. *Psychology and Aging, 27*(2), 474.