DISSERTATION

THE INFLUENCE OF FEEDBACK ON PREDICTIONS OF FUTURE MEMORY PERFORMANCE

Submitted by

Danielle Sitzman

Department of Psychology

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Fall 2013

Doctoral Committee:

Advisor: Matthew Rhodes

Anne Cleary
Deana Davalos
Dan Robinson

ABSTRACT

THE INFLUENCE OF FEEDBACK ON PREDICTIONS OF FUTURE MEMORY PERFORMANCE

The current experiments explored metacognitive beliefs about feedback. In Experiment 1, participants studied Lithuanian-English word pairs, took an initial test, were either shown correct answer feedback, right/wrong feedback, or no feedback. They then made a judgment of learning (JOL) regarding the likelihood of answering this item correctly on a later test. Participants were tested on the same word pairs during the final test. Although average JOLs were higher for items in the correct answer feedback condition, relative accuracy was impaired. Experiment 2 explored participants' beliefs about feedback by having half of them make JOLs prior to seeing an item (PreJOLs), with only knowledge of whether feedback would be provided. Participants in both the regular JOL and preJOL conditions provided higher average JOLs for items in the feedback condition than items in the no feedback condition; however relative accuracy was decreased for the feedback condition. In Experiment 3, participants went through a procedure similar to Experiment 1 twice, with two lists of word pairs. Metacognitive accuracy did not improve from List 1 to List 2. Lastly, Experiment 4 used scaffolded feedback to increase metacognitive accuracy. Participants corrected more errors if they could generate the correct response with fewer letter cues. However, relative judgments were not more accurate than the previous experiments. In sum, the current experiments suggest that participants may have a general understanding of the benefits of feedback; however, feedback diminishes prediction accuracy for specific items.

TABLE OF CONTENTS

INTRODUCTION

Feedback is beneficial for correcting errors in memory. If a person answers a question incorrectly, informing that individual of the correct response greatly increases the likelihood of remembering that correct response on a later test. Although much work has examined the efficacy of different forms of feedback (Pashler, Cepeda, Wixted, & Rohrer, 2005), little research has examined whether people are aware of the benefits of feedback (but see Kornell & Rhodes, 2013). That is, less is known about whether people are aware that feedback increases the likelihood of correcting errors. Prior research has shown that people are not always accurate when predicting their future memory performance (Koriat, 1997, Koriat, Bjork, Sheffer, & Bar, 2004; Kornell & Bjork, 2009). For example, people often overestimate how much they will remember in the future and may also fail to appreciate how much certain techniques (e.g., additional study opportunities) will improve memory (Kornell & Bjork, 2009). One way to characterize such data is in terms of a stability bias: People view their memory performance as stable over time and fail to consider future forgetting or future improvements (Kornell & Bjork, 2009; Kornell, Rhodes, Castel, & Tauber, 2011).

The current experiments explored the influence of feedback on predictions of future memory performance. Kornell and Rhodes (2013) reported that, after taking a test (without feedback), participants could accurately distinguish between items they would or would not remember on a later test. However, when participants were provided with feedback after answering a question, they were unable to account for the improvements in memory from this feedback and thus were less accurate at indicating whether or not they would remember the correct response later. 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 improvements in memory due to feedback into their predictions of future memory performance.

Metamemory

Metamemory is often defined as thoughts (cognitions) concerning one's own memory (Koriat, 2007; Nelson & Narens, 1990). Frameworks of metamemory suggest that it is comprised of two main processes: monitoring and control (Koriat, 2007; Nelson, 1996; Nelson & Narens, 1990). Monitoring involves awareness of one's current state of knowledge relative to a desired state of knowledge. Control involves changes in behavior based on monitoring, such as whether to continue or discontinue a behavior (e.g., studying). Monitoring and control processes are assumed to work together to continually assess memory and update study behaviors based on those assessments (Koriat, 2007; Nelson, 1996; Nelson & Narens, 1990).

One way that research has attempted to understand monitoring is by asking participants to make predictions of their future memory performance, known as judgments of learning (JOLs; Nelson & Narens, 1990). Typically, participants will study something, such as a word or image, and then judge the likelihood that they will be able to remember that item on a future test either immediately or at a delay (e.g., Rhodes & Tauber, 2011). This subjective judgment is then compared to final test performance.

JOLs can be considered in terms of absolute and relative accuracy. Absolute accuracy involves comparing average JOLs to overall performance. For example, a participant may study a list of word pairs and predict, on average, that 70% of those words will be remembered. If, on a final test, that individual correctly remembers close to 70% of the word pairs, they would be said to be calibrated (i.e., absolute accuracy would be strong). If their actual performance was lower, perhaps around 50%, then JOLs would be considered overconfident. However, if their actual performance exceeded predictions (e.g., 90%) then a participant's judgments would be considered underconfident.

Whereas absolute accuracy refers to the correspondence between JOLs and performance, relative accuracy (i.e., resolution) refers to how well a person can differentiate between items they will or will not remember on a later test. This is typically measured via a non-parametric, gamma correlation (Nelson, 1984), between JOLs provided during an early study or testing session, and subsequent accuracy for a specific item on a later test. A positive correlation would indicate that participants were more likely to give higher JOLs to items that they answered correctly on a later test and lower JOLs to items that they answered incorrectly. However, negative correlations would indicate that participants gave higher JOLs to items that were more likely to be incorrect and low JOLs to items they were more likely to answer correctly. A weak correlation in either direction would signify that participants were unable to distinguish between items that they would or would not remember on a later test. Ideally, participants should have a strong positive correlation between JOLs and final test accuracy, indicating that they can effectively differentiate between items they will and will not remember. Overall, the goal of much metamemory research is to understand those factors that lead to accurate memory predictions (both in terms of absolute and relative accuracy), and factors that impair a person's ability to accurately assess their memory.

Early research suggested that predictions of future memory performance were based on the strength of a memory trace (Hart, 1965). Such direct-access theories hold that people have access to the strength of their memory and thus, should have fairly accurate predictions. Over time, various discrepancies between subjective and objective performance have been noted, indicating that people do not have direct access to their memory (Benjamin & Bjork, 1996;

Koriat 1997; Koriat et al., 2004). Accordingly, most researchers argue that judgments about future memory performance are inferences based on a variety of cues available. Sometimes these cues help a participant accurately assess memory; however, cues are frequently misleading, and thus, impair monitoring accuracy. Koriat (1997) has proposed three main categories of cues that influence a person's assessment of their memory: intrinsic, extrinsic, and mnemonic. Intrinsic cues involve information about the materials used in the experiment. For example, the related word-pair Cat-Dog provides different cues to judge future memory performance than the unrelated word-pair Dog-Apple. The relatedness of a word pair influences how easily participants feel that they have processed the information (but see Muller, Tauber, & Dunlosky, 2013). Thus, participants are more likely to give higher JOLs to a related word pair (e.g., *Cat-Dog*) than an unrelated word pair (e.g., *Dog- Apple*). Although people are more likely to remember related word pairs compared to unrelated pairs, participants may also attend to other intrinsic cues that are deceptive. For example, Rhodes and Castel (2008) manipulated the font size of words and found that participants were more likely to give higher JOLs to large font words even though size of font had no impact on memory performance. Therefore, intrinsic cues may sometimes be diagnostic of future memory performance; however, they can also be misleading.

The second set of cues, extrinsic cues, involves information regarding how an item was studied (Koriat, 1997). For example, this may include the number of times a specific item was studied, whether it was tested versus restudied, whether encoding was massed or spaced, how deeply the information was processed, etc. If a person studies the same word 5 times, they will generally rate that word as more memorable than a word studied only once. Lastly, mnemonic cues are internal and indicate to a participant how well they have learned the information. These cues can include things such as how quickly information comes to mind when trying to answer a question, how familiar one is with the information being tested, or memory for response accuracy. If a response comes to mind quickly, regardless of whether that response is accurate, participants are more likely judge that they will remember that response in the future (Koriat, 1997).

A person's current memory state (whether they know a piece of information) is a highly salient cue that influences predictions of future performance (Koriat et al., 2004; Nelson & Dunlosky, 1991). If a participant was shown the word *Cat* and asked to recall the word that had been paired with it during an earlier study phase (e.g., *Tree*), the ability to recall *Tree* will change predictions of future performance. If the word *Tree* is recalled, the participant will likely provide a high JOL, indicating that the word has a high probability of being remembered in the future. However, if the participant is unable to generate the word *Tree*, that individual is likely to provide a low JOL indicating he or she is unlikely to remember that word in the future. Thus, people generally display a stability bias in memory predicting that their performance will be the same in the future as it is at that current point in time. This stability bias entails that participants often fail to account for factors that may change their performance over time. For example, providing participants with opportunities to restudy the word pair *Cat-Tree*, or telling participants that the next test will not be for a long period of time (e.g. a year), has little influence on judgments (Koriat et al., 2004). In a series of experiments, Koriat et al. (2004) had participants learn a list of word pairs and then predict the likelihood of remembering those pairs after 10 min, 1 week, 1 month, or 1 year. Koriat et al. found that when the retention interval was manipulated between-subjects, predictions were highly similar. That is, participants predicted that they would remember a similar number of word pairs after 10 min as they would after 1 year and did not seem to take forgetting into account (but see Tauber & Rhodes, 2012a). However, when the retention interval was manipulated within-subjects, participants provided more accurate predictions. Koriat et al. (2004) suggested that manipulating the retention interval within-subjects activated participants' theories of forgetting, which then informed memory predictions.

Koriat et al.'s (2004) results indicate that participants are aware of forgetting only under limited circumstances, even though they may understand that they will forget information over time. To understand this type of discrepancy, previous research has distinguished between experience-based and theory-based predictions (Jacoby & Kelley, 1987; Koriat, 1997; Koriat et al., 2004). Experience-based predictions typically reflect mnemonic cues, such that participants use their immediate experience to infer what they will remember in the future. This immediate experience thus strongly influences predictions of future performance and will generally reflect a stability bias in memory. Theory-based predictions reflect people's understanding of how memory works (e.g., forgetting occurs over time), but these factors typically must be made salient to influence memory predictions. However, most participants understand that they will remember less information in 1 year compared with 10 min.

Other lines of research have also documented discrepancies between a person's memory beliefs and their item-by-item predictions of memory (JOLs). In a series of 12 experiments, Kornell and Bjork (2009; see also Kornell et al., 2011) had participants make predictions of learning (POLs). Participants were told that they would have somewhere between 1 to 4 study-test cycles for a list of words. During a first study trial, after each word pair, participants were informed of how many more study cycles they would have before the final test (either 1 or 4) and made a POL about the likelihood that they would remember the correct response on a final test. If participants were aware of the benefits of multiple study opportunities, then they should have predicted higher levels of performance the more subsequent trials they will have (higher POLs when given 4 study cycles compared with 1 cycle). Overall, memory was much better when participants were provided with multiple study-test cycles than a single study-test cycle. However, participants predicted little to no learning across trials (i.e., POLs were flat across trials). Interestingly, when directly asked about beliefs regarding studying, participants indicated that they believed more study opportunities would lead to better performance (Kornell & Bjork, 2009). Therefore, although participants understood that more studying is beneficial for memory, they were unable to incorporate this information into their memory predictions.

One reason participants may have difficulty incorporating theories of how memory works into predictions is that other information is more salient when making predictions. Finn and Metcalfe (2007; 2008) have suggested that participants may rely on *Memory for Past Test* (MPT) when predicting future memory performance if they had already been tested on the same material. That is, when judging the future memorability of information, people base their judgments on whether that information was recalled during a previous test.

Finn and Metcalfe examined the role of MPT in the Underconfidence with Practice (UWP) effect (Koriat, Sheffer, & Ma'ayan, 2002). UWP refers to the finding that participants become underconfident in their memory predictions across multiple trials with the same list. In these experiments, participants typically study a list of word pairs (e.g. *Dog-Apple*) and make a JOL of the likelihood of remembering the target word (*Apple*) when presented with the cue (*Dog*). After making a JOL for each item, participants are tested on all word pairs. On a second trial, participants again study the same word pairs, make JOLs for each item and then take a test. This generally continues from 2-5 trials. On the first trial, participants are typically overconfident with JOLs higher than their actual performance on the Trial 1 test. However, although performance increases across trials due to more study opportunities, participants' predictions either remain relatively stable or only increase slightly (leading to underconfidence). Finn and Metcalfe (2008) reported that JOLs were primarily based on the recall outcome for that item on the previous test. During the Trial 1 test, if an item was recalled, then that item was given a high JOL during Trial 2 and if it was not recalled during the Trial 1 test, that item was given a low JOL during Trial 2. However, by virtue of having another study opportunity, some items that were not recalled on test 1 were recalled on test 2, yielding underconfident JOLs. Recent work by Tauber and Rhodes (2012b) suggests that although other factors may influence JOLs (e.g., JOLs for the same item on a previous trial), MPT is the strongest predictor of JOLs across trials.

The reliance on MPT suggests that participants generally do not take new learning into account when predicting future memory performance, leading to a stability bias in memory (Finn & Metcalfe, 2007; 2008; Kornell & Bjork, 2009). That is, although participants may understand that more study opportunities will be helpful for memory, the MPT cue may overshadow these beliefs and guide item-by-item predictions. Thus, participants may predict little change in performance for other methods known to enhance future memory performance, such as feedback.

Feedback

Decades of research have shown that feedback is beneficial for memory, enhancing retention of correct responses and facilitating error correction (Butler, Karpicke, & Roediger, 2008; Kulhavy & Anderson, 1972; Pressey, 1950; Skinner, 1954). However, different types of feedback are not equivalently effective (Fazio, Huelser, Johnson, & Marsh, 2010; Pashler et al., 2005). Most research has demonstrated that feedback is especially effective when the correct response is shown (*correct answer feedback*). Thus, when a participant answers a question incorrectly and is then shown the correct response, that error is more likely to be corrected on a future test. When participants are only told whether or not they were correct (*right/wrong feedback*), the benefits of feedback are diminished, particularly for incorrect responses. Right/wrong feedback provides diagnostic information about the accuracy of an item, but does not provide the information necessary to update memory. Thus, because correct answer feedback provides both diagnostic information and information necessary to update memory, it is generally more effective than right/wrong feedback (Fazio et al., 2010; Pashler et al., 2005).

Feedback & Metamemory

Although separate literatures have examined metamemory and feedback, far less research has explored their interaction. The research that has been done has generally examined how retrospective confidence judgments (RCJs) influence participants' attention to feedback (Butler, Karpicke, & Roediger, 2008; Butterfield & Metcalfe, 2001; 2006; Fazio & Marsh, 2009; Kulhavy & Stock, 1989; Sitzman, Rhodes, & Tauber, in press). Unlike JOLs, RCJs are made after a person has provided an answer and reflect a participant's confidence that an answer is correct. High levels of confidence indicate that participants are fairly certain their response is correct, while lower confidence judgments indicate they are not at all confident in their response. Following feedback, participants are more likely to correct high confidence errors compared with low confidence errors, a finding termed the hypercorrection effect (Butterfield & Metcalfe, 2001; 2006). More recently, some research has suggested that prior knowledge may actually modify attention to feedback, while confidence may serve as a proxy for prior knowledge (e.g., Finn & Metcalfe, 2011; Sitzman et al., in press). Overall, this work indicates that there are various factors that may influence attention to feedback which subsequently impacts the effectiveness of feedback for different items.

Although some prior research has examined factors that may influence how feedback is processed (Butler et al., 2008; Butterfield & Metcalfe, 2001; 2006; Fazio & Marsh, 2009; Finn & Metcalfe, 2011; Kulhavy & Stock, 1989; Sitzman et al., in press), little work has examined how feedback may influence a person's ability to predict what information they will remember in the future (but see Kornell & Rhodes, 2013). Yet, upon receiving feedback, a learner frequently must use that feedback to assess whether they will be able to later remember that information. For example, if a student has a cumulative final exam, it is important to review previous exams and correct any errors in knowledge. However, it is unclear whether students understand that receiving feedback is beneficial for future memory performance. Reviewing feedback provides another study opportunity, enhances retention of correct information and allows errors to be corrected. For errors that are not easily corrected, feedback will help the student to identify information needing additional study. Thus, it is important to understand if, after receiving feedback, people can accurately differentiate between information they know and information they do not know.

Recently, Kornell and Rhodes (2013) examined how providing correct answer feedback influenced participants' ability to predict future test performance when compared with taking a test without feedback or only restudying information. In their first experiment, participants studied 36 weakly related word pairs and were assigned to restudy the word pairs, take a test without feedback, or take a test with feedback. For the test conditions, participants were shown the cue word and asked to provide the target. In the feedback condition, the entire cue-target pair was displayed after the participant attempted to provide a response. Immediately following each trial, participants were asked to make a JOL about the likelihood of later correctly remembering the word pair. Participants then completed a final test on all word pairs.

Consistent with previous research, recall on a final test was superior if participants received feedback (82.4% correct) on an initial test compared with an initial test without feedback (56.9%; Kornell & Rhodes, 2013). However, JOLs did not differ between the groups (60.1% and 60.2% respectively). Of primary interest is the correlation between JOLs provided to items on the initial test and performance on the final test (i.e., relative accuracy). Overall, participants in the test-only condition showed the strongest gamma correlations between JOLs and final test performance ($G = .85$), followed by participants in the feedback condition ($G = .55$) and, lastly, participants in the read-only condition ($G = .31$). Experiments 2 and 3 replicated these findings with more educationally relevant materials and procedures. Therefore, although performance was best when feedback was provided, feedback also reduced metacognitive accuracy compared with a condition where a test alone was provided (Kornell & Rhodes, 2013).

As noted previously, metacognitive research generally assumes that monitoring influences control processes. That is, participants should update their study methods and the information they choose to focus on depending upon how well they judge information to be learned. Thus, more accurate monitoring should beget more optimal control of study behaviors. In their fourth experiment, Kornell and Rhodes' (2013) participants made restudy choices instead of making JOLs. During the second phase of the experiment, after either restudying an item, taking a test without feedback, or taking a test with feedback, participants indicated whether they wanted to restudy that item before taking a test (although participants did not restudy any information). After receiving feedback, participants were less able to accurately determine which items needed to be restudied compared with participants who only received a test (without feedback). In sum, Kornell and Rhodes' results suggest that people are not fully aware of the benefits of feedback. When predicting future memory performance, participants who received feedback were less metacognitively accurate than participants who only took a test; yet, feedback led to much better final test performance than a test alone.

Current Experiments

The current experiments attempted to extend Kornell and Rhodes' findings (2013) and tested methods that may improve metacognitive accuracy following feedback. The first two experiments further explored metacognitive accuracy after feedback. Experiment 1 used a within-subjects design and additional forms of feedback. Experiment 2 investigated whether participants' theories of feedback effectiveness matched their item-by-item judgments of how feedback influences memory for a particular item. In these two experiments, it was expected that participants would display a global understanding that receiving feedback improves memory. However, like Kornell and Rhodes (2013), I expected item-by-item judgments to be less accurate when participants received correct answer feedback compared with items receiving right/wrong feedback or no feedback. Specifically, I anticipated that participants would be unable to predict which errors on an initial test would be corrected on a later test, replicating prior work showing discrepancies between participants' theories of memory and their actual item-by-item predictions (Koriat et al., 2004; Kornell & Bjork, 2009). It is likely that item-by-item judgments will be heavily influenced by the *MPT* heuristic (Finn & Metcalfe, 2007; 2008). That is, participants may give higher JOLs to items they answer correctly and lower JOLs to items that are either incorrect or they could not answer. However, these predictions are unlikely to be influenced by the feedback condition. Therefore, although feedback will help correct errors in certain conditions, JOLs are still likely to be low because they will reflect the current state of that item (i.e., it was not remembered) and fail to take into account improvements in performance due to feedback.

Experiments 3 and 4 aimed to increase participants' metacognitive accuracy following feedback. Participants may first need to experience the benefits of feedback before they can accurately assess how it will impact their performance. Therefore, Experiment 3 allowed participants to gain experience with feedback. Participants studied a list of word pairs, took an initial test where they received the various forms of feedback and then took a final test. Participants next completed the same procedure again with a new list of words. This practice may help participants more accurately gauge how feedback will influence later memory performance, perhaps overriding the *MPT* heuristic. However, practice in Experiment 3 may improve absolute accuracy, but not relative accuracy (resolution). That is, participants may increase their average JOLs to better reflect final test performance, but may not better differentiate between items that will be correct or incorrect on the final test. Therefore, Experiment 4 aimed to provide participants with cues that may allow them to better differentiate between errors on an initial test that would be corrected on a later test and those that would not. In sum, these experiments seek to understand people's beliefs about feedback and ways to increase their metacognitive accuracy when predicting how feedback will influence future memory performance.

EXPERIMENT 1

The main goal of Experiment 1 was to replicate Kornell and Rhodes' (2013) findings but with several important differences. First, Kornell and Rhodes used a between-subjects design for their feedback manipulation. Thus, participants received a test with feedback, a test without feedback, or a restudy opportunity. In the current experiment, the feedback condition was manipulated within-subjects. Participants are often much more sensitive to factors expected to influence memory if these factors are manipulated within-subjects rather than between-subjects. As Koriat et al. (2004) demonstrated, manipulating variables within-subjects can help to activate theories of memory that may be beneficial for more accurate judgments (but see Kornell & Bjork, 2009). Therefore, although Kornell and Rhodes (2013) demonstrated that participants were unable to predict the benefits of feedback in a between-subjects condition, a within-subjects condition may make the benefits of feedback more salient.

Secondly, Kornell and Rhodes (2013) compared correct answer feedback to a no feedback and restudy condition. However, various types of feedback have been explored over the past few decades. For example, a great deal of research has examined right/wrong feedback (Pashler at al., 2005). This type of feedback indicates whether an answer is correct, but does not provide the information necessary to update memory. Therefore, it is unclear whether all forms of feedback will lead to a metacognitive deficit, or if this only pertains to correct answer feedback. Previous work by Rhodes and Tauber (2011, Experiment 3) demonstrated that if participants were given right/wrong feedback before providing a JOL, they 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 1, participants either received correct answer feedback, right/wrong feedback, or no feedback. Unlike Kornell and Rhodes (2013), there was no restudy only condition.

Participants in the current experiment studied Lithuanian-English word pairs and were asked to provide the English translation when shown the Lithuanian word on a later test. These were educationally relevant materials that allowed for clear feedback regarding the correct response. Experiment 1 was expected to replicate Kornell and Rhodes' (2013) findings. That is, I predicted that after correct answer feedback, participants would be less able to differentiate between what they would and would not remember on the final test when compared with a test without feedback. In contrast, although right/wrong feedback is unlikely to lead to improvements in memory, it does provide information about the accuracy of a response. Thus, replicating Rhodes and Tauber (2011), participants should be likely to give high JOLs to items they answer correctly on test 1 and low JOLs to items they answer incorrectly. Because right/wrong feedback is unlikely to improve performance over time, accuracy on test 1 is a good predictor of accuracy for that item on test 2.

Methods

Participants

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

Materials

Materials consisted of 34 easy to moderate Lithuanian-Word pairs taken from normative data reported by Grimaldi, Pyc, and Rawson (2010). Difficulty was determined by the percentage of Grimaldi et al.'s participants who correctly recalled the English translation on an initial test. Word pairs used in the current experiment were recalled by 19%-56% of participants during an initial trial of Grimaldi et al.'s study.

Procedure

Participants studied a randomized list of 34 word pairs, with the first 2 and last 2 pairs serving as buffers accounting for primacy and recency effects that were not included in analyses. Each word pair was presented on the computer screen for 4s. After a 500ms 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. During the test, participants were shown the Lithuanian word for 10s and asked to recall the English translation aloud to an experimenter who coded the response. For each item, participants then received one of the following feedback conditions. For one-third of the items, participants received correct answer feedback; both the Lithuanian word and the English translation appeared on the screen for 5s. 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 to indicate whether or not the participant answered correctly. This was displayed for 5s. For the remaining third of the items, no feedback was presented. However, to equate time across items, "please wait for the next screen" was displayed on the computer screen for 5s. The feedback condition for an individual item was counterbalanced across participants. Within each trial, the order of presentation for items was uniquely randomized for each participant. After the feedback condition, participants were asked to make a JOL indicating the likelihood that, when shown the Lithuanian word, they would be able to recall the English translation on a final test. Their judgments were made on a scale of 0-100%, with 0 indicating no likelihood of recalling the

English translation and 100 being absolutely likely to recall 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. Given that confidence judgments were tangential to the focus of the current studies, they were not analyzed.

Results

Test Performance

The percentage of words correctly recalled on test 1 and test 2 (see Table 1) was examined via two one-way ANOVAs. As would be expected, performance on test 1 was equivalent among the feedback conditions, $F < 1$. However, on test 2, there was a difference among feedback conditions, $F(2, 82) = 45.37$, $p < .001$, $\eta^2_p = .53$. Items that received correct answer feedback ($M = 63.10$, $SE = 3.24$) were more likely to be remembered on test 2 than items receiving right/wrong feedback ($M = 35.71$, $SE = 3.12$), $t(41) = 7.56$, $p < .001$, $d = 1.33$, or items not receiving any feedback ($M = 33.57$, $SE = 3.06$), $t(41) = 8.54$, $p < .001$, $d = 1.44$. There was no difference in test 2 performance between items receiving right/wrong feedback and items not receiving feedback, $t < 1$ (see Figure 1).

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. 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. Participants were highly likely to remember correct responses regardless of whether the item had received no feedback (*M* = 95.24, *SE* = 1.80), right/wrong feedback (*M* = 92.73, *SE* = 3.30), or correct answer feedback (*M* = 92.86, *SE* = 3.60).

The percentage of errors on test 1 that were corrected on test 2 was also examined via a one-way ANOVA. Overall, the feedback condition influenced the percentage of errors corrected, *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 (*M* = 50.21, *SE* = 3.31) compared with items given right/wrong feedback (*M* = 6.38, *SE* = 1.46), *t*(41) = 12.78, *p* < .001, *d* = 2.58, and items given no feedback (*M* = 4.99, *SE* = 1.45), *t*(41) = 11.86, *p* < .001, *d* = 2.76. However, there was no difference in the percentage of errors corrected between the right/wrong feedback and no feedback conditions, *t* < 1.

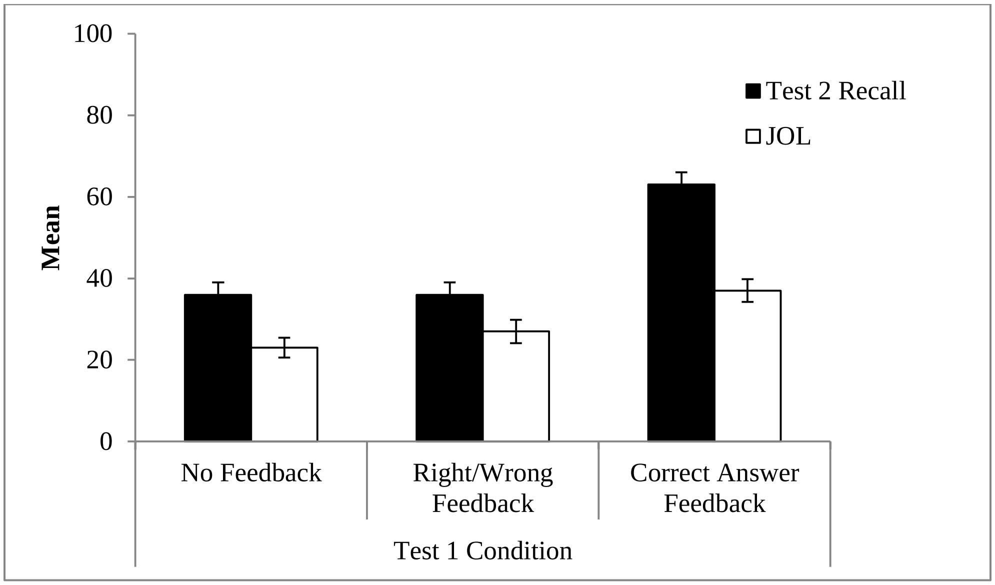

Figure 1. Recall on test 2 compare with average judgments of learning (JOL) provided during test 1 for Experiment 1. Error bars represent one standard error of the mean.

Memory Predictions

Absolute accuracy. Average JOLs (see Figure 1) given to items on test 1 were compared between the three feedback conditions. Overall, average JOLs differed as a function of feedback type, *F*(2, 82) = 14.30, *p* <.001, η^2_p=.26. Participants' average JOLs were higher following items in the correct answer feedback condition (*M* = 37.09, *SE* = 2.77) compared with items given right/wrong feedback (*M* = 26.92, *SE* = 2.88), *t*(41) = 3.47, *p* < .001, *d* = .56, and items that were given no feedback (*M* = 23.20, *SE* = 2.44), *t*(41) = 6.24, *p* > .001, *d* = .82. There was no difference in JOLs between items in the right/wrong feedback condition and the no feedback condition, *t*(41) = 1.31, *p* = .20, *d* = .22.

Two one-way ANOVAs were used to analyze average JOLs for each feedback condition based on whether the participant answered correctly during test 1. For correct responses, average JOLs differed based on the type of feedback the item received *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. However, average JOLs did not differ for items in the correct answer feedback or right/wrong feedback conditions, *t* < 1.

For items that were incorrect on test 1, average JOLs also differed as a function of feedback condition, *F*(2,82) = 48.55, *p* < .001, η^2_p = .54. On average, 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 items that did not receive feedback (*M* = 4.89, *SE* =

1.11), $t(41) = 7.54$, $p < .001$, $d = 1.18$. However, judgments 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. For each feedback condition, gamma correlations were computed between JOLs during test 1 and recall on test 2 (coded as 1 for correct and 0 for incorrect). Strong, positive correlations indicate that participants gave high JOLs to items that were correct on the final test and low JOLs to items that were incorrect on the final test.

All mean gamma correlations were reliably greater than zero, $t\text{s} \geq 7.55$, $p < .001$. Mean gamma correlations were compared across the three feedback conditions in a one-way ANOVA. Overall, mean gamma correlations differed as a function of the type of feedback, $F(2,68) = 18.29$, $p < .001$, $\eta^2{}_p = .35$. The mean correlation between JOLs and final test accuracy was reliably lower for items in the correct answer feedback condition ($G = .57$, $SE = .08$) compared with items in the right/wrong feedback condition ($G = .92$, $SE = .06$), $t(34) = 5.18$, $p < .001$ $d = .85$, and items in the no feedback condition ($G = .95$, $SE = .02$), $t(34) = 4.87$, $p < .001$, $d = 1.15$. However, there was no difference in relative accuracy for the right/wrong and no feedback conditions, $t < 1$ (see Figure 2).

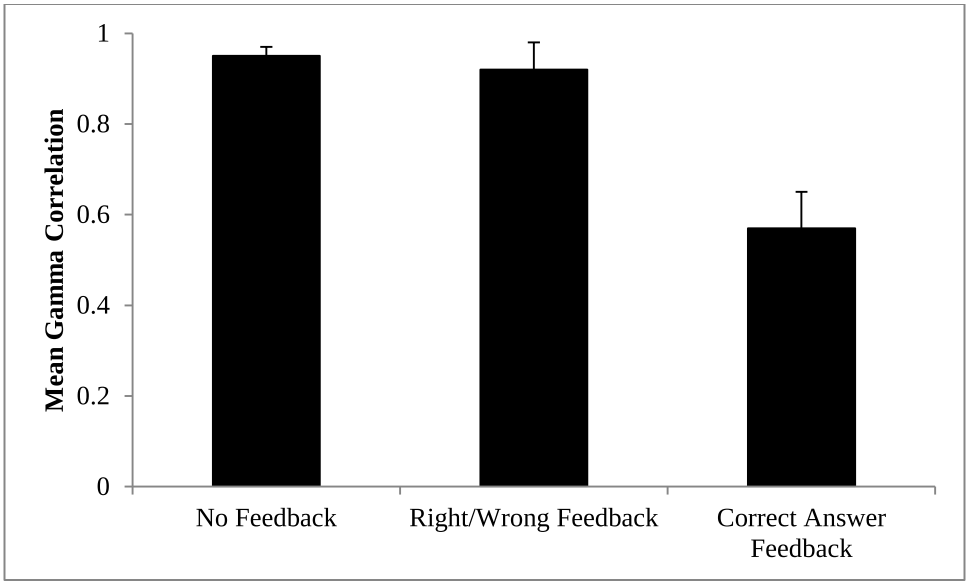

Figure 2. Mean gamma correlation between JOLs on test 1 and accuracy on test 2 for Experiment 1. Error bars represent one standard error of the mean.

Discussion

Overall, Experiment 1 replicated Kornell and Rhodes' (2013) results, with one notable exception. As expected, correct answer feedback led to better performance on test 2 compared with right/wrong feedback and no feedback; but, in contrast to Kornell and Rhodes, participants' JOLs predicted this pattern of results. On average, items in the correct answer feedback condition were given higher JOLs than items in the right/wrong and no feedback conditions. However, participants' predictions regarding the percentage of errors they would correct (20%) underestimated the percentage of errors they corrected on test 2 (50%). A paired samples t-test confirmed that the percentage of errors corrected on the final test (M = 50.21, SE = 3.31) was reliably greater than average JOL for incorrect items on the initial test (M = 20.75, SE = 2.40), $t(41) = 7.40$, $p < .001$, $d = 1.57$. Using a between-subjects design, Kornell and Rhodes reported identical predictions across feedback conditions. That is, participants' average JOLs for the no feedback condition did not differ from JOLs for the feedback condition. This was not replicated in the current study when using a within-subjects design.

Relative accuracy data were consistent with prior work by Kornell and Rhodes as item-by-item judgments were less accurate in the correct answer feedback condition compared with the no feedback and right/wrong conditions. For items in the right/wrong and no feedback conditions, gamma correlations were near unity (+1.0). In all, the results from Experiment 1 suggest that learning is greatest when correct answer feedback is provided; however, item-by-item metacognitive accuracy is lower with such feedback. That is, after correct answer feedback, participants are less accurate at differentiating between what they will and will not remember in the future.

EXPERIMENT 2

In general, Experiment 1 replicated Kornell and Rhodes' (2013) results. Although correct answer feedback led to the best performance on the final test, participants' item-by-item JOLs did not reflect those benefits. However, average JOLs indicated that participants may believe correct answer feedback is more beneficial than other forms of feedback. That is, although participants demonstrated a general understanding that correct answer feedback may be most beneficial for memory, they were unable to identify which specific errors would ultimately be corrected due to correct answer feedback. Therefore, the goal of Experiment 2 was to directly explore participants' beliefs about the benefits of feedback. Although participants may not accurately predict the benefits of feedback for a specific item, they may still have a general understanding of its benefits. As mentioned previously, it is important to distinguish between experience-based and theory-based judgments (Koriat et al., 2004). Thus, during Experiment 1, JOLs may have been heavily experience-based. That is, participants may have relied on their memory for the outcome of the recent test (MPT) as the basis for JOLs while not accounting for the feedback conditions. Experiment 2 attempted to make feedback more salient. By drawing attention to the feedback condition, it was expected that participants would shift to providing theory-based predictions that would be more predictive of final test performance than experience-based predictions.

Previous research by Kornell and Bjork (2009; see also Kornell et al., 2011) has shown that participants believe multiple study opportunities are beneficial for memory, but they do not take the number of study trials into account when making a prediction for a specific item. In several experiments, Castel (2008) explored this discrepancy between participants' beliefs about memory and their item-by-item judgments by examining participants' metacognition regarding the serial position effect (i.e., the finding that individuals are most likely to recall words presented at the beginning and end of a list). Although the serial position effect is well established, prior work had not examined whether participants are aware of this effect. Castel had participants study a list of 15 words and provide a JOL for each item, followed by a test of free recall for all items. Results from an initial experiment demonstrated that participants did not vary their JOLs based on a word's position in the list.

In follow-up experiments, participants were asked to provide pre-JOLs based only on a word's serial position (Castel, 2008). That is, participants were asked to predict the likelihood of recalling an item before they were shown that item, rather than make JOLs after studying an item. Before seeing the study word, participants were shown the serial position number and prompted to provide their pre-JOL. After making this pre-JOL, participants studied the word. Castel's rationale was that, while studying an item, participants may be focused on the characteristics of that particular item and pay less attention to cues that predict later performance (e.g., serial position). Thus, when the item is present, judgments are more likely to be experience-based. However, if participants are asked to provide JOLs before seeing an item, other cues may be more salient. Therefore, if the only information present when making a JOL is the serial position of that item, then judgments should be influenced by a person's theory of how serial position impacts memory. If participants understand the serial position effect, they should give higher JOLs to items at the beginning and end of the study list. Overall, Castel demonstrated that pre-JOLs were sensitive to serial position: Participants provided higher pre-JOLs for items at the beginning of the list and items at the end of the list, compared with items in the middle of the list, consistent with recall performance. Castel suggested that by soliciting pre-JOLs, participants were able to focus on extrinsic factors that are important for actual memory performance (e.g., serial position of an item). When these extrinsic factors are made salient, participants are able to incorporate relevant information into their memory predictions (see also Ariel & Dunlosky, 2011).

Experiment 2 used the pre-JOL method introduced by Castel (2008) to explore whether participants can take the benefits of feedback into account when correct answer feedback is made salient. Half of the participants in Experiment 2 were asked to make pre-JOLs (before seeing the item and receiving feedback) whereas the other half made JOLs in a similar manner to Experiment 1 (after feedback). In the pre-JOL condition, it was expected that judgments would be higher for items designated to receive feedback compared with items in the no feedback condition. In the JOL condition, I expected to replicate Kornell and Rhodes (2013). That is, JOLs should be similar for items receiving feedback compared with items in the no feedback condition. Alternatively, based on the results from Experiment 1, participants' average JOLs may be higher for items receiving feedback compared with items not receiving feedback.

Methods

Participants

Eighty four undergraduate students at Colorado State University completed this experiment for partial course credit.

Materials & Procedure

Participants studied the same Lithuanian-English word pairs used in Experiment 1. The procedure was identical to Experiment 1 with several exceptions. First, Experiment 2 and all subsequent experiments contained a test-only (no feedback) condition and a test with correct answer feedback condition. Given that correct answer feedback results in the greatest improvements to memory performance, the remainder of the experiments examined discrepancies between predictions and performance following correct answer feedback or no feedback.

In Experiment 2, half of the participants made JOLs for each item after the feedback/no feedback condition (JOL condition), while the other half of the participants made pre-JOLs before receiving feedback (pre-JOL condition). The JOL condition followed the same procedure as Experiment 1.

The pre-JOL condition was identical to the JOL condition, except for the timing of the JOL prediction. On the initial test, participants were first told whether or not that item would receive feedback and were asked to make a judgment about the likelihood that they would remember that item on a later test (before seeing the item). Next, they were shown the Lithuanian word and asked to recall the English translation. Finally, they were shown the correct response for the half of the items designated to receive feedback. As with participants in the JOL condition, participants in the pre-JOL condition received a final test on all the Lithuanian-English word pairs.

Results

Test Performance

The percentage of words correctly recalled on the initial and final tests (see Figure 3) was examined in a 2 (Judgment condition: JOL, Pre-JOL) x 2 (Feedback type: feedback, no feedback) x 2 (Test: test 1, test 2) mixed-factor ANOVA. Overall, there was a main effect of Test $F(1,82) = 453.80$, $p < .001$, $\eta^2_p = .85$. The percentage of words correctly recalled increased from test 1 ($M = 26.50$, $SE = 1.70$) to test 2 ($M = 43.40$, $SE = 2.10$). However, there was no difference based on whether participants made JOLs or Pre-JOLs, $F < 1$.

There was also a main effect of Feedback Type, *F*(1,82) = 123.18, *p* <.001, η^2_p = .60. Participants were more likely to correctly remember items in the feedback condition (*M* = 43.00, *SE* = 2.00) compared with items that did not receive feedback (*M* = 26.90, *SE* = 1.90). However, this did not differ as a function of JOL condition, *F*(1,82) = 1.74, *p* =.19, η^2_p =.02. Lastly, there was an interaction between test and feedback type, *F*(1,82) = 257.24, *p* <.001 , η^2_p = .76. As anticipated, there was no difference between the feedback condition (*M* = 27.40, *SE* = 1.90) and the no feedback condition (*M* = 25.60, *SE* = 1.90) on test 1, *t*(83) = 1.10, *p* = .27. However, on test 2, participants were more likely to correctly recall items that had received correct answer feedback (*M* = 58.60, *SE* .240) compared with items that had received no feedback (*M* = 28.3, *SE* = 2.00), *t*(83) = 16.67, *p* < .001.

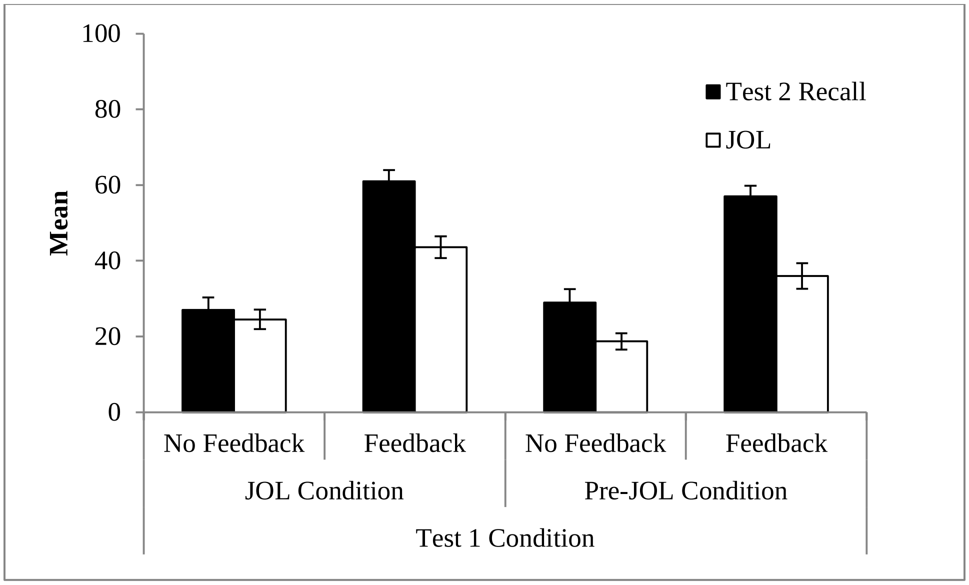

Figure 3. Recall on test 2 compared with average judgments of learning (JOL) provided during test 1 for Experiment 2. Error bars represent one standard error of the mean.

Memory Predictions

Absolute Accuracy. Average JOLs (Figure 3) were compared for each judgment condition. Two 2 (Feedback type: Feedback, no feedback) x 2 (Accuracy: correct, incorrect) ANOVAs were computed to examine JOLs in the JOL and pre-JOL conditions.

In the JOL condition, there was a main effect of feedback type, $F(1,35) = 56.81$, $p <.001$, $\eta^2_p =.62$. Participants gave higher JOLs to items in the feedback condition ($M = 57.35$, $SE = 2.37$) compared with items in the no feedback condition ($M = 39.20$, $SE = 2.38$). There was also a main effect of accuracy, $F(1,35) = 426.96$, $p <.001$, $\eta^2_p =.92$. Participants gave higher JOLs to items answered correctly ($M = 76.67$, $SE = 2.94$), compared with items answered incorrectly ($M = 19.88$, $SE = 1.88$). However, feedback type and accuracy did not interact, $F < 1$. Participants gave higher JOLs to correct items compared with incorrect items regardless of whether or not the item received feedback.

In the pre-JOL condition, there was a main effect of feedback condition $F(1,39) = 14.54$, $p <.001$, $\eta^2_p =.27$. Participants gave higher judgments to items in the feedback condition ($M = 24.39$, $SE = 1.21$) compared with items in the no feedback condition ($M = 18.73$, $SE = 2.25$). As anticipated, there was not a main effect of accuracy $F(1,39) = 1.94$, $p = .17$, $\eta^2_p = .047$. Pre-JOLs did not differ as a function of accuracy on test 1. Lastly, feedback type did not interact with accuracy, $F(1,39) = 2.99$, $p = .09$, $\eta^2_p =.07$.

Average judgments were compared between the JOL conditions in a 2 (Judgment condition: JOL, pre-JOL) x 2 (Feedback type: Feedback, no feedback) mixed-factor ANOVA. Overall, participants gave higher judgments to items in the feedback condition ($M = 39.78$, $SE = 2.23$) compared with items in the no feedback condition ($M = 21.21$, $SE = 1.68$), $F(1,82) = 91.90$. $p <.001$, $\eta^2_p =.53$. In addition, participants in the JOL condition ($M = 34.04$, $SE = 2.45$) gave marginally higher judgments than participants in the pre-JOL condition (*M* = 27.35, *SE* = 2.45), *F*(1,82) = 3.73, *p* = .06, η^2_p =.04. Importantly, feedback type and JOL condition did not interact, *F* < 1.

Relative Accuracy. For the JOL and pre-JOL conditions, gamma correlations were computed between JOLs during test 1 and performance on test 2 for the feedback and no feedback conditions (see Figure 4). The mean gamma correlations in the JOL condition replicated Experiment 1 and prior work by Kornell and Rhodes (2013). For items in the no feedback condition (*G* = .87, *SE* = .06) participants were highly accurate when asked to differentiate between what they would and would not remember on a later test. However, accuracy was reliably lower for items in the feedback condition (*G* = .50, *SE* = .07), *t*(35) = 4.07, *p* < .001, *d* = .99. Mean correlations for items in the No Feedback, *t*(35) = 15.51, *p* < .001, and Feedback conditions, *t*(35) = 7.02, *p* < .001, were reliably greater than zero.

In the pre-JOL condition, participants were asked to make JOLs before seeing an item and providing a response. Thus, relative accuracy was expected to be poor. In line with this, judgments did not reliably predict later test performance for items in the no feedback (*G* = -.15, *SE* = .10), *t*(32) = 1.60, *p* = .12, or feedback conditions (*G* = -.14, *SE* = 07), *t*(32) = 1.81, *p* = .08. Judgments did not reliably differ between feedback conditions, *F* < 1. There was no reliable difference in gamma correlations between feedback conditions, *t* < 1.

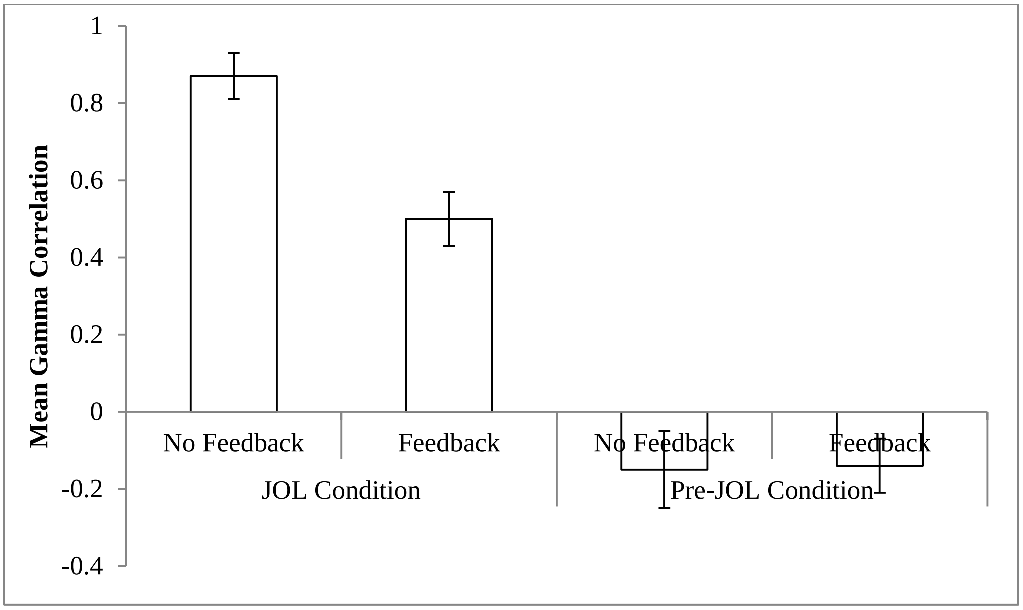

Figure 4. Mean gamma correlations between JOLs on test 1 and accuracy on test 2 for Experiment 2. Error bars represent one standard error of the mean.

Discussion

In both the JOL and PreJOL conditions, participants' average judgments indicated that they understood that receiving feedback would lead to better performance on the final test than not receiving feedback. It was expected that participants in the Pre-JOL condition would predict that feedback would lead to better final performance than no feedback. This follows previous work by Castel (2008) demonstrating that participants can make accurate memory predictions when they are focusing on information that is relevant to later test performance. However, participants in the JOL condition also gave higher average JOLs to items receiving feedback compared with items not receiving feedback. This contradicts previous work by Kornell and Rhodes (2013) but does replicate the pattern of results for absolute judgments in Experiment 1. Despite demonstrating knowledge that feedback would lead to better performance, participants in the JOL condition were unable to use this information to accurately differentiate between items they would and would not remember on a later test after receiving feedback on an initial test.

Thus, participants displayed a global understanding of the benefits of feedback but were unable to utilize this knowledge when predicting performance on specific items.

EXPERIMENT 3

Experiment 2 demonstrated that participants understand that feedback is beneficial for memory, but still have difficulty accounting for this when predicting future memory performance for specific items. Therefore, the goal of Experiment 3 was to provide participants with the opportunity to improve their metacognitive accuracy after receiving feedback. In Experiment 3, participants went through two lists. List 1 was the same as Experiment 1 (without the right/wrong feedback condition). However, in Experiment 3, participants completed this study-test-test cycle twice. At the end of the first study-test-test trial, participants continued on to List 2, a new Lithuanian-English word pair list. It may be that although participants appreciate that feedback is beneficial for memory, they do not fully understand the magnitude of improvement that feedback will have without any prior experience. Thus, providing them with practice during the first trial may allow them to adjust JOLs during the second trial. Therefore, in Experiment 3, I expected to find an overall improvement in absolute accuracy during List 2. That is, after completing List 1, average JOLs during the initial test for List 2 should more accurately reflect performance on the final test, especially in the feedback condition. It is less clear whether task experience will improve relative accuracy. Although participants may be more likely to give higher JOLs overall to feedback items during the second list, this does not necessarily mean that they will be better able to distinguish between items they will or will not answer correctly on the final test.

Methods

Participants

Participants consisted of 42 Colorado State University undergraduate students who completed the experiment for partial course credit.

Materials and Procedures

The materials and procedures were similar to Experiment 1. During List 1, Participants studied 30 Lithuanian-English word pairs. On an initial test they were shown the Lithuanian word and asked to provide the English translation. Following this, they either received correct answer feedback or waited for the next screen before making their JOL for that item. After a short distractor task, participants completed a final test. During List 2, participants completed the same process as List 1 with a new list of Lithuanian-English word pairs. The presentation order of lists was counterbalanced across participants.

Results

Test Performance

List 1. The percentage of items correctly recalled on test 1 and test 2 was examined via a 2 (Test: Test 1, Test 2) x 2 (Feedback: Feedback, No Feedback) repeated-measures ANOVA. Overall, participants correctly recalled a greater percentage of correct responses on test 2 (*M* = 50.20, *SE* = 2.80) than test 1 (*M* = 32.30, *SE* = 2.60), *F*(1, 41) = 251.94, *p* < .001, η^2_p= .86. As well, items that received feedback (*M* = 50.00, *SE* = 3.20) were more likely to be correctly recalled compared with items that did not receive feedback (*M* = 32.50, *SE* = 2.60), *F*(1, 41) = 51.49, *p* < .001, η^2_p= .56. These main effects were qualified by a reliable interaction, *F* (1, 41) = 195.76, *p* < .001, η^2_p= .83. As would be expected on test 1, there was no difference in correct recall for items given feedback (*M* = 32.86, *SE* = 3.21) and items not given feedback (*M* = 31.75, *SE* = 2.58), *t* < 1. However, on test 2, participants were more likely to correctly recall items that were initially given feedback (*M* = 67.14, *SE* = 3.43) compared with items that did not receive feedback on the initial test (*M* = 33.33, *SE* = 2.75), *t*(41) = 11.98, *p* < .001, *d* = 1.65.

The percentage of words correctly recalled 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 difference in the percentage of correct responses retained from test 1 to test 2 between items in the feedback (M = 91.67, SE = 3.25) and no feedback (M = 89.87, SE = 2.21) conditions, $t < 1$. However, participants corrected a greater percentage of errors when items were given feedback (M = 59.55, SE = 3.74) compared with items not given feedback (M = 6.93, SE = 1.68), $t(41) = 14.63$, $p < .001$, $d = 2.65$.

List 2. The percentage of items correctly recalled on test 1 and test 2 was examined in a 2 (Test: Test 1, Test 2) x 2 (Feedback: Feedback, No Feedback) repeated-measures ANOVA. Overall, participants correctly recalled a greater percentage of correct responses on test 2 (M = 60.90. SE = 3.20) compared with test 1 (M = 45.60, SE = 3.40), $F(1,41) = 166.38$, $p < .001$, η^2_p= .80. Items that received feedback (M = 61.00, SE = 3.40) were more likely to be correctly recalled compared with items that did not receive feedback (M = 45.50, SE = 3.70), $F(1,41) = 33.50$, $p < .001$, η^2_p= .45. Lastly, a reliable Test x Feedback type interaction was present, $F(1,41) = 127.18$, $p < .001$, , η^2_p= .76. On test 1, there was no difference in the percentage of items answered correctly for items given feedback (M = 47.14, SE = 3.81) and items not given feedback (M = 44.13, SE = 3.66), $t(41) = 1.05$, $p = .30$, $d = .12$. However, on test 2, participants were more likely to correctly recall items that were given feedback on the initial test (M = 74.92, SE = 3.22) compared with items that did not receive feedback (M = 46.83, SE = 3.78), $t(41) = 9.55$, $p < .001$, $d = 1.22$.

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 difference in the percentage of correct responses retained from test 1 to test 2 for items in the feedback condition (*M* = 96.48, *SE* = 1.51) and items in the no feedback condition (*M* = 95.02, *SE* = 1.48), *t* < 1. For test 2, participants corrected a greater percentage of errors when items were given feedback (*M* = 61.34, *SE* = 3.89) compared with items not given feedback (*M* = 9.67, *SE* = 2.46), *t*(41) = 11.59, *p* < .001, *d* = 2.44[1].

[1] Memory performance was compared between List 1 and List 2 to examine the influence of practice on test performance in an omnibus analysis. A 2 (List: List 1, List 2) x 2 (Test: Test 1, Test 2) x 2 (Feedback type: Feedback, No Feedback) repeated-measures ANOVA compared the percentage of English translations correctly recalled in List 1 and List 2. Overall, participants correctly recalled a greater percentage of English translations during List 2 (*M* = 53.30, *SE* = 3.30) compared with List 1 (*M* = 41.30, *SE* = 2.60), *F*(1,41) = 22.78, *p* < .001, η^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, η^2_p = .89. Lastly, items in the Feedback condition (*M* = 55.50, *SE* = 3.00) were more likely to be correctly recalled than items in the No Feedback condition (*M* = 39.00, *SE* = 2.70), *F* (1,41) = 72.33, *p* <.001, η^2_p = .64.

The interaction between Test and Feedback type was reliable, *F*(1, 41) = 287.85, *p* < .001, η^2_p = .88. On test 1, there was no difference between the percentage of English translations correctly recalled for items in the feedback condition (*M* = 40.00, *SE* = 3.10) compared with items in the no feedback condition (*M* = 37.90, *SE* = 2.70). On test 2, items in the feedback condition were more likely to be recalled (*M* = 71.00, *SE* = 3.10) compared with items in the no feedback condition (*M* = 40.10, *SE* = 2.80). Lastly, this was qualified by a 3 way interaction among List, Test, and Feedback type, *F*(1,41) = 6.30, *p* = .02, η^2_p = .13. The analyses reported in the text can be viewed as describing this interaction.

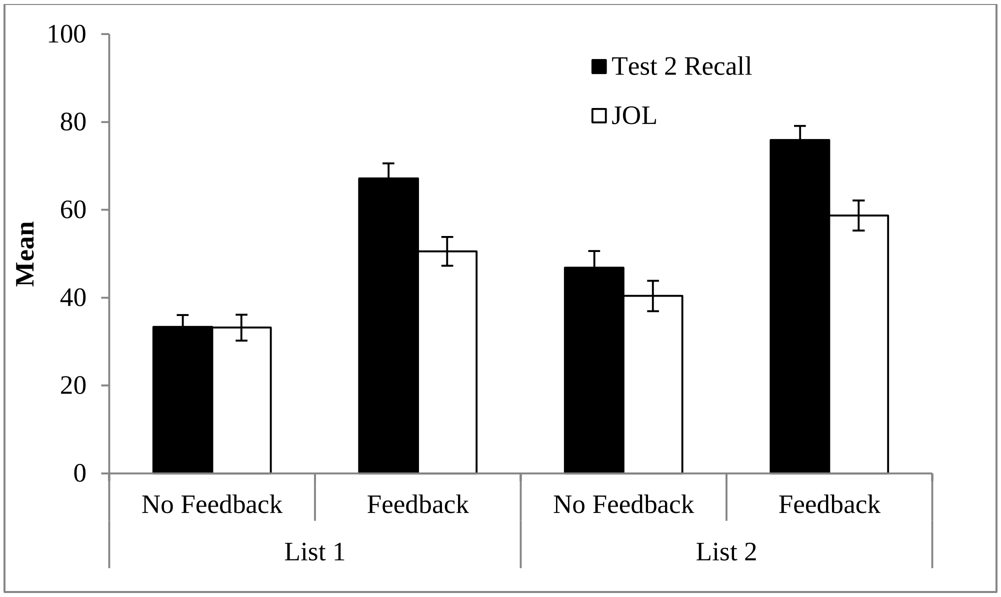

Figure 5. Recall on test 2 compared with average judgments of learning (JOL) provided during List 1 and List 2 for Experiment 3. Error bars represent one standard error of the mean.

Memory predictions: Absolute Accuracy

For each list of words, a 2 (Feedback type: Feedback, no feedback) x 2 (Accuracy: Correct, Incorrect) repeated-measures ANOVA was used to examine average JOLs provided for correct and incorrect responses on the initial test.

List 1. On the initial test for List 1, participants provided higher JOLs for items that were answered correctly (M = 76.04, SE = 2.56) compared with items that were answered incorrectly (M = 27.97, SE = 2.30), F(1,38) = 607.03, p < .001, η^2_p= .94. JOLs were also higher for items in the feedback condition (M = 60.03, SE = 2.83) compared with items in the no feedback condition (M = 43.98, SE = 2.20), F(1,38) = 44.65, p < .001, η^2_p= .54. Lastly, there was a reliable Feedback Type x Text Accuracy interaction, F(1,38) = 8.56, p < .01, η^2_p= .18. For correct items, participants provided reliably higher JOLs for items in the feedback condition (M = 81.54, SE = 2.87) compared with items in the no feedback condition (M = 70.53, SE = 3.00), t(38) = 3.84, p <

.001, *d* = .60. For items that were incorrect on the initial test, participants also provided higher JOLs for items in the feedback condition (*M* = 38.51, *SE* = 3.21) compared with items in the no feedback condition (*M* = 17.43, *SE* = 2.23), *t*(41) = 7.07, *p* < .001, *d* = 1.19.

List 2. On the initial test for List 2, participants provided higher JOLs for items that were answered correctly (*M* = 78.91, *SE* = 3.07) compared with items that were answered incorrectly (*M* = 26.95, *SE* = 2.15), *F*(1,41) = 476.42, *p* < .001, η^2_p= .92. JOLs were also higher for items in the feedback condition (*M* = 61.55, *SE* = 2.59) compared with items in the no feedback condition (*M* = 44.31, *SE* = 2.53), *F*(1,41) = 77.61, *p* < .001, η^2_p= .65. Lastly, there was a reliable interaction between these two variables, *F*(1,41) = 10.20, *p* < .01, η^2_p= .20. For correct items, participants provided reliably higher JOLs for items in the feedback condition (*M* = 84.60, *SE* = 2.84) compared with items in the no feedback condition (*M* = 73.22, *SE* = 3.64), *t*(41) = 5.14, *p* < .001, *d* = .51. For items that were incorrect on the initial test, participants also provided higher JOLs in the feedback condition (*M* = 38.51, *SE* = 2.82) than the no feedback condition (*M* = 15.40, *SE* = 2.45), *t*(41) = 7.50, *p* < .001, *d* = 1.35[2].

[2] Average JOLs were compared between List 1 and List 2 to determine if judgments changed following practice. A 2 (List: List 1, List 2) x 2(Feedback Type: Feedback, No Feedback) x 2(Accuracy: Correct, Incorrect) repeated-measures ANOVA compared the average JOLs provided on Test 1 between List 1 and List 2. Overall, average JOLs on List 2 (*M* = 54.94, *SE* = 52.00) were reliably higher than average JOLs on List 1 (*M* = 52.00, *SE* = 2.23), *F*(1,38) = 4.79, *p* = .04, η^2_p = .112. As well, average JOLs were 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, η^2_p = .66. Lastly, participants provided higher average JOLs for items that were answered correctly (*M* = 78.91, *SE* = 2.29) compared with items that were answered incorrectly (*M* = 28.04, *SE* = 2.09), *F*(1,38) = 834.44, *p* < .001, η^2_p = .956.

There was a reliable interaction between the List and Accuracy, *F*(1,38) = 6.24, *p* = .02, η^2_p = .14. There was no difference between average JOLs on incorrect responses for List 1 (*M* = 27.97, *SE* = 2.31) and List 2 (*M* = 28.10, *SE* = 2.10). However, for correct responses, average JOLs were greater on List 2 (*M* = 81.79, *SE* = 2.44) when compared with List 1 (*M* = 76.04, *SE* = 2.56). There was also a reliable interaction between Feedback type and accuracy. For items in the feedback condition, participants provided reliably greater JOLs for both correct (*M* = 84.05, *SE* = 2.30) and incorrect (*M* = 39.25, *SE* = 2.77) when compared with JOLs for correct (*M*= 73.67, *SE*

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) x 2 (Feedback Type: Feedback, No Feedback) repeated-measures ANOVA (see Figure 6). If practice improves relative accuracy, then gamma correlations for the feedback condition should increase from List 1 to List 2. Overall, there was not a reliable difference in correlations based on List, $F(1, 30) = 2.08$, $p = .16$, $\eta^2_p = .07$. The gamma correlations between JOLs on the initial test and accuracy on the final test did not differ in magnitude between list 1 ($M = .76$, $SE = .04$) and list 2 ($M = .82$, $SE = .03$). However, there was a reliable effect of feedback condition. Gamma correlations were stronger for items in the no feedback condition ($M = .91$, $SE = .02$) compared with items in the feedback condition ($M = .67$, $SE = .04$), $F(1,30) = 27.94$, $p < .001$, $\eta^2_p = .48$. There was no reliable interaction, $F < 1$. On List 1, there was a stronger correlation between test 1 JOLs and test 2 accuracy for items in the no feedback condition ($G = .87$, $SE = .04$) compared with items in the feedback condition ($G = .65$, $SE = .06$), $t(30) = 3.16$, $p < .001$, $d = .77$. This pattern was replicated for List 2. Gamma correlations were reliably greater for items in the no feedback condition ($G = .94$, $SE = .05$) compared with items in the feedback condition ($G = .69$, $SE = .05$), $t(30) = 5.18$, $p < .001$, $d = 1.19$. All gamma correlations reported were reliably greater than zero, $ts \geq 10.90$.

= 2.59) and incorrect responses ($M = 16.83$, $SE = 2.25$) in the no feedback condition. There was not a reliable interaction between List and Feedback type, $F < 1$, and the three way interaction between List, Feedback type, and Accuracy was not reliable, $F < 1$.

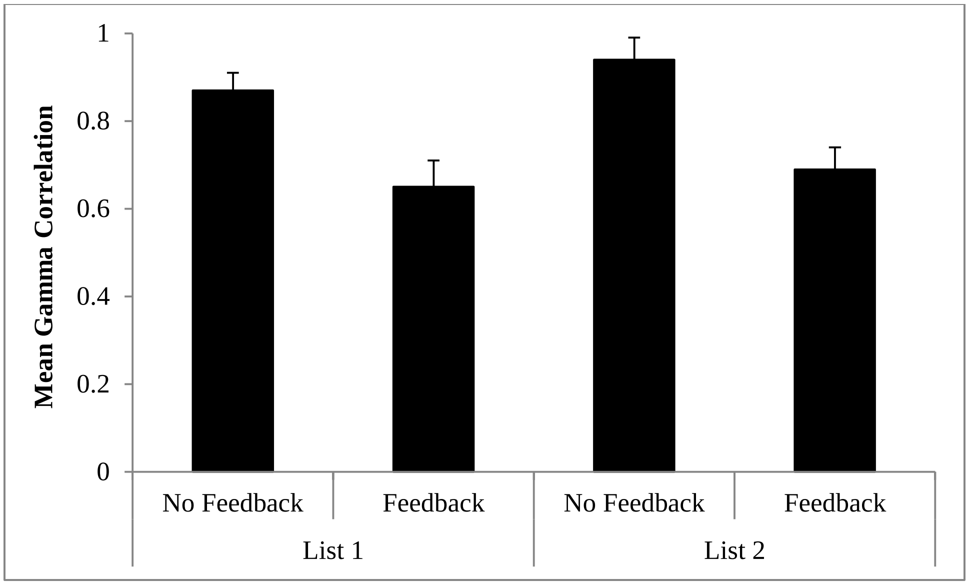

Figure 6. Mean gamma correlations between JOLs on test 1 and accuracy on test 2 for List 1 and List 2 in Experiment 3. Error bars represent one standard error of the mean.

Discussion

Consistent with 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. This was reflected in average JOLs for both List 1 and List 2. Even following practice, JOLs were underestimates of the likelihood of improvement for errors when given feedback. Practice also did not improve relative accuracy across lists. For both of the lists, gamma correlations between JOLs and final test accuracy were stronger for items in the no feedback condition than items in the feedback condition. The magnitude of these correlations did not improve from List 1 to List 2. Overall, participants demonstrated that they understand that feedback will ultimately lead to better final test performance; however, feedback still impaired their ability to predict the future accuracy of an individual item on a later test. Although practice on the task led to improved memory performance on the final test for list 2, it did not lead to improvements in either absolute or relative metacognitive accuracy.

EXPERIMENT 4

In the previous three experiments, participants demonstrated that they have a general understanding that feedback is beneficial for memory. However, although participants understand that feedback is beneficial, they may have difficulty differentiating between errors that will be corrected and those that will not. Thus, the main goal of Experiment 4 was to provide participants with a cue that would help them to determine which errors would be corrected by feedback on a later test and which errors would likely remain uncorrected.

Across experiments, participants have generally predicted that there is a low likelihood that errors will be corrected on a later test (i.e., JOLs have averaged about 25% for feedback items and about 15% for no feedback items). Although these judgments are highly predictive of later test performance in the no feedback condition, they fail to accurately account for error correction that will happen as the result of feedback. Indeed, in the first three experiments, approximately 50% of errors on test 1 were corrected on test 2 when participants were given feedback. Thus, to demonstrate their understanding of feedback, participants may need cues that indicate the likelihood of an error being corrected on a later test.

In a series of experiments, Finn and Metcalfe (2010) explored the efficacy of various forms of feedback. Finn and Metcalfe had participants answer general knowledge questions and, after an error, receive one of four types of feedback. Correct answer feedback involved showing participants the correct answer after an error. Minimal feedback involved informing participants when they answered incorrectly and allowed them one more attempt to answer the question. In the answer-until-correct (AUC) feedback condition, participants were shown 6 options after an error and asked to pick the correct response. If they picked the wrong option, it turned red and they were told to pick again. They continued to choose among options until they picked the correct answer. In the scaffolded feedback condition, participants were initially shown the first letter of the correct response and then asked to generate the correct answer. If they generated the correct answer, they moved on to the next question. If they could not generate the correct answer, they were shown the second letter and again asked to generate the correct response. This procedure continued until participants either generated the correct response or the entire correct answer was displayed.

Overall, Finn and Metcalfe (2010) found that standard correct answer feedback and scaffolded feedback led to a greater proportion of correct responses on a final test than minimal feedback or AUC multiple choice feedback. However, scaffolded feedback, especially after a longer retention interval, was better for retention than correct answer feedback. They suggested that scaffolded feedback capitalizes on the benefits of retrieval practice (Bjork, 1975) while participants are trying to generate the correct response. Of interest to Experiment 4 are the results from the scaffolded feedback condition when considering final test performance by the number of letter cues participants needed before remembering the correct answer. Overall, 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 (1 or 2 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. Finn and Metcalfe's (2010) results suggest that when fewer cues are presented, more "effort" is required to retrieve the correct response, and thus, the benefits of retrieval are more substantial than if more cues are provided to the participants (Carpenter & DeLosh, 2006).

Such findings suggest that participants are more likely to correct errors if they are able to generate the correct answer with fewer cues than if they require more cues, or the entire word, to be presented. 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. After errors, if participants can generate the correct response after 1 or 2 letter cues, then they should be more likely to correctly remember the answer on the final test than if they require more cues or the entire word. If they are sensitive to the number of letters needed, then participants should give higher JOLs to items that only require 1 or 2 letter cues to generate the correct response, and lower JOLs to items that require more cues or the entire word. Alternatively, participants may feel that generating the answer after 1 or 2 letter cues is more difficult than generating an answer when most or all of the word is already presented. Previous research suggests that predictions of future memory performance are often higher for items that are easily generated compared with items that may take longer or more effort to generate (Benjamin, Bjork, & Schwartz, 1998). Therefore, participants in Experiment 4 may regard answers generated after 1 or 2 letter cues to be more difficult, and thus less memorable, when compared with items that were generated after most or all of the word was already displayed. If this is true, then participants will give higher JOLs to items that required most or all of the cues to generate, and lower JOLs to items where the correct answer was generated after 1 or 2 letter cues.

Methods

Participants

Participants consisted of 60 undergraduate students at Colorado State University who completed the experiment for partial course credit.

Materials and Procedures

The materials and procedures were identical to the previous experiments except for the feedback procedure during the initial test. Specifically, feedback during Experiment 4 was presented in a scaffolded manner similar to Finn and Metcalfe (2010). After a correct response, participants were shown the correct English translation for 5s before moving on 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 5s. If they could not correctly recall the 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 Experiment 1; participants completed a distractor task and then took the final test.

Results

Test Performance

A paired samples t-test compared the percentage of correct answers recalled on test 1 and test 2. Overall, participants correctly recalled a greater percentage of English translations on test 2 (M = 55.12, SE = 2.91) compared with test 1 (M = 33.45, SE = 2.65), t(59) = 18.75, p < .001, d = .98.

Final test performance as a function of number of feedback cues needed. A one-way ANOVA compared the percentage of errors corrected on the final test when 1, 2, or more than 3 letters were needed to generate the correct response on the initial test (see Table 3 for frequency data). Overall, the percentage of errors corrected 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 could generate the correct response after 1 letter cue, they corrected a greater percentage of errors (*M* = 50.77, *SE* = 4.16) than if they required 2 letter cues to generate the correct response, (*M* = 38.68, *SE* = 3.85), *t*(59) = 2.67, *p* = .01, *d* = .39, and if they required 3 or more letter cues (*M* = 37.02, *SE* = 2.77), *t*(59) = 3.41, *p* < .001, *d* = .48. However, there was not a reliable difference in the percentage of errors corrected if participants required 2 letter cues to generate the correct response or 3 or more letter cues, *t* < 1.

Memory Predictions

A one-way ANOVA was used to compare average JOLs provided for answers generated after 1, 2, 3, or 4 letter cues. Overall, JOLs differed based on the number of letter cues needed to generate the correct answer, *F*(3,165) = 97.13, *p* < .001, $\eta^2{}_p$= .64. When participants could generate the correct response after 1 letter cue (*M* = 39.04, *SE* = 2.82), they provided reliably greater JOLs than when they generated the correct response after 2 letter cues (*M* = 27.22, *SE* = 2.64), *t*(55) = 8.68, *p* < .001, *d* = .57, 3 letter cues (*M* = 20.86, *SE* = 2.33), *t*(55) = 10.99, p < .001, *d* = .91, and 4 letter cues (*M* = 13.35, *SE* = 2.26), *t*(55) = 11.95, *p* < .001, *d* = 1.31. On average, JOLs provided after 2 letter cues were reliably greater than JOLs after 3 letter cues, *t*(55) = 4.89, *p* < .001, *d* = .33, and JOLs after 4 letter cues *t*(55) = 8.39, *p* < .001, *d* = .74. Lastly, JOLs provided after 3 letter cues were reliably greater than JOLs provided after 4 letter cues, *t*(55) = 7.60, *p* < .001, *d* = .44.

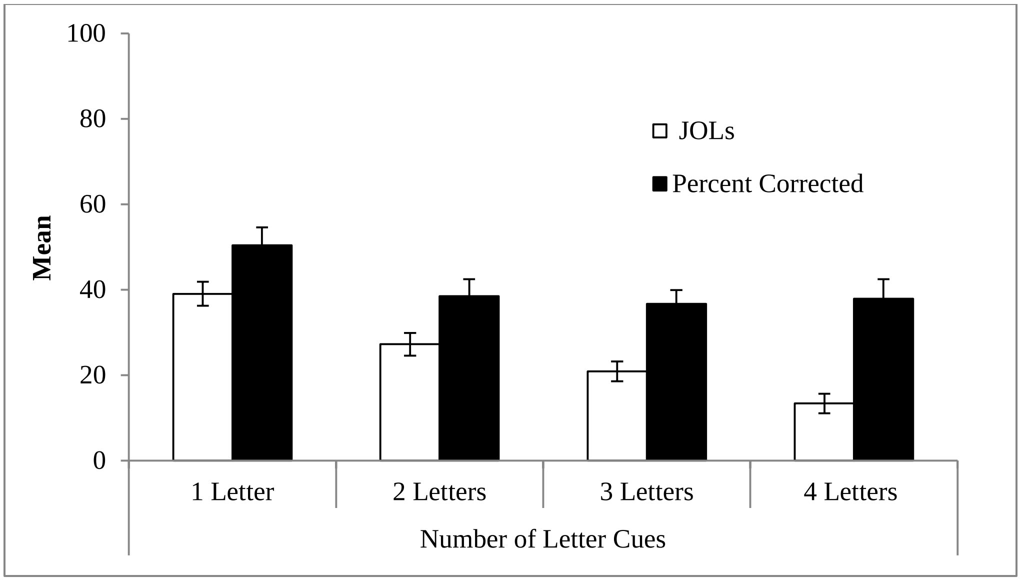

Figure 7. Average judgments of learning (JOLs) compared to the percent of error corrected based on the number of letter cues needed during feedback for Experiment 4. Error bars represent standard error of the mean.

Relative Accuracy. A gamma correlation was computed to compare JOLs during the initial test to final test accuracy. Overall, there was a reliable positive correlation between the JOLs participants provided during the initial test and the accuracy of items on the final test (G = .61, SE = .03), t(58) = 19.76, p < .001. The gamma correlation for Experiment 4 was also compared to the correlations from Experiment 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 for items in the feedback condition, F(4, 189) = 1.62, p = .17.

Gamma correlations were also computed to examine the relationship between JOLs and final test accuracy for items that were initially incorrect. Comparisons were made between items where the participant only required 1 or 2 letter cues to generate the correct response, and items where participants required 3 or more letter cues to generate the correct response. Overall, there was a positive correlation between JOLs and final test accuracy for items that required 1 or 2 letter cues (*G* = .24, *SE* = .07), *t*(53) = 3.39, *p* = .001, and for items that required 3 or 4 letter cues (*G* = .17, *SE* = .08), *t*(53) = 2.19, *p* =.03. However, these correlations did not reliably differ, *t* < 1.

Discussion

Consistent with previous research by Finn and Metcalfe (2010), participants were more likely to correct errors when they were able to generate the correct response after 1 cue compared with items that required multiple letter cues be shown before the correct response was generated. Mean 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. However, this 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) in the no feedback conditions, 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. Although participants' average JOLs are correctly predicting the pattern of results, item-by-item judgments did not more effectively differentiate between items that would or would will not be corrected on the later test compared with the previous three 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. Yet, after receiving feedback, participants are less able to differentiate between information they will and will not remember on a later test. Tests without feedback, however, lead to strong metacognitive accuracy when predicting performance on a future test. The goal of the current experiments was to explore beliefs about the influence of feedback on memory performance and examine methods that may enhance metacognitive accuracy following feedback.

In Experiment 1, participants learned Lithuanian-English word pairs and took an initial test. Some items received correct answer feedback whereas other items received right/wrong feedback or no feedback at all. Overall, the results showed that correct answer feedback led to the greatest improvement in memory performance on the final test, but decreased metacognitive accuracy compared with the right/wrong feedback and no feedback conditions. In line with Kornell and Rhodes (2013), after receiving feedback, participants were least accurate at differentiating between what they would and would not know on a future test. In terms of absolute accuracy, I anticipated that average JOLs would not differentiate between items in the feedback condition and items in the no feedback condition (replicating Kornell & Rhodes). Therefore, in order to distinguish between experience-based and theory-based JOLs, participants in the preJOL condition in Experiment 2 made judgments before seeing the item, with only information regarding the feedback condition available. This might allow participants to use more theory-based judgments as they would have few other cues available as the basis of their JOLs. Unexpectedly, in both the JOL conditions (Experiment 1 and 2) and the PreJOL condition, participants provided higher JOLs to items in the feedback condition compared with items in the no feedback condition. Thus, participants' knowledge of the benefits of feedback influenced the magnitude of JOLs. However, on an item-by-item level, participants were still unable to identify which errors they would correct on a later test.

Experiment 3 aimed to improve metacognitive accuracy by allowing participants to gain experience with the task. Although participants understand that feedback is beneficial for memory performance, they may have difficulty predicting the benefits of feedback without first experiencing the task. Thus, in Experiment 3, participants performed the task twice. Consistent with the previous experiments, participants provided higher JOLs to items in the feedback condition compared with the no feedback condition on List 1. However, gamma correlations remained reliably lower for items that received feedback compared with items that did not receive feedback. On list 2, neither relative nor absolute accuracy changed. Participants still predicted that they were more likely to remember items that received feedback compared with items that did not receive feedback, but remained less accurate when predicting the future accuracy of specific items following feedback.

Experiments 1 through 3 demonstrated that participants have a global understanding that feedback is beneficial for memory; however, they display difficulty identifying which items will benefit from feedback and which items will remain unlearned. Experiment 4 used scaffolded feedback in order to cue participants to errors from the initial test that were more likely to be corrected on a later test. If participants could generate the correct English translation after 1 or 2 letter cues, they would be more likely to correctly remember that item on a later test than if they required the majority of the word to generate the correct response. Overall, participants in Experiment 4 recalled a greater percentage of English translations on the final test if they could generate the correct response for errors after 1 letter cue compared to more letter cues. Their absolute judgments predicted this pattern with average JOLs negatively related to the number of cues needed. However, compared to the previous three experiments, item-by-item judgments did not more accurately discriminate between items that would be corrected on a later test and those that would not.

In sum, these experiments attempted to replicate and expand upon Kornell and Rhodes' (2013) findings that participants are unaware of the benefits of feedback. Using a within-subjects design, the current experiments suggest that participants do have a general understanding that feedback is beneficial for memory performance. Across experiments, participants corrected approximately 50% of their errors from test 1 to test 2. Average JOLs indicated that participants believed they were more likely to remember items in the feedback condition than items in the no feedback condition, but JOLs in the feedback condition were underestimates of how well they would perform on the final test. Despite demonstrating a global understanding of feedback, after receiving feedback on errors during the initial test, participants were unable to identify which errors were going to benefit from feedback on the final test.

Stability Bias

Previous research suggests that people view memory as stable over time, a finding referred to as the stability bias (Koriat et al., 2004; Kornell & Bjork, 2009; Kornell et al., 2011). Often, metamemory predictions fail to take into account future forgetting or future learning. In the current experiments, relative judgments displayed a stability bias, but absolute judgments did not. That is, participants predicted that feedback would differentially impact performance on the final test, but they were unable to accurately identify which items would specifically benefit from feedback.

One possible explanation is participants' judgments reflected the memory for past test heuristic (Finn & Metcalfe, 2007; 2008). That is, participants' may have used their most recent test performance as a basis for JOLs and failed account for factors that improve memory on a subsequent test (e.g., additional study). To explore this possibility, I computed gamma correlations between accuracy on test 1 and JOLs. A strong, positive correlation between accuracy on test 1 and JOLs would indicate that people are giving high JOLs to items they answered correctly and low JOLs to items they answered incorrectly. In the no feedback condition, past test performance is generally an appropriate cue for future test performance given that little change in learning would be anticipated. However, in the feedback condition, a strong correlation would indicate that participants are not incorporating the benefits of feedback into their judgments of future performance and basing JOLs on current recall. Overall, for both the no feedback and feedback conditions across all experiments, gamma correlations between test 1 accuracy and JOLs were above +.93 (see Table 4). Performance on test 1 was closely associated with subsequent JOLs regardless of whether feedback was provided. Thus, participants may be largely relying on their current test performance to predict later test performance, ignoring factors that may change memory performance (e.g., feedback)[3].

These results are similar to previous research (e.g., Kelley & Jacoby, 1996; Koriat et al., 2004; Kornell & Bjork, 2009) reporting dissociations between experience-based and theory-based JOLs. Theory-based JOLs are often based upon people's understanding of how their memory works, whereas experience-based JOLs are influenced by current processing of items (e.g., accuracy of a response on a test, how fluent information appears, etc.). As mentioned

[3] It is important to note that while memory for past test plays a large role in JOLs, previous research suggests that other factors may also influence predictions in addition to past test performance (Ariel & Dunlosky, 2011; Tauber & Rhodes, 2012b).

previously, when participants have multiple study-test cycles, they often base their judgments on their most recent experience with the item (whether or not they answered it correctly on the previous test, Finn & Metcalfe, 2007; 2008), failing to take into account learning that will occur due to an additional study opportunity. One’s current memory state is a salient cue (Koriat et al., 2004; Nelson & Dunlosky, 1991) and thus increases the likelihood that participants JOLs will be influenced by the current experience (experience-based JOLs).

Based on Finn and Metcalfe’s results (2007; 2008), one may mistakenly conclude that participants do not appreciate the benefits of further study opportunities. However, Kornell and Bjork (2009) demonstrated that participants may express different global beliefs of memory than their item-by-item judgments would suggest. After studying an item, Kornell and Bjork informed participants that they would have 1-4 more times to study that item, and then asked them to predict the likelihood they would remember that item on a later test. Item-by-item judgments did not differentiate between the number of study opportunities, suggesting that participants may not understand that they are more likely to remember an item if they study it four times compared with one time. However, when asked directly if they thought more study opportunities were beneficial for memory, participants reported that, overall, items given four study opportunities should be remembered better than items studied once (see also Kornell et al., 2011). Kornell and Bjork suggest that people have a basic understanding of how their memory works (e.g., more study opportunities leads to better memory performance), but those theories of memory need to be explicitly activated. These findings echo earlier work by Koriat et al. (2004) and Kelley and Jacoby (1996) demonstrating discrepancies between theory-based and experience-based predictions depending on how questions are framed or the type of knowledge participants possess before making judgments for themselves or others.

In the current experiments, absolute judgments appear to reflect participants' theories of feedback. Participants assigned higher JOLs to items in the feedback condition compared with items in the no feedback condition, indicating that they were more likely to remember items on a later test if they received feedback on an initial test. Unlike Kornell and Bjork (2009), these theories of feedback influenced memory predictions without explicit direction to focus on the feedback condition. The PreJOL condition in Experiment 2 was designed to specifically cue participants to attend to the feedback condition when making judgments. However, in the standard JOL conditions across all experiments, average JOLs were higher for items in the feedback condition compared with items not receiving feedback. Alternatively, item-by-item judgments appear to reflect experience-based processes related to recall failure or success on the first test but did not take into account the benefits of feedback. Thus, in the current experiments, participants' theories of memory (e.g., that feedback will lead to better performance) influenced where they anchored their judgments but did not influence the distribution of JOLs between correct and incorrect items, hindering relative accuracy.

Such data complement recent research by England and Serra (2012) examining underconfidence with practice. They demonstrated that telling participants that a task is going to be easy or hard changed where they anchored their JOLs (average JOLs were higher when participants were told the task was easy), but did not change their relative accuracy (gamma correlations did not differ between the two groups). That is, although theories of memory change where participants anchor on the scale, they do not appear to largely influence how judgments are distributed. Accordingly, for individual items in the feedback condition, participants may be less able to accurately monitor their knowledge and subsequently control their behavior.

Delayed JOLs

A great deal of research has focused on exploring methods that will lead to the most accurate predictions of future memory performance. The methodologies used in the no feedback condition mirror a related literature examining delayed JOLs (dJOLs). In experiments focusing on the accuracy of JOLs, participants generally study a list of word pairs (e.g. Dog- Apple) and make a JOL for each pair about the likelihood that they will remember the target word (e.g., Apple) in the pair when given the cue (e.g., Dog) on a later test. Some pairs receive immediate JOLs whereas JOLs are delayed for other items. When making JOLs, participants are sometimes shown both words in the pair and asked to make a JOL, or participants are only shown the cue. Finally, participants are given a test on the word pairs and relative accuracy is assessed by comparing JOLs at study to accuracy on a test. In sum, a large body of literature has shown that when dJOLs are made with only the cue present, they predict final test performance with higher levels of accuracy (gammas near +1.0) than immediate JOLs or conditions where both the cue and target are present (Dunlosky & Nelson, 1992). This benefit to relative accuracy is referred to as the delayed JOL effect (for a review, see Rhodes & Tauber, 2011b). According to the monitoring dual memories account (Nelson & Dunlosky, 1991), the delayed JOL condition allows participants to test their memory and discriminate between what they do and do not know. In the immediate JOL condition and conditions where both the cue and target are present, access to the target word may be too readily available and thus, not provide a true test of whether the correct target would be retrievable from memory at a later point in time.

In the current experiments, the delayed JOL effect is evident in the no feedback conditions. Participants were highly accurate at predicting future test performance when feedback was not provided. However, feedback decreased this accuracy. To my knowledge, Kornell and Rhodes (2013) have provided the only other test of delayed JOLs following correct answer feedback. The current results bolster Kornell and Rhodes' findings to further suggest that tests without feedback enhance metacognitive accuracy, consistent with the benefits of delaying JOLs. However, providing feedback can decrease metacognitive accuracy, eliminating the delayed JOL effect, while also enhancing memory performance. Although feedback is generally regarded as good for memory, it is also important to consider that it may have somewhat detrimental effects for metamemory and thus, potentially impair effective control of study behaviors.

Monitoring and Control Processes in Memory

A primary tenant of metamemory frameworks is that when people accurately understand their memory (monitoring) they can properly control their learning behaviors. These monitoring and control processes continually work together to assess and update memory (Koriat, 2007; Nelson, 1996; Nelson & Narens, 1990). In paradigms that allow participants to choose items for restudy, participants are likely to choose items deemed less well learned (regardless of actual learning; e.g., Rhodes & Castel, 2009). Thus, if participants are given control over their study, their memory performance should improve if they choose the correct information to restudy. For example, Kornell and Metcalfe (2006) had participants answer general knowledge questions, provided them with the correct response to the question, and then allowed participants to pick half of the items for restudy. The researchers either honored their choices and allowed participants to restudy the selected items or dishonored their choices and had participants restudy the items that were not chosen for restudy. Ultimately, memory performance was better when participants restudy choices were honored. These results suggest that participants were effectively monitor memory and choose appropriate items for restudy, enhancing memory performance. If participants can accurately differentiate between what they do and do not know (relative accuracy), then they will efficiently allocate their resources (e.g., spend more time studying unlearned material). However, learning will not be ideal if an individual's subjective assessment of memory differs from actual memory performance. Therefore, if participants can accurately predict how feedback will influence items on a test, then they will more effectively allocate their study time to errors on the initial test that will not be corrected by feedback alone.

Kornell and Rhodes (2013) demonstrated that not only were participants unable to predict the benefits of feedback, but they also did not make optimal restudy choices when allowed to select items they would like to study again. In various situations, such as educational settings, people are expected to learn material thoroughly and efficiently. For example, students often take multiple courses, are expected to do well on assessments in all courses, and have limited time. Some errors on an initial test will be corrected after feedback alone (e.g., 50% of errors in the current experiments were corrected following feedback). However, it is likely that some information has not been learned and feedback alone may not be enough to fix those errors. Thus, to efficiently learn the most information, students should focus their study efforts on items that will not be corrected due to feedback alone. Experiment 4 suggests that participants may have some ability to understand which errors are more likely to be corrected after feedback. On average, participants provided higher JOLs for items that were generated after 1 or 2 letter cues compared with items generated after 3 or 4 letter cues. If people can capitalize on this knowledge, they should make study choices that would focus their time on items that are unlikely to be corrected by feedback alone.

In order to improve memory, people must be able to both accurately monitor their current memory state and engage in effective control behaviors that will increase learning for unlearned information. Future work should examine how feedback influences participants' control behaviors. For example, the methods of Experiment 4 could be paired with a restudy option. Thus, after scaffolded feedback participants could have the option to pick items for restudy. If scaffolded feedback allows participants to better differentiate between what they do and do not know, then participants should choose to restudy items that required more letter cues to answer correctly than items that only required 1 or 2 letter cues. This would indicate that participants can use their theories of memory to effectively control behaviors and thus, improve memory for information that is not learned well.

Limitations and Future Directions

There are several limitations of the current experiments which could most likely be remedied with further research. The current experiments were intended to replicate and extend upon previous work by Kornell and Rhodes (2013) which used a between-subjects design. In contrast, the current experiments used a within-subjects design under the assumption that manipulating feedback within-subjects would increase participants' sensitivity to the effects of feedback on memory (e.g., Koriat et al., 2004). However, the current experiments did not directly compare a between-subjects condition with a within-subjects condition. Average JOLs were higher for feedback items than non-feedback items in the current experiments, yet using a between-subjects design, Kornell and Rhodes did not find a difference in average JOLs. Because the current experiments did not directly compare a within-subjects and a between-subjects design, it is uncertain whether using a within-subjects design increased sensitivity to the feedback manipulation, or whether the materials used in these experiments led to different judgments than the materials used by Kornell and Rhodes. This issue could be addressed with future research that implements a between-subjects version of Experiment 1.

Another limitation of the current experiments may have been the difficulty of the items used, specifically when considering the methods outlined in Experiment 4. The original Metcalfe and Finn study (2010) utilized scaffolded feedback with general knowledge questions. Therefore, their task most likely relied heavily on semantic memory, whereas the Lithuanian-English word pairs relied more on episodic memory and were relatively difficult to learn. In Experiment 4, participants' average JOLs were negatively related to the number of letter cues needed to generate the correct response. Judgments were reliably higher for items that needed fewer cues compared with items that needed more cues. However, error correction did not necessarily follow this pattern. Participants were more likely to correct errors if they could generate the correct response after 1 letter cue compared with more letter cues, but there was no difference in error correction for responses generated after 2 or more letter cues. This may be due to the difficulty of the task or it may be due to the fact that the majority of the correct responses were words with 4 letters or less. Thus, further work should examine the efficacy of scaffolded feedback for both semantic and episodic memory tasks.

In addition to scaffolded feedback, further research should explore other methods of feedback that are likely to differentially impact error correction. For example, Lhyle and Kulhavy (1987) presented correct answer feedback or scrambled feedback. In the scrambled feedback condition, the correct response was scrambled and participants had to unscramble the answer before moving on to the next item. Ultimately, participants were more likely to correct errors for items where they had to unscramble the correct response compared with items where the correct response was presented normally. Similar work on the generation effect has found that participants are more likely to remember the correct target word when they generate that word during study compared with when they just study the target and cue at the same time

(Jacoby, 1978; Slamecka & Graf, 1978). Further research should explore how participants monitor their memory performance following various forms of feedback. If certain types of feedback are more beneficial than others, can participants predict that? In Experiment 1, average JOLs were higher for items in the correct answer feedback condition compared with items in the right/wrong feedback condition, but it is not clear how predictions would differ for other forms of feedback (e.g., scrambled feedback). For example, Castel, Rhodes, and Friedman (2013) reported that participants' JOLs are sensitive to any form of generation, even when generating information is not helpful for memory performance. Thus, participants' JOLs may change for feedback conditions that require further generation or manipulation of information.

Across all four experiments, item-by-item JOLs did not effectively differentiate between errors that would be corrected on a later test and those that would not. When studying items individually, the influence of feedback may be hard to predict. Not only are participants trying to determine the fate of that specific item, but they are also trying to decide which items, out of all their errors, will be corrected and which ones will not. Thus, it may be beneficial for future research to explore different methods of eliciting metamemory predictions that would allow participants to consider all of the incorrect responses at once. For example, instead of asking participants to make item-by-item judgments, they might be more accurate when asked to identify which errors they will correct relative to the other errors. Participants could be shown all of their incorrect responses at once and asked to identify which items they think will be corrected on a later test. This may serve two purposes. First, participants could consider all their errors at once and make a relative decision regarding which ones may be corrected on a later test. Second, this procedure may also increase the saliency of feedback. It may remind participants that some of their errors will be corrected and then, hopefully, they would be able to use that information to provide more accurate predictions.

Conclusion

In sum, the current experiments replicate and expand upon previous research indicating that people are unaware of the benefits of feedback (Kornell & Rhodes, 2013). The four experiments demonstrate that people do 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 items will benefit from feedback. Because relative accuracy was diminished after feedback (compared with no feedback), this suggests that participants would not make ideal restudy choices if given the opportunity. Further work is needed to explore how feedback influences participants' behaviors following a test.

Table 1

Percentage of items correctly recalled across experiments.

	No Feedback				**Feedback**			
	Test 1	**Test 2**	**Retained**	**Corrected**	**Test 1**	**Test 2**	**Retained**	**Corrected**
Experiment 1	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)
Experiment 2								
JOL	25.40 (2.70)	27.30 (2.90)	90.89 (3.41)	5.83 (1.05)	27.94 (2.70)	60.64 (3.40)	90.10 (3.78)	51.27 (3.51)
PreJOL	25.87 (2.70)	29.20 (2.90)	96.49 (1.82)	6.70 (1.57)	26.83 (2.70)	56.51 (3.40)	91.42 (3.18)	45.85 (3.75)
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)

Note. Numbers in parentheses represent the standard error of the mean.

Table 2

Gamma correlations across experiments

	Gamma	
	No Feedback	**Feedback**
Experiment 1	.95 (.02)	.57 (.08)
Experiment 2		
JOL	.87 (.06)	.50 (.07)
PreJOL	-.15 (.10)	-.14 (.07)
Experiment 3		
List 1	.87 (.04)	.65 (.06)
List 2	.94 (.05)	.69 (.05)
Experiment 4		.66 (.03)

Note. Parentheses represent standard error of the mean.

Table 3

Number of letter cue in the words, number needed to generate the correct response, and percentage corrected in Experiment 4

	Number of Letters							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Number of Words in Each Letter Group	0	0	4	15	8	1	0	2
Average Number of Cues Needed to Generate Answer	3.69 (.23)	4.83 (.33)	7.57 (.41)	3.28 (26)	1.30 (.10)	0	0	0
Percent Corrected on Test 2	50.77 (4.16)	38.68 (3.85)	36.04 (3.07)	37.92 (4.50)	28.57 (18.44)	0	0	0

Note. Numbers in parentheses refer to the standard error of the mean.

Table 4

Mean gamma correlations between test 1 accuracy and JOLs.

	No Feedback	Feedback	t	p
Experiment 1	.97 (.05)	.94 (.02)	1.59	.12
Experiment 2				
JOL Condition	.96 (.01)	.99 (.04)	1.98	.06
Experiment 3				
List 1	.94 (.10)	.98 (.05)	.71	.48
List 2	.94 (.02)	.98 (.01)	2.52	.02
Experiment 4		.96 (.01)		

Note. The t score represents the t value for the paired t-test comparing gamma correlations for the feedback and no feedback conditions (*p* represents the p value for this test). The numbers within parentheses represent the standard error of the mean.

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.

Benjamin, A. S., & Bjork, R. A. (1996). Retrieval fluency as a metacognitive index. *Implicit Memory and Metacognition*, 309-338.

Benjamin, A. S., Bjork, R. A., & Schwartz, B. L. (1998). The mismeasure of memory: When retrieval fluency is misleading as a metamnemonic index. *Journal of Experimental Psychology: General*, *127*(1), 55.

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.

Butterfield, B., & Metcalfe, J. (2001). Errors committed with high confidence are hypercorrected. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*, 1491-1494.

Butterfield, B., & Metcalfe, J. (2006). The correction of errors committed with high confidence. *Metacognition Learning*, *1,* 69-84.

Bjork, R. A. (1975). Retrieval as a memory modifier. In R. Solso (Ed.), *Information processing and cognition: The Loyola Symposium* (pp. 123-144). Hillsdale, NJ: Lawrence Erlbaum Associates.

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.

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.

Castel, A. D., Rhodes, M. G., & Friedman, M. C. (2013). Predicting memory benefits in the production effect: The use and misuse of self-generated distinctive cues when making judgments of learning. *Memory & Cognition*, *41*(1), 28-35.

England, B. D., & Serra, M. J. (2012). The contributions of anchoring and past-test performance to the underconfidence-with-practice effect. *Psychonomic Bulletin & Review*, *19*(4), 715-722.

Fazio, L. K., Huelser, B. J., Johnson, A., & Marsh, E. J. (2010). Receiving right/wrong feedback: Consequences for learning. *Memory*, *18*(3), 335-350.

Fazio, L. K., & Marsh, E. J. (2009). Surprising feedback improves later memory. *Psychonomic Bulletin & Review*, *16*(1), 88-92.

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.

Hart, J. T. (1965). Memory and the feeling-of-knowing experience. *Journal of Educational Psychology*, *56*(4), 208.

Hartwig, M. K., & Dunlosky, J. (2012). Study strategies of college students: Are self-testing and scheduling related to achievement?. *Psychonomic Bulletin & Review*, *19*(1), 126-134.

Jacoby, L. L. (1978). On interpreting the effects of repetition: Solving a problem versus remembering a solution. *Journal Of Verbal Learning & Verbal Behavior*, *17*(6), 649-667.

Jacoby, L. L., & Kelley, C. M. (1987). Unconscious influences of memory for a prior event. *Personality and Social Psychology Bulletin*, *13*(3), 314-336.

Kelley, C. M., & Jacoby, L. L. (1996). Adult egocentrism: Subjective experience versus analytic bases for judgment. *Journal of Memory and Language*, *35*(2), 157-175.

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General*, *126*(4), 349.

Koriat, A. (2007). Metacognition and consciousness. In P. D. Zelazo, M. Moscovitch, & E. Thompson (Eds.), *The Cambridge Handbook of Consciousness* (pp. 289-325). Cambridge, UK: Cambridge University Press.

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.

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.

Kornell, N., & Bjork, R. A. (2007). The promise and perils of self-regulated study. *Psychonomic Bulletin & Review*, *14*(2), 219-224.

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.

Kornell, N., & Metcalfe, J. (2006). Study efficacy and the region of proximal learning framework. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(3), 609.

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.

Kulhavy, R. W., & Stock, W. A. (1989). Feedback in written instruction: The place of response certitude. *Educational Psychology Review, 1,* 279- 308.

Lhyle, K. G., & Kulhavy, R. W. (1987). Feedback processing and error correction. *Journal Of Educational Psychology*, *79*(3), 320-322.

Metcalfe, J., & Finn, B. (2011). People's hypercorrection of high-confidence errors: Did they know it all along?. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(2), 437.

Mueller, M. L., Tauber, S. K., & Dunlosky, J. (2013). Contributions of beliefs and processing fluency to the effect of relatedness on judgments of learning. *Psychonomic Bulletin & Review, 20*, 378-384.

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin*, *95*(1), 109.

Nelson, T. O. (1996). Consciousness and metacognition. *American Psychologist*, *51*(2), 102.

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.

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. *The Psychology of Learning and Motivation*, *26*, 125-141.

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.

Pressey, S. L. (1950). Development and appraisal of devices providing immediate automatic scoring of objective tests and concomitant self-instruction. *Journal of Psychology, 29,* 417-447.

Rhodes, M. G., & Castel, A. D. (2008). Memory predictions are influenced by perceptual information: evidence for metacognitive illusions. *Journal of Experimental Psychology: General*, *137*(4), 615.

Rhodes, M. G., & Castel, A. D. (2009). Metacognitive illusions for auditory information: Effects on monitoring and control. *Psychonomic Bulletin & Review,16*(3), 550-554.

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.

Rhodes, M. G., & Tauber, S. K. (2011b). The influence of delaying judgments of learning on metacognitive accuracy: a meta-analytic review. *Psychological Bulletin*, *137*(1), 131.

Sitzman, D. M., Rhodes, M. G., & Tauber, S. K. (in press) Prior knowledge is more predictive of error correction than subjective confidence. *Memory and Cognition.*

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

Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal Of Experimental Psychology: Human Learning And Memory*, *4*(6), 592-604.

Tauber, S. K., & Rhodes, M. G. (2012a). Measuring memory monitoring with judgments of retention (JORs). *The Quarterly Journal of Experimental Psychology*, *65*(7), 1376-1396.

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.