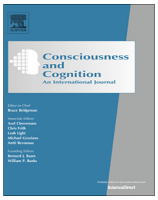
Consciousness and Cognition 29 (2014) 131–140

Contents lists available at ScienceDirect

Consciousness and Cognition

journal homepage: www.elsevier.com/locate/concog



Improving metacognitive accuracy: How failing to retrieve practice items reduces overconfidence

Tyler M. Miller [a,*], Lisa Geraci [b]

[a] *Department of Psychology, South Dakota State University, United States*
[b] *Department of Psychology, Texas A&M University, United States*

ARTICLE INFO

Article history:
Received 14 April 2014

Keywords:
Metacognition
Overconfidence
Retrieval practice

ABSTRACT

People often exhibit inaccurate metacognitive monitoring. For example, overconfidence occurs when people judge that they will remember more information on a future test then they actually do. The present experiments examined whether a small number of retrieval practice opportunities would improve participants' metacognitive accuracy by reducing overconfidence. Participants studied Lithuanian–English paired associates and predicted their performance on an upcoming memory test. Then they attempted to retrieve one or more practice items (or none in the control condition) and made a second prediction. Experiment 1 showed that failing to retrieve a single practice item lead to improved subsequent performance predictions – participants became less overconfident. Experiment 2 directly manipulated retrieval failure and showed that again failure to retrieve a single practice item significantly improved subsequent predictions, relative to when participants successfully retrieved the practice item. Finally, Experiment 3 showed that additional retrieval practice opportunities reduced overconfidence and improved prediction accuracy.

1. Introduction

The term metacognition refers to a person's knowledge about the quality and accuracy of their own cognition. Flavell (1979) formalized the term and investigated the phenomena from a developmental perspective. In the earliest studies, children were asked to study a list of items until they believed they had memorized them completely. The younger children said they had completely memorized the items and were ready for the test when they actually had not memorized the words (Flavell). Studies today show a similar pattern for college-aged participants. They predict that they will earn a much better score on a given test than they actually do (for one example see Hacker, Bol, Horgan, & Rakow, 2000). People are also overconfident in other abilities, including their driving ability (Knouse, Bagwell, Barkley, & Murphy, 2005), dating popularity (Preuss & Alicke, 2009), ability to complete projects before deadlines (Buehler, Griffin, & Ross, 1994), and even their gun-safety knowledge (Ehrlinger, Johnson, Banner, Dunning, & Kruger, 2008).

Although some have claimed that overconfidence is adaptive (Gramzow, Willard, & Mendes, 2008), a large literature suggests that there are costs to overconfidence and that being metacognitively accurate is useful (e.g., Thiede, Anderson, & Therriault, 2003; Thiede, Dunlosky, Griffin, & Wiley, 2005). Accurate metacognition is particularly beneficial in educational contexts. For example, more accurate metacognition is associated with better academic performance (Everson & Tobias, 1998) and memory performance (Thiede et al., 2003, 2005). Other studies have confirmed the link between accurate

[*] Corresponding author. Address: Department of Psychology, South Dakota State University, Brookings, SD 57007, United States.
E-mail address: tyler.miller@sdstate.edu (T.M. Miller).

http://dx.doi.org/10.1016/j.concog.2014.08.008
1053-8100/© 2014 Elsevier Inc. All rights reserved.

metacognition and improved performance (e.g., Nelson, Dunlosky, Graf, & Narens, 1994). This finding makes sense – students who can accurately determine which information they do not know can then focus their efforts on learning that information to perform well on an upcoming test. Students who do not know which information they know well enough for the test may be inefficient with their study time and may focus unnecessary efforts on already-learned-material. And they may stop studying prematurely, falsely believing that they understand material that they do not yet know well enough for the test.

Because accurate metacognition can lead to good memory performance, researchers have attempted to improve metacognitive accuracy. In many of these intervention studies, participants are encouraged to practice making performance predictions (Kelemen, Winningham, & Weaver, 2007) either with or without feedback (Lichtenstein & Fischoff, 1980; Miller & Geraci, 2011) and incentives for accuracy (Ehrlinger et al., 2008; Miller & Geraci, 2011). In the classroom, students have been given entire practice tests to help them prepare for the test with instructions to use their performance on the practice test to identify strengths and weaknesses in their understanding of the content (Hacker et al., 2000). In other cases, people have been asked to generate keywords after reading expository texts and asked to self-select texts for additional study (Thiede et al., 2003). Finally, people have been asked to engage in self-reflection in which they formulated plans to prepare for the next exam after they received feedback about the inaccuracy of their predictions (Hacker et al., 2000). While a few of these attempts have been modestly successful at improving metacognitive monitoring accuracy, the general theme of this research is that metacognitive ability is highly resistant to change. And when there is improvement, sometimes it only helps the students who need it the least; that is, it helps the highest performers (Kelemen et al., 2007; Hacker et al., 2000). The present study examined a different method for improving metacognitive accuracy. Here, we examined whether practice retrieving a single test item (Experiments 1 and 2) or more test items (Experiment 3) could lead to more accurate metacognitive monitoring. To examine this issue, we used a novel paradigm in which participants studied a list of paired-associates for a later memory test, made a performance prediction, attempted to retrieve one or more practice items, and then made a second performance prediction. The idea was that providing participants with direct experience answering even a single practice test item may improve the accuracy of their performance predictions. Experiment 1 examined whether a single practice opportunity would lead to more accurate memory performance predictions. Experiment 2 directly manipulated retrieval failure and to directly examine the role of retrieval failure on metacognition. Experiment 3 examined the effect of a mixed record of retrieval practice success and failure on participants' performance predictions and metacognitive accuracy.

1.1. The effects of testing on metacognition

A large body of research now demonstrates the memory advantages of taking a practice test versus restudying material prior to taking a final memory test (for reviews see Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013; Roediger, Putnam, & Smith, 2011). It has also been suggested that testing has other benefits. For example, testing produces better organization of knowledge (Zaromb & Roediger, 2010) and it reduces proactive interference (Szpunar, McDermott, & Roediger, 2008).

One can ask at least two questions regarding the relationship between testing and metacognition. One, are people aware of the benefits of testing and two, does testing improve metacognition? The literature is somewhat mixed with respect to the first question – are people aware of the benefits of testing. In one survey study that asked participants to free-report study strategies, results showed that students did not report testing themselves as a preferred study strategy; rather they reported non-testing strategies like rereading their notes or textbooks. The survey also included a question that forced respondents to choose a testing or restudying strategy. Again, restudying was the more popular choice of study strategy. Taken together, the results suggest that students have little to no awareness of the benefits of testing because they do not use it in practice (Karpicke, Butler, & Roediger, 2009). In contrast, other research indicates that people may be aware of the benefits of testing (Tullis, Finley, & Benjamin, 2013). In the Tullis et al. study, participants reported judgments of learning (JOL) for items that had been previously tested or restudied. In some conditions, the participants' JOLs indicated that they were aware of the mnemonic benefits of testing because participants believed tested items would be remembered more than restudied items. Participants in this study were even aware that the benefit of testing interacts with the retention interval because they assigned comparatively higher JOLs for re-studied items than tested items on an immediate test and vice versa for the delayed test – a result that suggests that participants may be aware of the memory advantages of tested material. Other work suggests that participants are aware of the benefits of testing (Kornell & Bjork, 2007). When given a choice of how to study items, participants start in a simple "presentation" mode, but then switch to a testing strategy.

Thus, the evidence regarding whether people have good metacognitive awareness of the benefits of testing is equivocal. What about the other question? Does testing improve students' metacognition? Thus far there is little direct evidence to answer this question. In one study that examined the effect of retrieval practice on performance and metacognitive accuracy, participants were assigned to one of four conditions where study and testing circumstances were manipulated (Karpicke & Roediger, 2008). During the learning phase, participants in the standard condition studied foreign language word pairs, were tested on the entire list, studied the entire list again, and were tested on the entire list again for a total of four study-test cycles. In the drop-out conditions, recalled items were treated differently – they were dropped from subsequent study sessions, dropped from subsequent test sessions, or dropped from both subsequent study and test sessions. Participants in the drop-out conditions also completed four study-test cycles. Following the learning phase, all participants made a metacognitive monitoring judgment – a prediction – about how they would perform on the memory test one week later. Results showed that participants in the standard learning condition and in the drop-out condition in which recalled items were dropped from study, but not from testing, had more than double the level of recall performance compared to the other learning conditions. However, performance predictions did not differ across conditions. Participants in all conditions predicted that they would remember about half the items, leading the authors to note that, "…students exhibited no awareness of the mnemonic effects of retrieval practice, as evidenced by the fact that they did not predict they would recall more if they had repeatedly recalled the list of vocabulary words than if they only recalled each word one time. (Karpicke & Roediger, 2008; p. 968). This finding indicates, again, that participants have little awareness of the mnemonic benefits of testing and it also suggests that testing had little effect on metacognitive accuracy. Because the Karpicke and Roediger paper was focused on the role of testing in affecting later memory performance, and not metacognition, the design of the study does not permit us to know whether testing improves metacognition because all participants in the study provided metacognitive judgments after retrieval practice had already occurred. There were no un-tested participants to use as a comparison. Thus, the goal of the current studies was to directly examine how practice tests can improve metacognition.

The now-classic experiments on delayed JOLs (dJOL) are relevant to this discussion on the effects of testing on metacognitive accuracy. In paired associate learning paradigms when participants make delayed JOLs rather than immediate JOLs, their relative accuracy is comparatively high (Nelson & Dunlosky, 1991). It has been suggested delaying JOLs leads to highly accurate monitoring because participants attempt to retrieve the target item (covertly) before they make the JOL, especially when only the cue is provided at JOL (Dunlosky & Nelson, 1992). Therefore, if the JOL and the covert retrieval attempt are made at some delay then it will be most diagnostic of memory performance on the test, which also happens at a delay, because it is not contaminated with the contents of short-term memory. This theoretical mechanism has been termed the monitoring dual memories hypothesis (Nelson & Dunlosky, 1991; also see Rhodes & Tauber, 2011 for a review). Thus, the improvement in JOL accuracy when the JOLs are made after a delay rather than immediately can be said to provide indirect evidence that retrieval practice improves metacognition.

1.2. The present experiments

The present experiments were designed to directly examine whether practice retrieving items can lead to more accurate metacognitive monitoring. In particular, we examined whether retrieval practice with a single item (Experiment 1) would improve subsequent performance predictions. We also examined the role of failure (vs. success) in retrieving a single practice item on subsequent judgments (Experiment 2) and how much failure is needed to improve the accuracy of participants' predictions (Experiment 3). In each of the three experiments, participants were asked to make global performance predictions before and after engaging in retrieval practice or no retrieval practice.

2. Experiment 1

2.1. Method and materials

2.1.1. Participants

Eighty young adults (53% male) between the ages of 18 and 21 ($M = 19.11$, $SE = .09$) participated in the experiment in return for partial course credit. Participants were undergraduate students with 12–16 years of education ($M = 13.14$, $SE = .10$). Across conditions, participants did not differ in average age or education (F's < 1.5).

2.1.2. Design

Experiment 1 used a 2 Condition (Retrieval Practice and No Retrieval Practice Control) × 2 Performance Prediction (Original and Adjusted) mixed randomized repeated model study design. Participants were randomly assigned to either a retrieval practice condition ($n = 40$) or a control (no retrieval practice control) condition ($n = 40$) and they made two performance predictions. The dependent variable of interest was the accuracy of each performance prediction. Accuracy was assessed by calculating a difference score (performance prediction minus recall performance) for each prediction. As such, a positive number indicates that the participant was overconfident.

2.1.3. Materials and procedure

We used a subset of 40 Lithuanian–English paired associates from norming data provided by Grimaldi, Pyc, and Rawson (2010). Items were chosen based on the probability they were recalled on the memory test during trial one in Grimaldi et al. (2010). The average probability of recall on trial one was .23, but items from the entire range were selected (.04–.49 probability of recall).

Each paired-associate was presented via computer for a fixed-duration of 10 s. After presentation of all 40 paired-associates, participants made a self-paced global performance prediction as a total number (0–40) of English equivalents they believed they would recall on the memory test. Following the first prediction, participants in the retrieval practice condition were given one practice item, the Lithuanian practice item "sesuo – ______", and were instructed to recall the English equivalent of the item by typing their answer. Participants did not receive feedback about the accuracy of their recall attempt. The Lithuanian–English paired associate "sesuo-sister" was chosen because Grimaldi et al. (2010) reported that participants recalled this item nearly 50% of the time after one study session. We selected this item so that we could also examine the influence that retrieval practice success or failure had on metacognitive accuracy in a supplementary analysis.

Following retrieval practice, participants were asked to make a second, adjusted performance prediction, based on their experience attempting to recall the English equivalent of "sesuo." Participants were told their adjusted performance prediction should be as accurate as possible and could go up, down, or stay the same. Participants in the no retrieval practice control condition simply made two consecutive predictions without the intervening retrieval practice item but heard the same instruction that their prediction could go up, down, or stay the same. Following the adjusted performance prediction, participants took the final recall test and completed a demographic questionnaire.

2.2. Experiment 1 results

We predicted an interaction whereby participants in the retrieval practice condition would have more accurate adjusted performance predictions compared to their original predictions and that participants in the no retrieval practice control condition would have equivalent original and adjusted predictions. We also predicted that recall performance on the memory test at the end of the experiment (not on the retrieval practice items) would be equivalent across conditions.

2.2.1. Metacognitive accuracy analysis

Participants in both conditions were slightly overconfident about their future memory performance in both their original and adjusted performance predictions and retrieval practice did not significantly influence those predictions. Results of the repeated measures ANOVA indicated non-significant main effects of performance prediction, $F(1,78) = 1.17$, $MSE = 0.01$, $p = .28$, $\eta^2{}_p = .02$ and condition, $F(1,78) = 0.39$, $MSE = 0.02$, $p = .53$, $\eta^2{}_p = .01$. Of interest was the interaction term, which was also non-significant, $F(1,78) = 0.94$, $MSE = 0.01$, $p = .34$, $\eta^2{}_p = .01$. A separate univariate ANOVA indicated equivalent recall performance between conditions ($F < 1$; see Table 1 for means and standard errors and Fig. 1).

2.2.2. Supplementary analysis

Participants in the Retrieval Practice condition recalled the English equivalent of the Lithuanian cue "sesuo" at a rate higher than was expected based on the previous literature (Grimaldi et al., 2010). That is, 70% of the participants recalled the English equivalent "sister". We completed a supplementary analysis to determine the influence of retrieval practice success or failure on adjusted performance predictions. We split the data for the Retrieval Practice condition into two groups: a group that correctly recalled the English equivalent and a second group that did not recall the English equivalent. Participants who correctly recalled the practice item did not change their performance predictions ($M = .30$, $SE = .03$ to $M = .31$, $SE = .04$) and their memory performance was $M = .28$ ($SE = .03$). In contrast, participants who failed to retrieve the practice item decreased their performance predictions ($M = .30$, $SE = .05$ to $M = .19$, $SE = .03$; $F(1,38) = 13.90$, $MSE = 0.01$, $p = .001$, $\eta^2{}_p = .27$), becoming more accurate about future memory performance given that their memory performance was $M = .21$ ($SE = .04$). In terms of metacognitive monitoring accuracy, retrieval practice failure affected participants' monitoring accuracy, whereas retrieval practice success did not. In particular, retrieval failure prior to the final memory test allowed participants to adjust their predictions. That is, after retrieval failure, participants reduced their predictions and became less overconfident. In Experiment 2, we directly manipulated retrieval success and failure to examine their effects on metacognitive monitoring accuracy.

3. Experiment 2

3.1. Method and materials

Results from Experiment 1 suggested that retrieval practice itself does not improve metacognitive accuracy – rather retrieval failure does. In Experiment 2, we directly examined the effect of retrieval success and failure on metacognitive monitoring accuracy. To do this, we manipulated the difficulty of the retrieval practice item. Again, all participants

Table 1
Means for retrieval practice performance, recall performance and performance predictions for Experiments 1, 2, and 3 expressed as a proportion of total items.

Condition	Retrieval practice	Recall	Original prediction	Adjusted prediction
Experiment 1				
Control	n/a	0.28	0.33	0.33
Easy RP	0.70	0.26	0.30	0.28
Experiment 2				
Easy RP	0.67	0.25	0.34	0.31
Difficult RP	0.27	0.28	0.34	0.26
Experiment 3				
Control	n/a	0.21	0.32	0.34
Easy RP	0.62	0.25	0.37	0.34
Medium RP	0.41	0.22	0.33	0.24
Difficult RP	0.13	0.26	0.33	0.19

Note: Standard errors for all recall and prediction values range from .01 to .03 and .02 to .07 for all retrieval practice values.

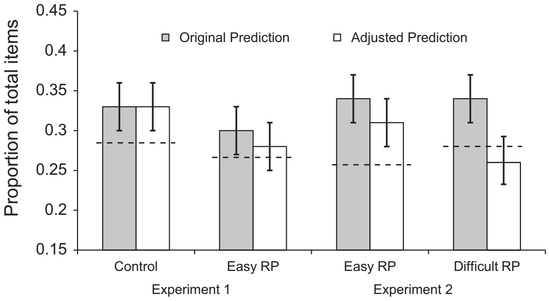


Fig. 1. Proportion of total items predicted for original and adjusted predictions among participants in Experiments 1 and 2. Dashed lines indicate mean recall performance for each condition. Error bars indicate standard error.

completed a one-item practice test, but the item that was practiced was either an easy or a difficult item (based on recall data provided in Grimaldi et al., 2010). The assumption was that more participants would fail to retrieve the difficult practice item than the easy practice item and we could test the hypothesis that retrieval practice failure is more beneficial for metacognitive accuracy than retrieval practice success.

3.1.1. Participants

Ninety younger adults (52% male) between the ages of 18 and 29 ($M = 19.74$, $SE = .21$) participated in return for partial course credit. Participants were undergraduate students with between 12 and 16 years of education ($M = 13.74$, $SE = .11$). Across conditions, participants did not differ in average age or education (Fs < 1.5).

3.1.2. Design

Experiment 2 used a 2 Condition (Easy Retrieval Practice and Difficult Retrieval Practice) × 2 Performance Prediction (Original and Adjusted) mixed randomized repeated model. Participants were randomly assigned to either the Easy ($n = 45$) or the Difficult retrieval practice condition ($n = 45$) and made two performance predictions. The dependent variable of interest was the accuracy of each performance prediction. Accuracy was assessed by calculating a difference score (performance prediction minus recall performance).

3.1.3. Materials and procedure

We used the same 40 Lithuanian–English paired associates as Experiment 1 from Grimaldi et al. (2010). Participants in the Easy Retrieval Practice condition were asked to recall the English equivalent of "sesuo" (sister) as before and participants in the Difficult Retrieval Practice condition were asked to recall the English equivalent of "muilas" (soap). Previous work showed that participants recalled the English equivalent of "sesuo" 50% of the time and the English equivalent of "muilas" 4% of the time after one study session (Grimaldi et al., 2010). Note that participants were not told that they had received an "easy" or a "difficult" retrieval practice item. The procedure for Experiment 2 was nearly identical to the procedure for Experiment 1 with the key exception that all participants in Experiment 2 attempted to retrieve one item (either an easy or a difficult item) between their original and adjusted performance predictions.

3.2. Experiment 2 results

Based on the supplementary analysis of Experiment 1 we predicted that participants in the Difficult Retrieval Practice condition would have more accurate performance predictions following retrieval practice than participants in the Easy Retrieval Practice condition. We also predicted that participants in the two conditions would have similar recall performance.

3.2.1. Metacognitive accuracy analysis

First, we examined performance on the practice item. Results showed that participants in the Easy Retrieval Practice condition correctly recalled the English equivalent of the Lithuanian cue "sesuo" 67% of the time (a 33% failure level), whereas those in the Difficult Retrieval Practice condition correctly recalled the English equivalent of the Lithuanian cue "muilas" 27% of the time (a 73% failure level), $F(1,88) = 16.85$, $MSE = 0.21$, $p < .001$, $\eta^2_p = .16$. Turning to the effect of retrieval practice difficulty on prediction accuracy, results showed that while the original performance predictions for participants in both conditions were nearly identical, participants in the Difficult Retrieval Practice condition adjusted their performance predictions downward more than participants in the Easy Retrieval Practice condition, $F(1,88) = 4.03$, $MSE = 0.03$, $p < .05$, $\eta^2_p = .04$. Participants who attempted to recall the difficult practice item significantly improved their metacognitive accuracy while participants who attempted to recall the easy practice item did not improve their metacognitive accuracy. The results also indicated a main effect of performance prediction, $F(1,88) = 1.17$, $MSE = 0.10$, $p < .001$, $\eta^2_p = .13$, showing an overall increase in accuracy, but no main effect of condition, $F(1,88) = 1.98$, $MSE = 0.13$, $p = .16$, $\eta^2_p = .02$. A separate univariate ANOVA indicated equivalent recall performance between conditions ($F < 1$; see Table 1, Fig. 1).

Experiment 2 showed that participants in the Difficult Retrieval Practice condition adjusted their predictions more than participants in the Easy Retrieval Practice condition. Thus, retrieval failure improved metacognitive accuracy. Furthermore, Experiment 2 confirmed that a single instance of retrieval practice failure was beneficial for metacognitive accuracy.

However, the results from Experiment 2 do not explain *why* retrieval practice failure is better for metacognitive accuracy than retrieval success. It has been suggested that practice items must be diagnostic of the final memory test (i.e., the *diagnosticity assumption*; Dunlosky, Rawson, & McDonald, 2002). Given that participants were overconfident about their performance, one could infer that the difficult retrieval practice item was diagnostic of the memory test; therefore participants had practice with a test item that was representative, or diagnostic, of the final test. Thus, the difficult test condition benefitted participants' metacognitive accuracy.

While it is true that, on average, participants in the Difficult Retrieval Practice condition experienced failure at a similar level on the practice test as they did on the final test, individual participant's accuracy during retrieval practice was experienced a bit differently. For each participant, performance on the practice item was either 100% or 0% – each participant correctly recalled the retrieval practice item or failed to retrieve the practice item. To examine whether taking a difficult practice test improved predictions because the test was more diagnostic of the final test than was an easy practice test, we examined the effect of multiple instances of retrieval practice on metacognitive accuracy in Experiment 3. In particular, we examined how a mixed record of success and failure with retrieval practice would influence the accuracy of subsequent performance predictions.

4. Experiment 3

4.1. Method and materials

Given the results from Experiments 1 and 2, showing that retrieval practice failure is more beneficial than retrieval practice success for improving metacognitive accuracy, one might predict that a higher "dose" of a failure would be more beneficial for metacognitive accuracy than a lower dose. Alternatively, one might predict that too much retrieval practice failure would reduce participant's performance predictions excessively, leading to metacognitive errors of underconfidence. A third prediction based on the *diagnosticity assumption* (Dunlosky et al., 2002) would lead one to care less about the overall amount of success or failure of retrieval practice but rather, how similar performance was on the retrieval practice test and the final memory test. The prediction would be that the more similar the performance on the practice items was to performance on the final test the better participants' metacognitive accuracy would be. Experiment 3 was designed to tease apart these potential explanations for why retrieval failure improved predictions.

4.1.1. Participants

Two hundred and seventy-four younger adults (26% male) between the ages of 17 and 28 ($M = 18.85$, $SE = .10$) participated in return for partial course credit. Participants were undergraduate students with between 12 and 16 years of education ($M = 12.50$, $SE = .05$). Participants did not differ in average age or education across conditions (F's < 1.0).

4.1.2. Design

Experiment 3 used a 4 Condition (No Retrieval Practice Control, Easy Retrieval Practice, Medium Retrieval Practice, Difficult Retrieval Practice Difficult) × 2 Performance Prediction (Original and Adjusted) mixed randomized repeated model. Participants were randomly assigned to either the No Retrieval Practice Control ($n = 68$), Easy ($n = 69$), Medium ($n = 70$), or difficult retrieval practice ($n = 67$) and made two performance predictions. The dependent variable of interest was the accuracy of each performance prediction. Accuracy was assessed by calculating a difference score (performance prediction minus recall performance).

4.1.3. Materials and procedure

The materials and procedure were nearly identical to the previous experiments with the key exceptions being that participants in the retrieval practice conditions attempted to recall four practice items, rather than one practice item (as was the case in Experiments 1 and 2). The four items were all easy items, all difficult items, or a combination of easy and difficulty items to make up the medium difficulty condition according to norms reported in Grimaldi et al. (2010). All participants studied the 40 Lithuanian–English paired associates and made an original performance prediction. Participants in the retrieval practice conditions then attempted to recall four practice items. Then all participants made a second, adjusted performance prediction with the instructions that the prediction could increase, decrease, or stay the same, but should be as accurate as possible.

4.2. Experiment 3 results

4.2.1. Metacognitive accuracy analysis

First, retrieval practice performance depended on the condition – easy, medium, or difficult ($F(3,270) = 124.13$, $MSE = 0.04$, $p < .001$, $\eta^2 = .58$). Participants in the Easy Retrieval practice condition had the most retrieval practice success

($M = 0.62$) followed by participants in the Medium Retrieval Practice condition ($M = 0.41$), and finally, participants in the Difficult Retrieval Practice had the least amount of success ($M = 0.13$). Results from the repeated measures ANOVA indicated a significant interaction, $F(3,270) = 28.32$, $MSE = 0.17$, $p < .001$, $\eta^2_p = .24$, see Fig. 2, as well as significant main effects of performance prediction, $F(1,270) = 80.11$, $MSE = 0.48$, $p < .001$, $\eta^2_p = .23$, and condition, $F(3,270) = 8.87$, $MSE = 0.43$, $p < .001$, $\eta^2_p = .09$. Follow-up simple main effect analyses indicated that participants' original performance predictions were nearly equivalent ($F < 1.5$) and that participants' adjusted predictions were significantly different between conditions, $F(3,270) = 22.80$, $MSE = 0.56$, $p < .001$, $\eta^2_p = .20$. Tukey post hoc tests indicated that participants in the Medium and Difficult Retrieval Practice conditions had adjusted predictions that were significantly different than all other conditions. Participants in the Easy Retrieval Practice condition had adjusted predictions that were not significantly different than the adjusted predictions of Control condition participants.

Dunlosky et al. (2002) argued that the most effective type of retrieval practice would be one that yields similar performance as the final test (i.e., the *diagnosticity assumption*). The most similar performance between retrieval practice and memory test performance occurred in the Difficult Retrieval Practice condition. On average, participants in this condition correctly retrieved less than one practice item (0.13; see Table 1) and correctly retrieved approximately a quarter of the items on the final test (0.26). Yet, participants in the Difficult Retrieval Practice condition did not have the most accurate adjusted performance predictions ($M = .19$. Rather, participants in the Medium Retrieval Practice condition had the most accurate adjusted performance predictions ($M = .24$) despite the fact that their practice test performance was considerably higher (0.41) than final test performance (0.22).

The findings from Experiment 3 support the findings from Experiment 2; specifically, participants who experienced retrieval failure on a practice test changed their performance predictions the most. However, in terms of benefits to metacognitive accuracy, there appears to be a point of diminishing returns for retrieval practice failure. Participants in the Difficult Retrieval Practice condition, who experienced the most retrieval practice failure, revised their performance predictions too much – becoming underconfident about future memory performance and less accurate than participants in the Medium Retrieval Practice condition. That is, the metacognitive error of adjusted performance predictions was greater for participants in the Difficult condition compared to participants in the Medium condition. Thus, it appears that a moderate amount of failure on a practice test is necessary to improve metacognitive accuracy.

5. General discussion

The present experiments examined how testing influences metacognition in a multiple-prediction paradigm in which participants make performance predictions before and after retrieval practice. The results from Experiment 1 indicated that retrieval practice alone did not improve participants' adjusted, or second, performance predictions. Only participants who had failed to retrieve the practice test item showed improved metacognitive accuracy because these participants decreased their performance predictions. Participants who had successfully retrieved the practice item did not change their performance prediction. In Experiment 2 retrieval practice difficulty was manipulated and results showed that participants who were given a difficult practice item that they tended to fail to retrieve had more accurate performance predictions than participants who were given an easy practice item that they tended to get correct. Results confirmed that retrieval failure and not simply retrieval practice led to improved metacognitive accuracy by reducing performance predictions. Interestingly, when participants failed to retrieve the one-item practice test, they did not adjust their prediction downward by one-item, which might be expected. Rather, they adjusted their predictions downward by as many as 4 items in both Experiments 1 and 2. Results from Experiment 3 confirmed the previous findings – participants who failed to retrieve the most practice items adjusted their predictions the most. Experiment 3 also indicated that there is a point where too much failure lead participants to adjust their predictions excessively – with participants moving from being overconfident to being underconfident.

An innovation of the present set of experiments is that changes in metacognitive reports can be examined directly by asking participants to make more than one performance prediction following a relatively short run of practice items. In the

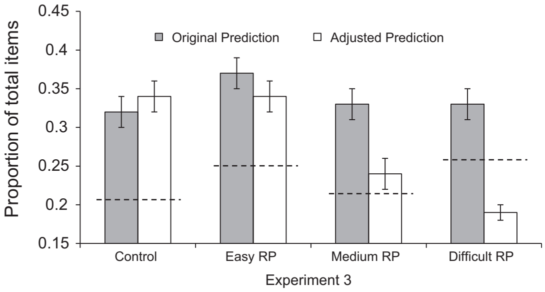


Fig. 2. Proportion of total items predicted for original and adjusted predictions among participants in Experiment 3. Dashed lines indicate mean recall performance for each condition. Error bars indicate standard error.

present experiments participants made performance predictions before and after retrieval practice. Requiring participants to make multiple performance predictions in a repeated measures design could uncover other factors that lead to metacognitive improvements. This method could be used in experiments that seek to improve participants' metacognitive accuracy after incentives, other practice, feedback, and other interpolated events.

The current findings are similar to research on the underconfidence-with-practice effect, because in both cases, participants are less overconfident after a second judgment. For example, in studies showing an underconfidence-with-practice effect, participants study a list of items and make item-level JOLs, and then are given a cued-recall memory test. Immediately following these events, the study-JOL-test cycle is repeated. However, unlike the current experiments, in studies showing an under-confidence-with-practice effect, participants become less overconfident (or more underconfident) with repeated trials because they underestimate how much they are learning during practice (Koriat, Sheffer, & Ma'yan, 2002). Participants' memory performance increases from Trial 1 to Trial 2, but their JOLs only increased slightly (e.g., Kornell & Bjork, 2009). Thus, in the under-confidence-with-practice research, participants' learning changes (increases) and predictions are relatively stable. However, in the current experiments participants' predictions changed but their learning stayed the same (because they only had one chance to study the paired associates). Thus, while the end results (improved metacognitive accuracy) are similar, the mechanism and the procedure are quite different across the two paradigms. The current experiments demonstrate that metacognition can be improved following a handful of retrieval failures, and even following a single retrieval failure of a non-studied item.

The reason participants adjusted or maintained their performance prediction after retrieval practice cannot be determined unequivocally from the present experiments. Both performance predictions (original and adjusted) are made based on analytic and/or non-analytic inferential processes (or experience-based and/or theory-based) (Kelley & Jacoby, 1996; Koriat, Bjork, Sheffer, & Bar, 2004). Because the reported experiments were not designed to evaluate the degree to which analytic and non-analytic processes were at work, the following discussion is speculative. Participants in the present experiment may have adjusted their performance predictions based on analytic inferential processes, meaning that they used their beliefs or theories about the factors that influence memory performance to inform their predictions. For example, they may have thought to themselves "I missed two of the four retrieval practice items. If I didn't know the English equivalents during retrieval practice, then I won't know them on the final test. I should adjust my performance prediction downward by two items." Participants may have also used nonanalytic inferential processes and changed their predictions based on their subjective experiences while attempting to recall the practice. Retrieval fluency is one nonanalytic process that may influence participants to change or maintain their performance prediction (Benjamin, Bjork, & Schwartz, 1998). If the solution to the retrieval practice item came to the participant quickly (it was highly fluent), then participant's subjective experience with the test was one of ease of retrieval. Ease of retrieval could lead the participant to increase their test performance prediction. If the solution to the practice item did not come to the participant quickly, or at all, (it was less fluent), this could lead the participant to decrease the test performance prediction.

Previous work that has investigated these two classes of inferential processes – analytic and nonanalytic – and has suggested that, while both processes could influence participants' predictions, analytic processes exert more influence on JOLs (Matvey, Dunloksy, & Guttentag, 2001). The results from Experiments 2 and 3 in the present experiments may reflect both classes of inferential processes influencing performance predictions. One might expect that if analytic inferential processes were the only process that influenced participants' adjusted performance predictions, then participants would adjust their performance prediction downward by exactly the same number of items they failed to retrieve during practice. For example, the participant would think to themselves "I don't remember this item right now so it is likely I won't remember this item later during the memory test," and downgrade their prediction by one item. But nonanalytic processes, or subjective experience, may also be influencing a participant's performance prediction such that participants think to themselves "retrieving this practice item was difficult and I can expect all of the items to be more difficult to retrieve on the memory test than I originally thought," and downgrade their prediction much more than one item. The relative contribution of both analytic and nonanalytic inferential processes for performance predictions will need to be further examined in future research.

One alternative hypothesis for the present pattern of results is that any success or failure experience influences metacognitive monitoring judgments. In other words, it might be that participants did not gain additional information from their retrieval practice per se, but that the outcome itself of either success of failure influenced them to change their judgments. As such, even an unrelated failure experience would lead participants to decrease their performance predictions. Geraci and Miller (2013) demonstrated that previous success or failure on an unrelated task influenced subsequent memory performance for older adults. Participants who experienced task success had superior memory performance compared to participants who experienced no prior task experience or prior task failure. In addition, prior task success appears to increase memory predictions for older adults (Geraci, Hughes, Miller, & De Forrest, submitted for publication). So, we think that it is certainly possible that failing at a task or a portion of a task, could reduce participants' overall confidence in their abilities and lead them to lower their subsequent performance predictions.

6. Conclusions

The present experiments showed that retrieval practice failure had significant benefits for monitoring accuracy, but future research should examine whether retrieval success and failure influences metacognitive control. Nelson and

Narens' (1990) model of metacognition, in which monitoring processes inform control processes, would predict that improved monitoring accuracy would lead to more effective control of future study. The relationship between more accurate monitoring and improved control processes has been observed in correlational and experimental studies on the topic (e.g., Nelson et al., 1994; Thiede, 1999; Thiede et al., 2003). In the present experiment, if participants were given the opportunity to restudy, a norm-of-study, and sufficient motivation to improve their performance, one would predict that students who failed to retrieve the practice item would make restudy decisions that would lead to improved performance. In contrast, participants who successfully retrieved the practice items may choose to study less, which could be a less than optimal decision. Note that because retrieval practice failure engendered underconfidence among participants in Experiments 1–2 and in the condition that produced the most failure (difficult retrieval practice) in Experiment 3, there could be an ideal amount of failure necessary for participants to not only improve calibration but also to guide them to the best possible study decisions.

The benefits of accurate metacognitive monitoring cannot be understated; accurate monitoring is associated with better performance outcomes in ecologically valid contexts (Everson & Tobias, 1998) and causally related to better performance outcomes in laboratory situations (Thiede et al., 2003). But often methods to improve participants' prediction accuracy have had limited success. Results from the present experiments indicate that retrieval practice failure is one factor that has a significant positive impact on participants' metacognitive monitoring and is a useful method to improve metacognitive accuracy. A final message that can be gleaned from the present experiments is that if students are going to take a practice test, they should take a difficult one. Taking a difficult practice test will reduce overconfidence and could lead to additional study efforts. Further research will identify the ideal amount of retrieval practice failure for improving prediction accuracy, but for now, erring on the side of more difficult practice tests may be a prudent decision.

References

Benjamin, A. S., Bjork, R. A., & Schwartz, B. L. (1998). The mismeasure of memory: When retrieval fluency is misleading as metamnemonic index. *Journal of Experimental Psychology: General, 127*, 55–68.

Buehler, R., Griffin, D., & Ross, M. (1994). Exploring the ''planning fallacy'': Why people underestimate their task completion times. *Journal of Personality and Social Psychology, 67*, 366–381.

Dunlosky, J., & Nelson, T. O. (1992). Importance of the kind of cue for judgments of learning (JOL) and the delayed-JOL effect. *Memory and Cognition, 20*, 374–380.

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising direction from cognitive and educational psychology. *Psychological Science in the Public Interest, 14*, 4–58.

Dunlosky, J., Rawson, K. A., & McDonald, S. L. (2002). Influence of practice tests on the accuracy of predicting memory performance for paired associates, sentences, and text material. In T. J. Perfect & B. L. Schwartz (Eds.), *Applied metacognition* (pp. 68–92). Cambridge, UK: Cambridge University Press.

Ehrlinger, J., Johnson, K., Banner, M., Dunning, D., & Kruger, J. (2008). Why the unskilled are unaware: Further exploration of (absent self-insight) among the incompetent. *Organizational Behavior and Human Decision Processes, 105*, 98–121.

Everson, H. T., & Tobias, S. (1998). The ability to estimate knowledge and performance in college: A metacognitive analysis. *Instructional Science, 26*, 65–79.

Flavell, J. H. (1979). Metacognition and cognitive monitoring. *American Psychologist, 34*, 906–911.

Geraci, L., Hughes, M.L., Miller, T.M., De Forrest, R., 2014. The effect of prior task success on older adults' memory performance: Examining the influence of different types of task success (submitted for publication).

Geraci, L., & Miller, T. M. (2013). Improving older adults' memory performance using prior task success. *Psychology and Aging, 28*, 340–345.

Gramzow, R. H., Willard, G., & Mendes, W. B. (2008). Big tales and cool heads: Academic exaggeration is related to cardiac vagal reactivity. *Emotion, 8*, 138–144.

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*, 634–642.

Hacker, D. J., Bol, L., Horgan, D., & Rakow, E. A. (2000). Test prediction and performance in a classroom context. *Journal of Educational Psychology, 92*, 160–170.

Karpicke, J. D., Butler, A. C., & Roediger, H. L. (2009). Metacognitive strategies in student learning: Do students practise retrieval when they study on their own? *Memory, 17*, 471–479.

Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrieval on learning. *Science, 319*, 966–968.

Kelemen, W. L., Winningham, R. G., & Weaver, C. A. III, (2007). Repeated testing sessions and scholastic aptitude in college students' metacognitive accuracy. *European Journal of Cognitive Psychology, 19*, 689–717.

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

Knouse, L. E., Bagwell, C. L., Barkley, R. A., & Murphy, K. R. (2005). Accuracy of self-evaluation in adults with ADHD: Evidence from a driving study. *Journal of Attention Disorders, 8*, 221–234.

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

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

Kornell, N., & Bjork, R. A. (2007). The promise and perils of self-regulated study. *Psychonomic Bulletin and Review, 14*, 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*, 449–468.

Lichtenstein, S., & Fischoff, B. (1980). Training for calibration. *Organizational Behavior and Human Performance, 26*, 149–171.

Matvey, G., Dunloksy, J., & Guttentag, R. (2001). Fluency of retrieval at study affects judgments of learning (JOLs): An analytic or non-analytic basis for JOLs? *Memory and Cognition, 29*, 222–233.

Miller, T. M., & Geraci, L. (2011). Training metacognition in the classroom: How incentives and feedback influence exam predictions. *Metacognition and Learning, 6*, 303–314.

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

Nelson, T. O., Dunlosky, J., Graf, A., & Narens, L. (1994). Utilization of metacognitive judgments in the allocation of study during multitrial learning. *Psychological Science, 5*, 207–213.

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. H. Bower (Ed.). *The psychology of learning and motivation* (Vol. 26, pp. 125–173). New York: Academic Press.

Preuss, G. S., & Alicke, M. D. (2009). Everybody loves me: Self-evaluations and metaperceptions of dating popularity. *Personality and Social Psychology Bulletin, 35*, 937–950.

Rhodes, M. G., & Tauber, S. K. (2011). The influence of delaying judgments of learning on metacognitive accuracy: A meta-analytic review. *Psychological Bulletin, 137*, 131–148.

Roediger, H. L., Putnam, A. L., & Smith, M. A. (2011). Ten benefits of testing and their applications to educational practice. In J. Mestre & B. Ross (Eds.), *Psychology of learning and motivation: Cognition in education* (pp. 1–36). Oxford: Elsevier.

Szpunar, K. K., McDermott, K. B., & Roediger, H. L. (2008). Testing during study insulates against the buildup of proactive interference. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*, 1392–1399.

Thiede, K. W. (1999). The importance of monitoring and self-regulation during multitrial learning. *Psychonomic Bulletin and Review, 6*, 662–667.

Thiede, K. W., Anderson, M. C. M., & Therriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. *Journal of Educational Psychology, 95*, 66–73.

Thiede, K. W., Dunlosky, J., Griffin, T. D., & Wiley, J. (2005). Understanding the delayed-keyword effect on metacomprehension accuracy. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*, 1267–1280.

Tullis, J. G., Finley, J. R., & Benjamin, A. S. (2013). Metacognition of the testing effect: Guiding learners to predict the benefits of retrieval. *Memory and Cognition, 41*, 429–442.

Zaromb, F. M., & Roediger, H. L. (2010). The testing effect in free recall is associated with enhanced organizational processes. *Memory and Cognition, 38*, 995–1008.