PsyCh Journal

INSTITUTE OF PSYCHOLOGY
Chinese Academy of Sciences

PsyCh Journal **10** (2021): 751–766
DOI: 10.1002/pchj.481

Effect of feedback type on enhancing subsequent memory: Interaction with initial correctness and confidence level

Lingwei Wang, and Jiongjiong Yang

School of Psychological and Cognitive Sciences and Beijing Key Laboratory of Behavior and Mental Health, Peking University, Beijing, China

Abstract: Feedback is an important factor to enhance subsequent memory, showing that memory performance increases after the feedback than after the no feedback condition during retrieval practice. However, most studies have provided answers as feedback and only examined memory accuracy. It is unclear whether memory is enhanced over time when other types of feedback (e.g., correct/incorrect) is given. In addition, during retrieval practice, participants' responses differ in correctness and confidence level. To what extent these initial memory features interact with feedback type to influence subsequent memory accuracy and confidence level remains unclear. In this study, to address these questions, participants learned a series of sentences, then during the retrieval practice phase, different types of feedback—feedback with correct/incorrect and answer (CA-feedback), feedback with answer (A-feedback), feedback with correct/incorrect (C-feedback), and no feedback—were given after they performed a cued-recall test and rated the confidence. After retention intervals of 5 min, 1 day, and 7 days, they took final tests, followed by the confidence rating. The results showed that different types of feedback influenced subsequent memory and forgetting by different mechanisms. The CA-feedback and A-feedback enhanced memory performance by correcting initial errors and increasing the confidence of correct trials, but the corrected memory was more easily forgotten from 5 min to 7 days. Compared to A-feedback, the CA-feedback maintained the corrected memory after 1 day. The C-feedback did not correct initial errors but slowed the forgetting rate and reduced the confidence of incorrect trials. This study highlighted the interaction between feedback type and initial memory features (correctness, confidence) to influence subsequent memory performance, including memory accuracy and confidence level.

Keywords: confidence; episodic memory; feedback; forgetting; testing effect

Correspondence Prof. Jiongjiong Yang, School of Psychological and Cognitive Sciences, Peking University, Beijing 100871, China. Email: yangjj@pku.edu.cn

Received 27 October 2020. Accepted 26 July 2021.

Many studies have shown that retrieval practice is an efficient way to enhance memory performance. The information that was practiced beforehand is finally better remembered than that which was only restudied. This phenomenon is called the "testing effect" (Kornell & Vaughn, 2016; Roediger & Butler, 2011; Roediger & Karpicke, 2006; Rowland, 2014). The testing effect has been identified steadily with different experimental materials (e.g., word lists, pairs, pictures, and proses) and different test forms (e.g., free recall, paired association, and recognition) (Roediger & Butler, 2011; Roediger & Karpicke, 2006; Rowland, 2014). An important feature of the testing effect is that the longer the delay time (e.g., days or weeks later), the higher the testing effect is (Roediger & Karpicke, 2006; Spitzer, 1939; Wheeler, Ewers, & Buonanno, 2003). Compared to the restudy condition, the retrieval practice induces stronger reactivation of the target item (Antony, Ferreira, Norman, & Wimber, 2017; Dudai, 2004; Roediger & Butler, 2011) and triggers the reconsolidation process, which leads to the enhancement of memory representations (Carpenter, 2011; Roediger & Butler, 2011) and the connections with contextual information (Karpicke, Lehman, &

Aue, 2014; Kornell & Vaughn, 2016), especially after a longer delay.

Feedback is one of the important factors to enhance the testing effect. Compared to the condition of no feedback, giving feedback during retrieval practice further improves subsequent memory accuracy, and this improvement is maintained for 2 to 7 days (Butler, Karpicke, & Roediger, 2008; Butler & Roediger, 2008; Pashler, Cepeda, Wixted, & Rohrer, 2005). In addition, feedback can correct the individual's confidence of memory in the final test. The confidence level is a metacognitive index to reflect individuals' cognition of their own memory abilities and the vividness of recalling memory contents (Chua, Schacter, & Sperling, 2008; Dunlosky & Metcalfe, 2009; Yonelinas, 1994). People usually have high confidence for correct responses and low confidence for incorrect responses. The metacognition of memory is biased if individual's confidence for the correct response is low (Roediger, Wheeler, & Rajaram, 1993) or the confidence for the incorrect response is high (Butterfield & Metcalfe, 2001, 2006). Studies have shown that feedback can correct metacognitive bias (Butler et al., 2008; Roediger & Butler, 2011). For example, in a study by Butler et al. (2008), after participants responded to the general knowledge questions and rated their confidence, they received feedback with correct answers or no feedback. Two days later, they took the final test and again made a confidence rating for every question. The results showed that memory accuracy was significantly improved after answer feedback in the final test, and the correlation between the final confidence and final accuracy was high. This suggests that answer feedback helps participants better discriminate between correct and incorrect responses and improves confidence level.

Previous studies have mainly applied feedback with correct answers (A-feedback) (Butler et al., 2008; for a review, see Butler & Woodward, 2018). Knowing whether the responses are or are not correct is another type of feedback, but there is no consensus on whether correct/incorrect feedback (C-feedback) has similar effects to feedback with answer (A-feedback) (Fazio, Huelser, Johnson, & Marsh, 2010; Marsh, Lozito, Umanath, Bjork, & Bjork, 2012; Pashler et al., 2005), especially after longer intervals. For example, after participants learned a series of Luganda–English pairs, they performed a cued-recall practice with A-feedback, C-feedback, or no feedback (Pashler et al., 2005). The results showed that in the final test on the same day and 7 days later, compared to the no-feedback condition, memory accuracy was significantly improved after A-feedback, but not after C-feedback. Differently, Marsh et al. (2012) asked participants to answer multiple-choice questions, then gave A-feedback, C-feedback, or no feedback. After 5 min and 2 days, the participants took the final short-answer test. The results showed that memory accuracy was improved by both A-feedback and C-feedback. Thus, it is necessary to clarify whether the effects of A-feedback and C-feedback differ on subsequent memory. In addition, it is unclear whether presenting CA-feedback simultaneously would further enhance memory performance. Clarifying this question is important because it helps us understand whether the effects of different feedback types could be additive to influence subsequent memory over time.

To explore the role of feedback type in the testing effect, one important factor is whether the responses are correct during the retrieval practice phase (i.e., initial response) (Butler et al., 2008; Fazio et al., 2010; Marsh et al., 2012; Pashler et al., 2005). The correctness of the initial response reflects memory strength of a specific stimulus (Halamish & Bjork, 2011; Kornell, Bjork, & Garcia, 2011). If the initial response is correct, the A-/CA-feedback and C-feedback should have similar effects on subsequent memory enhancement. However, if the initial response is incorrect, the A-/CA-feedback could provide correct answers, but C-feedback could not, which may lead to different effects on subsequent memory. Current studies have inconsistent findings on this issue. For example, Pashler et al. (2005) showed that the final accuracy of initially correct trials was not significantly different between feedback and no-feedback conditions. However, the final accuracy of initially incorrect trials was significantly improved after A-feedback rather than that after C-feedback and no-feedback. In contrast, in a study by Butler et al. (2008), after choosing an answer to a general knowledge question, participants received A-feedback or no feedback. The results showed that not only the accuracy of initially incorrect trials but also the accuracy of initially correct trials was improved after A-feedback. In addition, some studies have found that errors that were corrected after A-feedback gradually recovered as incorrect after 7 days (e.g., Butler, Fazio, & Marsh, 2011; Metcalfe & Miele, 2014), but it is unknown whether the recovery also happens after other feedback types. Therefore, it is necessary to differentiate the trials that are initially responded to correctly and incorrectly, and examine their memory performance after different types of feedback.

In addition to initial correctness, another factor that should be taken into consideration is the confidence of responses in the retrieval practice phase (Butler et al., 2008; Fazio et al., 2010). Only a few studies have considered this factor and have found that the effect of feedback was influenced by confidence level during the retrieval practice. For example, Butler et al. (2008) found that compared to the no-feedback condition, the accuracy of initially correct trials with low confidence was significantly improved after A-feedback whereas that with high confidence remained relatively stable. Then in a study by Fazio et al. (2010), after participants learned a series of Luganda–English pairs, they performed a cued-recall practice, rated the confidence, and received A-feedback, C-feedback, or no feedback. Their results showed that compared to the no-feedback condition, the accuracy of initially correct trials with low confidence was improved after A-feedback and C-feedback whereas that with high confidence remained similar. Note that although some studies had required participants to rate their confidence in the final test, to our knowledge, none of them had analyzed the confidence in the final test, so it is unknown whether feedback type and confidence level during retrieval practice play roles in the final confidence.

In sum, the main objective of this study was to explore to what extent the initial correctness and confidence level influence the effect of feedback type on subsequent memory performance. Participants learned a series of sentences that described features of unfamiliar objects. Then during the retrieval practice phase, participants were asked to recall by cues and rate their confidence, followed by different types of feedback, including CA-feedback, A-feedback, C-feedback, and no feedback. The final tests were performed 5 mins, 1 day and 7 days later. According to correctness and confidence level during the retrieval practice phase, the initial trials were divided into four levels: initially incorrect trials with low or high confidence, and initially correct trials with low or high confidence. Both memory accuracy and confidence in the final test were analyzed. Furthermore, forgetting rates after different types of feedback were estimated with memory accuracy at 5 min as baseline, and both forgetting rates of 5 min to 1 day and 5 min to 7 days were analyzed. It is generally accepted that memory consolidation goes through different stages (Dudai et al., 2015; Moscovitch, Cabeza, Winocur, & Nadel, 2016; Walker et al., 2003). The initial consolidation occurs within 5 min to 6 hr after memory acquisition, during which the interference of competitive memory is resisted. The second stage of consolidation occurs at the first night of sleep, during which the memory is protected against subsequent interference or decay and restored. After that, memory enters a longer period of stabilization. Therefore, three different retention intervals of 5 min, 1 day, and 7 days are used to determine memory performance at different consolidation stages and have been widely applied in studies of retrieval practice (for reviews, see Roediger & Karpicke, 2006; Rowland, 2014) and feedback (e.g., Pashler et al., 2005).

We hypothesized that feedback type would interact with initial correctness and confidence level to influence subsequent memory accuracy and confidence. When the initial trials are correct, different types of feedback have similar effects on memory performance. When the initial trials are incorrect, different types of feedback would influence subsequent memory by different mechanisms. As feedback provides correct answers, memory accuracy is increased by correcting errors after A-feedback and CA-feedback (Butler et al., 2008; Marsh et al., 2012; Pashler et al., 2005). In this case, the descriptions are reactivated during retrieval practice and have to be recombined with the corrected information to form a unitized episode. However, as the corrected information is only presented once and does not undergo a sufficient reconsolidation process (Dudai, 2012), the corrected information would recover to be incorrect after 7 days (Butler et al., 2011), and memory accuracy after A-feedback and CA-feedback declines significantly, which leads to a significant decrease in overall memory accuracy after A-feedback and CA-feedback. In contrast, although C-feedback has little effect on correcting errors, the overall memory accuracy is mainly derived from the initially correct trials, which would be maintained over time, leading to less forgetting after C-feedback. As confidence represents memory vividness and recollection, when A-feedback and CA-feedback provide the correct information, the final confidence of initially correct trials with low confidence would increase. At the same time, as C-feedback provides the information of correctness, the final confidence of initially correct trials with low confidence would increase, and the final confidence of initially incorrect trials with high confidence would decrease.

Materials and methods

Participants

Twenty-six participants (8 males; *M* age = 21.69 ± 2.43 years) were recruited for the study. The overall sample size was based on a prior power analysis (G*Power 3.1.9.6; University of Kiel, Germany). To obtain adequate power (i.e.,

$\alpha = .05$, $1 - \beta = .95$) and detect a moderate effect size (i.e., $f = .25$) for the interaction of feedback type (4) and retention interval (3), we would need a total sample of at least 18 participants for the experiment. All participants were native Chinese speakers, were right-handed, and gave written informed consent in accordance with the procedures and protocols, which were approved by the Review Board of School of Psychological and Cognitive Sciences, Peking University.

Materials

Two within-subjects factors were included in the study: type of feedback (CA-feedback, A-feedback, C-feedback, no feedback), and retention interval (5 min, 1 day, 7 days).

Forty-eight sentences that described 48 unfamiliar words were used as materials (Chen et al., 2018). To confirm that the words were unfamiliar and that the familiarity was matched across conditions, 19 participants (11 males; *M* age = 22.6 ± 2.58 years) who did not participate in the experiment rated to what extent they were familiar with the words (l = *most unfamiliar*, 7 = *most familiar*). The mean word familiarity was 2.93 ± 0.65.

The unfamiliar words were used to generate sentences that described their features. Each sentence contained the name of the category to which the word belongs, two perceptual feature descriptions, and two functional feature descriptions (Table 1). Based on standards applied in the study by McRae et al. (2005), the perceptual features were defined as information that can be seen or perceived such as color, shape, and odor, and the functional features were defined as information related to perceptual-irrelevant features such as their usage and location. Position of the perceptual and functional descriptions within the sentence was counterbalanced across the sentences. For example, the sentence that describes the word "*skunk*" was as follows: "*The skunk is a species of animal. The color of its fur is black and white. Its ears are short and round. It is a nocturnal animal. It has a lifespan of about 20 years*." Each sentence during the study contained 36.23 ± 4.22 Chinese characters (including punctuation), and the average length for each description was 8.25 ± 1.64 characters.

Table 1
Examples of the Sentences and Descriptions in Different Phases

Sentences and Descriptions in Different Phases
Study phase
The skunk is a species of animal. The color of its fur is black and white. Its ears are short and round. It is a nocturnal animal. It has a lifespan of about 20 years.
Retrieval practice phase
The color of skunk's fur is ( ).
The skunk's ears are ( ).
The skunk is a ( ) animal.
The skunk has a lifespan of about ( ) years.
Answer feedback
The color of skunk's fur is black and white.
The skunk's ears are short and round.
The skunk is a nocturnal animal.
The skunk has a lifespan of about 20 years.
Final test phase
The color of skunk's fur is ( ).
The skunk's ears are ( ).
The skunk is a ( ) animal.
The skunk has a lifespan of about ( ) years.

Each sentence was divided into four short descriptions to be used for the retrieval practice and final test phases. For each description, one keyword was omitted. For example, the descriptions of "*skunk*" to be retrieved were: "*The color of skunk's fur is ( ).*" "*The skunk's ears are ( ).*" "*The skunk is a ( ) animal.*" "*The skunk has a lifespan of about ( ) years*." The mean logarithmic frequency (Friederic & Frisch, 2000) for the keywords was 8.48 ± 0.51, and the mean word length was 1.85 ± 0.46 characters.

In addition, to control for the baseline level of the cue-recall test, another 18 participants (6 males; *M* age = 22.67 ± 1.53 years) who did not participate in the experiment filled in the blanks of the sentences without the study phase. The average baseline accuracy was 0.18 ± 0.11.

In total, 48 sentences were randomly divided into four groups to be used as materials for four types of feedback. Each group was then randomly divided into three sets to be used as materials for three retention intervals. Each sentence was divided into four descriptions during retrieval practice (48 total for each feedback type). Thus, each condition (e.g., CA-feedback at 5 min) had 16 descriptions for analysis, which was similar to previous studies (e.g., Butler et al., 2008; Butler & Roediger, 2008; Fazio et al., 2010; Pashler et al., 2005). The four groups and three sets had no significant differences in average baseline accuracy and various lexical-semantic features such as word familiarity, frequency and number of strokes for the words, frequency and word length of the keywords, and sentence length, *p*s > .500. The groups and sets were counterbalanced; thus, each group had an equal opportunity of being used at different types of feedback, and each set had an equal opportunity of being used at different retention intervals.

Procedure

The experiment included three phases: study, retrieval practice, and final test. During the study phase, participants learned all the sentences. They were presented with each of the 48 sentences for 20 s, during which they read the sentence silently and tried to remember it (Figure 1). After the sentence disappeared, they rated to what extent they could remember the sentence on a scale of 1 (*not at all*) to 6 (*completely*). The sentences were presented in a pseudo-random order so that no more than three sentences under the same condition were continuously presented to reduce order effect and fatigue effect (e.g., Butler et al., 2008; Fazio et al., 2010; Metcalfe & Finn, 2012). After the study phase, participants were asked to perform a distractor task for 2 min.

During the retrieval practice phase, each sentence was divided into four descriptions. The category description was not presented. Each description was presented for 5 s with one keyword omitted (Table 1). The participants were asked to recall the omitted keyword, give an oral report, and rate the confidence of their response ranging from 1 (*lowest*) to 6 (*highest*). Then each type of feedback was presented for 3 s. For CA-feedback, according to the participant's response, a "correct" or an "incorrect" feedback and the complete description were presented simultaneously. For A-feedback, the complete description was presented regardless of whether the participant's response was correct or incorrect. For C-feedback, according to the participant's response, a "correct" or an "incorrect" feedback was presented. For no feedback, only a fixation point was presented. The descriptions for each feedback type were presented in each of four blocks, and the four blocks were presented in a Latin square design. The four feedback-type blocks were

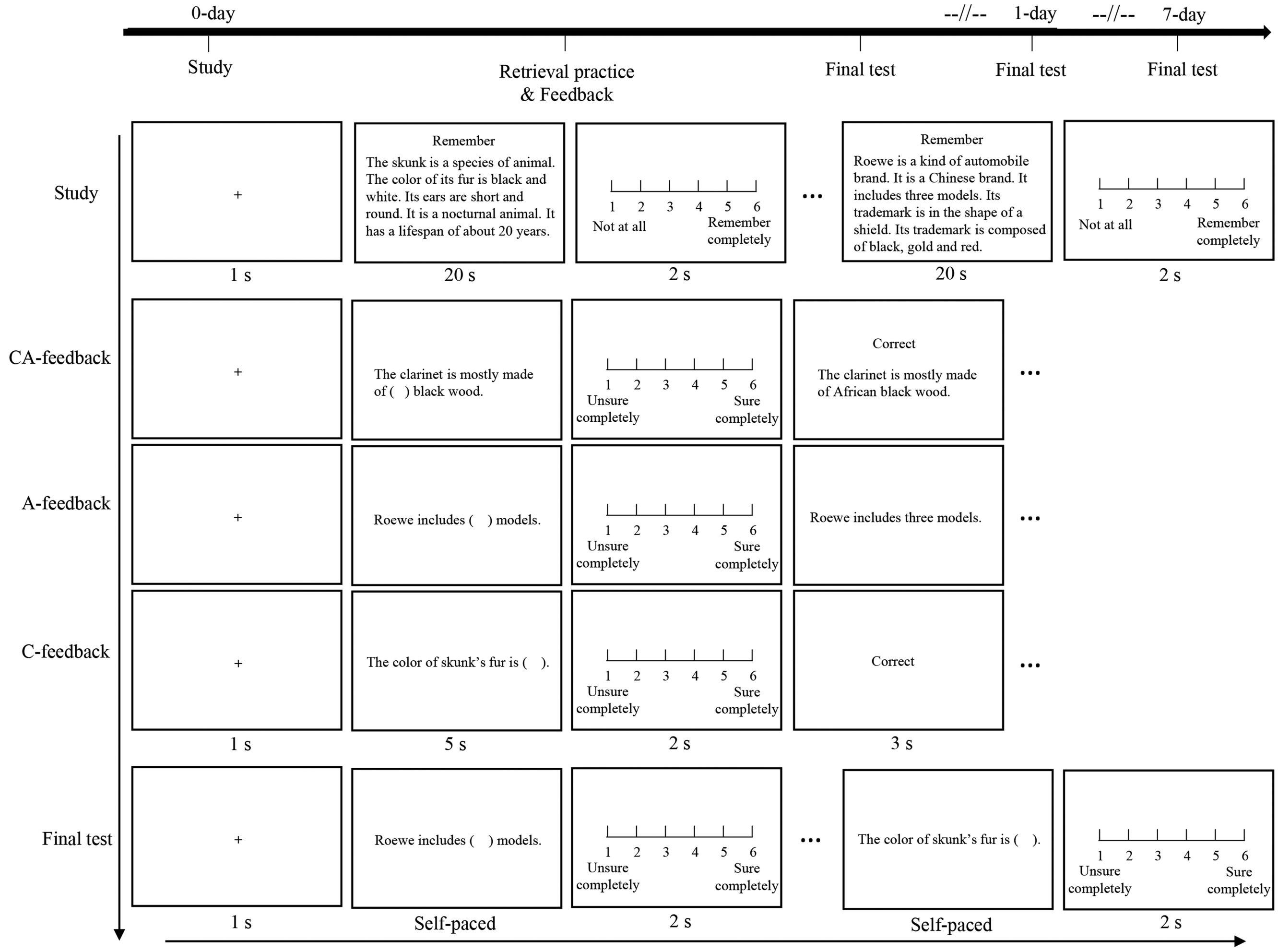
0-day
--//-- 1-day
--//-- 7-day
Study
Retrieval practice & Feedback
Final test
Final test
Final test
Study
CA-feedback
A-feedback
C-feedback
Final test
+
Remember
The skunk is a species of animal. The color of its fur is black and white. Its ears are short and round. It is a nocturnal animal. It has a lifespan of about 20 years.
1 2 3 4 5 6
Not at all
Remember completely
...
Remember
Roewe is a kind of automobile brand. It is a Chinese brand. It includes three models. Its trademark is in the shape of a shield. Its trademark is composed of black, gold and red.
1 2 3 4 5 6
Not at all
Remember completely
1 s
20 s
2 s
20 s
2 s
+
The clarinet is mostly made of ( ) black wood.
1 2 3 4 5 6
Unsure completely
Sure completely
Correct
The clarinet is mostly made of African black wood.
...
+
Roewe includes ( ) models.
1 2 3 4 5 6
Unsure completely
Sure completely
Roewe includes three models.
...
+
The color of skunk's fur is ( ).
1 2 3 4 5 6
Unsure completely
Sure completely
Correct
...
1 s
5 s
2 s
3 s
+
Roewe includes ( ) models.
1 2 3 4 5 6
Unsure completely
Sure completely
...
The color of skunk's fur is ( ).
1 2 3 4 5 6
Unsure completely
Sure completely
1 s
Self-paced
2 s
Self-paced
2 s

Figure 1. Procedure of the study, practice, and final test phases in the experiment. During the study phase, participants were presented with sentences that describe unfamiliar words, and were asked to read and remember them. During the retrieval practice phase, the participants were asked to recall the missing keywords, then rate their confidence. Four types of feedback were presented in blocks. During the final test phase, the participants were asked to recall the missing keywords, then rate their confidence again. The Chinese sentences were translated into English for illustration purposes.

counterbalanced across descriptions and participants. Descriptions within the same feedback type were presented in a pseudorandom order so that no more than three descriptions under the same condition were continuously presented. After each feedback-type block, participants were asked to perform a distractor task for 2 min.

After all the feedback blocks, participants performed final cued-recall tests at intervals of 5 min, 1 day, and 7 days. During each final test phase, the procedure was similar to that during the retrieval practice phase. Each description was presented with one keyword omitted, and participants recalled and orally reported the omitted keyword. Then they rated the confidence of their response. Different from that during the retrieval practice phase, no feedback was given. In each interval, there were 64 descriptions, and the four descriptions for the same word (e.g., "skunk") were presented continuously. The descriptions for different words were presented in a pseudorandom order so that no more than three descriptions under the same condition were continuously presented.

The participants had separate opportunities to practice study, retrieval, and final test trials before the formal phases.

Data analysis

Accuracy was calculated as the proportion of correct responses to the descriptions of the total descriptions at each retention interval and for each feedback type. The initial accuracy in the retrieval practice was analyzed using repeated measures analysis of variance (ANOVA), with feedback type (AC-feedback, A-feedback, C-feedback, no feedback) as a within-subjects factor. The initial confidence value in the retrieval practice was analyzed using repeated measures ANOVA, with feedback type and initial correctness (correct, incorrect) as within-subjects factors. The final accuracy was analyzed using repeated measures ANOVA, with feedback type and retention interval (5 min, 1 day, 7 days) as within-subjects factors. To control for the influence of the first test accuracy, the forgetting rates within 1 and 7 days were calculated as (accuracy at 5 min – accuracy at 1 day)/(accuracy at 5 min) and (accuracy at 5 min – accuracy at 7 days)/(accuracy at 5 min), and were analyzed using repeated measures ANOVA, with feedback type as a within-subjects factor. The forgetting rates also were compared with chance level (0) to determine whether memory was significantly forgotten at 1 day or at 7 days.

To clarify the interaction of feedback type, initial correctness, and retention interval for memory accuracy, the final accuracy was analyzed using repeated measures ANOVA, with feedback type, initial correctness in the retrieval practice (initially incorrect, initially correct), and retention interval as within-subjects factors. The forgetting rate was also analyzed using repeated measures ANOVA, with feedback type and initial correctness as within-subjects factors.

To clarify the interaction of feedback type and initial performance for the final memory, the confidence of the initial trials first was divided into two levels (1–3 as low confidence and 4–6 as high confidence). The initial trials then were divided into four levels (initially incorrect trials with low or high confidence and initially correct trials with low or high confidence). The retention interval was not used as an independent variable in this analysis because at each interval, some levels of initial trials had a small number of items, especially for the levels of initially incorrect trials with high confidence and initially correct trials with low confidence (Butler et al., 2008; Fazio et al., 2010). The data at three intervals were thus combined and analyzed. The final high-confidence accuracy (Butler et al., 2008) and final confidence were analyzed separately using repeated measures ANOVA, with feedback type, initial correctness, and initial confidence (low, high) as within-subjects factors.

One participant did not return for the 7-day final test, so his data were excluded. Another participant's confidence levels were not recorded in the 5-min final test due to a program failure, so his data of confidence were excluded. After dividing the initial correctness and confidence, there were some missing values. When there is a missing value in one condition, the whole data of the participant had to be excluded in the ANOVA analysis, which leads to a smaller sample size and weaker analysis power. Therefore, we used expectation maximization imputation to replace the missing values (Rashid & Gupta, 2019; Schafer & Graham, 2002) to ensure that a sufficient number of participants was included in the ANOVA analysis. The average number of missing values was 1.7 for initially incorrect trials with high confidence, 1.5 for initially correct trials with low confidence when memory accuracy was analyzed, and 3 for initially correct trials with low confidence when the final correct confidence was analyzed (see supplementary materials, Tables S1–S3). The number of missing values was similar to that in previous studies (Butler et al., 2008;

Fazio et al., 2010). There were no missing data for the other two levels (i.e., initially incorrect trials with low confidence and initially correct trials with high confidence). Note that the results were similar when the participants with missing values were excluded from analysis (see supplementary materials). To estimate the effect size of each analysis, η_p^2 was calculated. Post hoc pairwise comparisons were Bonferroni-corrected. $p < .050$, two tailed.

Results

Initial performance during retrieval practice

For the accuracy of retrieval practice, the ANOVA with feedback type as a factor showed that there was no significant effect of feedback type, $F(3,\ 72) = 2.03$, $p = .117$, $\eta_p^2 = .08$ (Figure 2A). Further analysis also showed that there was no significant difference between either of the two feedback conditions, *p*s > .200. For the confidence value of retrieval practice, the ANOVA with feedback type and initial correctness as factors showed that there was only a significant effect of initial correctness, $F(1,\ 25) = 253.65$, $p < .001$, $\eta_p^2 = .91$ (Figure 2B), showing that the confidence value of initially correct trials was higher than that of initially incorrect trials. There was no significant effect of feedback type, $F(3,\ 75) = 1.04$, $p = .380$, $\eta_p^2 = .04$, or the interaction, $F(3,\ 75) = 0.21$, $p = .889$, $\eta_p^2 = .01$. Further analysis also showed that there was no significant difference between either of the two feedback conditions, *p*s > .400. The results suggested that the initial accuracy and confidence value of retrieval practice were well-matched before feedback.

Interaction of feedback type and retention interval for remembering and forgetting

For the accuracy of the final test, the ANOVA with feedback type and retention interval as factors showed that there was a significant effect of feedback type, $F(3,\ 72) = 51.31$, $p < .001$, $\eta_p^2 = .68$. The accuracy after CA-feedback and A-feedback was significantly higher than that after C-feedback and no feedback, *p*s < .001. The accuracy after C-feedback was also higher than that after no feedback, $p = .058$. The results suggested that compared to the no-feedback condition, different types of feedback improved memory performance. In addition, the accuracy decreased significantly over time, $F(2,\ 48) = 87.94$, $p < .001$, $\eta_p^2 = .79$, and there was a significant interaction between feedback type and interval, $F(6,\ 144) = 8.07$, $p < .001$, $\eta_p^2 = .25$ (Figure 3A). Further analysis showed that the accuracy decreased significantly from 1 day to 7 days after different feedback conditions, *p*s < .001, but it did not decrease significantly from 5 min to 1 day after CA-feedback and C-feedback conditions, *p*s > .600. The results suggested that feedback with correct/incorrect information made memory forgotten less at 1 day.

The effect of C-feedback would be underestimated due to stronger effects of CA-feedback and A-feedback. Therefore, we did another repeated measures ANOVA with only C-feedback and no feedback included in the levels of feedback type. The results showed that the accuracy after

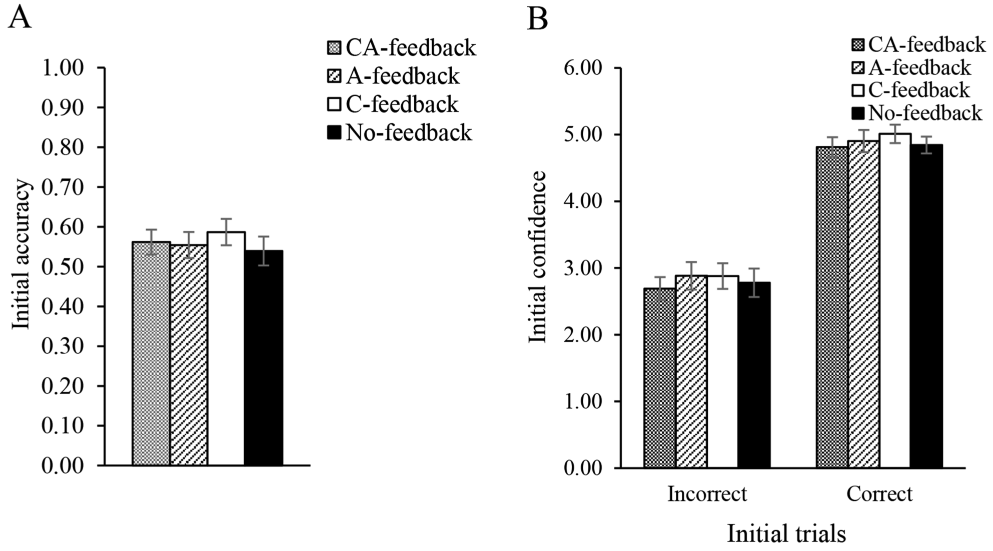

Figure 2. Results of initial accuracy (A) and confidence (B) in the retrieval practice phase. There was no significant difference between either condition for initial accuracy and initial confidence. The error bars represent the *SE*s of the means.

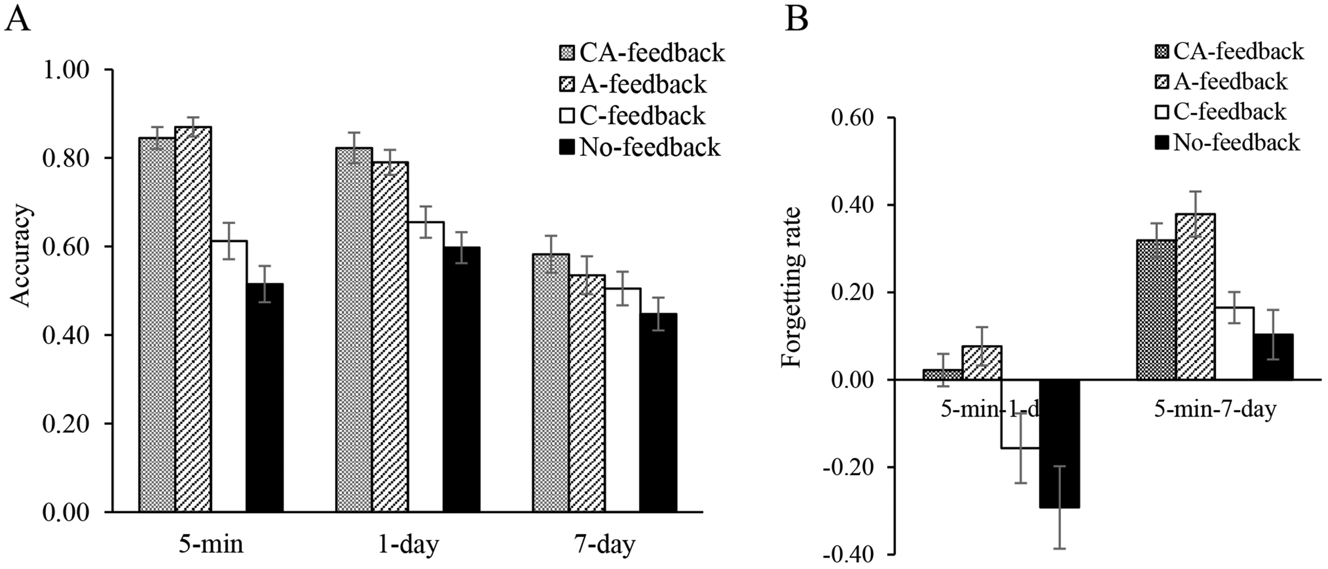

Figure 3. Results of accuracy (A) and forgetting rate (B). CA-feedback and A-feedback improved the memory accuracy, and this improvement remained 7 days later. The forgetting rate was slower after C-feedback. The error bars represent the *SE*s of the means.

C-feedback was significantly higher than that after no feedback, $F(1, 24) = 7.91$, $p = .010$, $\eta_p^2 = .25$. There was no significant interaction between feedback type and retention interval, $F(1.99, 47.68) = 0.69$, $p = .507$, $\eta_p^2 = .03$, and the improvement after C-feedback remained for 7 days, $p = .042$. The results suggested that compared to the no-feedback condition, the improvement after C-feedback was also maintained for 7 days.

To further explore the effect of feedback type on forgetting, the forgetting rates of 1 day and 7 days were analyzed. For the forgetting rate of 1 day, there was a significant effect of feedback type, $F(2.14, 51.30) = 5.82$, $p = .004$, $\eta_p^2 = .20$. The forgetting rate after CA-feedback, $p = .073$, and A-feedback, $p = .022$, was higher than that after no feedback whereas any of the other two had no significant difference, *p*s > .150. The forgetting rate was at chance level (0) after CA-feedback, $p = .555$, but higher than chance level after A-feedback, $p = .095$ (Figure 3B, left). For the forgetting rate of 7 days, there was also a significant effect of feedback type, $F(3, 72) = 9.06$, $p < .001$, $\eta_p^2 = .27$. Further analysis showed that the forgetting rate after CA-feedback and A-feedback was significantly higher than that after C-feedback and no feedback, *p*s < .030, whereas the former two and the latter two both had no significant difference, *p*s = 1.00 (Figure 3B, right). The forgetting rate was higher than chance level (0) after each feedback type with answer and/or correct/incorrect information, *p*s < .001. The results suggested that compared to the no-feedback condition, although CA-feedback and A-feedback improved subsequent memory performance, the memory decreased significantly over time whereas the memory after C-feedback remained relatively stable.

Interaction of feedback type, retention interval, and initial correctness on remembering and forgetting

We next included initial correctness as a factor to explore its effect and interaction with feedback type and retention interval on the final memory accuracy. The results showed that there was a significant three-way interaction, $F(6, 126) = 6.72$, $p < .001$, $\eta_p^2 = .24$. For the accuracy of initially incorrect trials, there were significant main effects of feedback type, $F(3, 63) = 74.27$, $p < .001$, $\eta_p^2 = .78$, and retention interval, $F(2, 42) = 66.65$, $p < .001$, $\eta_p^2 = .76$. Their interaction was significant as well, $F(6, 126) = 12.08$, $p < .001$, $\eta_p^2 = .37$ (Figure 4A). Further analysis showed that from 5 min to 7 days, the accuracy after CA-feedback and A-feedback was significantly higher than that after C-feedback and no feedback, *p*s < .030. There was no significant difference between CA-feedback and A-feedback, *p*s > .500, or between C-feedback and no feedback, *p*s > .290. However, when CA-feedback and A-feedback were directly compared, the accuracy of initially incorrect trials was significantly higher after CA-feedback than that after A-feedback at 1 day, $p = .052$, whereas they were comparable at 7 days, $p = .720$. In addition, the CA-feedback and A-feedback conditions differed in their time comparisons. The accuracy after CA-feedback only decreased significantly from 1 to 7 days, $p < .001$, whereas the accuracy after A-feedback decreased significantly from 5 min to 1 day, and from 1 to 7 days, *p*s < .030. The accuracy after C-feedback and no feedback did not decrease significantly over time, *p*s > .200. Accuracy of initially correct trials decreased significantly over time, $F(1.60, 38.43) = 32.28$, $p < .001$, $\eta_p^2 = .57$, but there was no significant effect of feedback type, $F(2.48, 59.44) = 1.35$, $p = .268$, $\eta_p^2 = .05$, or their interaction, $F(3.03, 72.71) = 1.01$, $p = .395$, $\eta_p^2 = .04$ (Figure 4B). The results suggested that CA-feedback and A-feedback could correct initially incorrect trials during retrieval practice. In addition, the corrected memory after CA-feedback remained stable from 5 min to 1 day whereas that after A-feedback decreased from 5 min to 7 days. Different types of feedback had no significant difference on memory of initially correct trials.

For the forgetting rate of 1 day, ANOVA with feedback type and initial correctness as factors showed a significant effect of feedback type, $F(2.42, 53.14) = 3.09$, $p = .045$, $\eta_p^2 = .12$, and a marginally significant interaction, $F(2.30,$

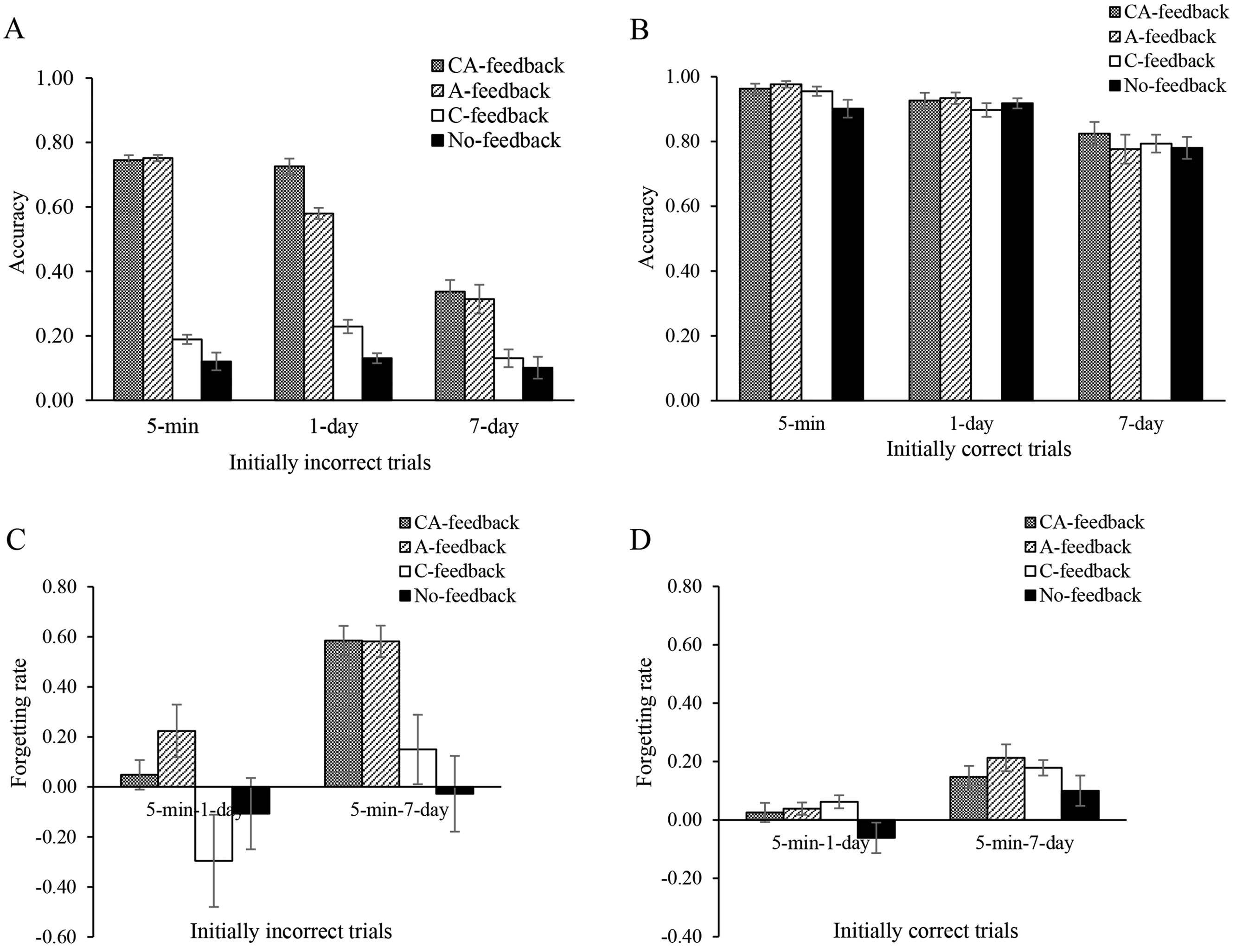

Figure 4. Results of accuracy (A, B) and forgetting rate (C, D) for initially incorrect and correct trials. The CA-feedback and A-feedback enhanced memory performance by correcting initial errors, but the corrected memory was easier to forget at 7 days. Compared to A-feedback, the CA-feedback maintained the corrected memory after 1 day. The error bars represent the *SE*s of the means.

50.65) = 2.81, p = .062, η_p^2 = .11. Further analysis showed that for the initially incorrect trials, there was no significant difference between either feedback condition, ps > .100; however, the forgetting rate of A-feedback was significantly higher than chance level, p = .052, and the forgetting rates of the other three feedback conditions were at chance level, ps > .200 (Figure 4C, left). When an outline (>3 *SD*) was excluded, the results showed that the forgetting rate after A-feedback was higher than that after no feedback, p = .024, C-feedback, p = .009, and CA-feedback, p = .095, whereas the latter three had no significant differences, ps > .350. This was consistent with the results of time comparisons of memory accuracy (Figure 4A). For the initially correct trials, there was no significant difference between either of the two conditions, ps > .100 (Figure 4D, left). The forgetting rate of each feedback type was comparable to chance level, ps > .050. The results suggested that although A-feedback could correct initial errors, the corrected memory representations were more easily forgotten at 1 day whereas feedback with correct/incorrect information maintained the corrected memory at shorter intervals.

For the forgetting rate of 7 days, ANOVA with feedback type and initial correctness as factors showed a significant effect of feedback type, F(1.90, 41.56) = 9.07, p = .001, η_p^2 = .29, and a significant interaction, F(1.91, 41.94) = 5.26, p = .010, η_p^2 = .19. The forgetting rate of initially incorrect trials was significantly higher than that of initially correct trials after CA-feedback and A-feedback, ps < .001. Further analysis showed that for the initially incorrect trials, the forgetting rate after CA-feedback and A-feedback was higher than that after no feedback, ps < .020, and C-feedback, ps < .075, whereas the former two and the latter two both had no significant difference, ps = 1.00 (Figure 4C, right). The forgetting rates after CA-feedback and A-feedback were both significantly higher than chance level, ps < .001, whereas after C-feedback and no feedback, they were at chance level, ps > .200. For the initially correct trials, the feedback type had no significant effect on forgetting rate, ps = 1.00, and the forgetting rates of each feedback condition were significantly higher than chance level, ps < .050 (Figure 4D, right). Thus, the memory for the initially correct trials remained stable at 1 day, but was forgotten at 7 days. The results suggested that although CA-feedback and A-feedback could correct initial errors, the corrected memory representations were more easily forgotten at 7 days whereas the forgetting of C-feedback was relatively slow.

Interaction of feedback type, initial correctness with confidence on remembering

In addition to initial correctness, the confidence level in the retrieval practice phase also influenced the effect of feedback type on memory accuracy. To ensure that the final memory was remembered, we first analyzed the accuracy of trials with final high confidence. The results showed that there was a significant three-way interaction, F(3, 69) = 5.68, p = .002, η_p^2 = .20. For the initially incorrect trials, the ANOVA with feedback type and initial confidence as factors showed a significant effect of feedback type, F(2.48, 57.02) = 65.56, p < .001, η_p^2 = .74. The accuracy after CA-feedback and A-feedback was significantly higher than that after C-feedback and no feedback, ps < .001, but there was no significant difference between CA-feedback and A-feedback, p = .552, or between C-feedback and no feedback, p = 1.00 (Figure 5A). The accuracy of high confidence was higher than that of low confidence, F(1, 23) = 3.88, p = .061, η_p^2 = .14. There was no significant interaction between feedback type and confidence level, F(2.39, 55.03) = 2.20, p = .111, η_p^2 = .09. The result suggested that CA-feedback and A-feedback corrected initially incorrect trials, regardless of the initial confidence level.

For the initially correct trials, the ANOVA showed significant effects of feedback type, F(2.45, 56.31) = 7.67, p = .001, η_p^2 = .25, and confidence level, F(1, 23) = 86.15, p < .001, η_p^2 = .79. There was also a significant interaction between them, F(2.25, 51.65) = 3.11, p = .047, η_p^2 = .12. For the condition of low confidence, the accuracy after CA-feedback, $M_{\text{difference}}$ = 0.26, p = .029, and A-feedback, $M_{\text{difference}}$ = 0.21, p = .034, was significantly higher than that after no feedback, and the accuracy after CA-feedback, $M_{\text{difference}}$ = 0.26, p = .051, and A-feedback, $M_{\text{difference}}$ = 0.21, p = .298, uncorrected p = .049, was also higher than that after C-feedback. There was no significant difference between CA-feedback and A-feedback or between C-feedback and no feedback, ps = 1.00. For the condition of high confidence, the accuracy after CA-feedback, $M_{\text{difference}}$ = 0.05, p = .166, uncorrected p = .026, and A-feedback, $M_{\text{difference}}$ = 0.09, p = .013, was

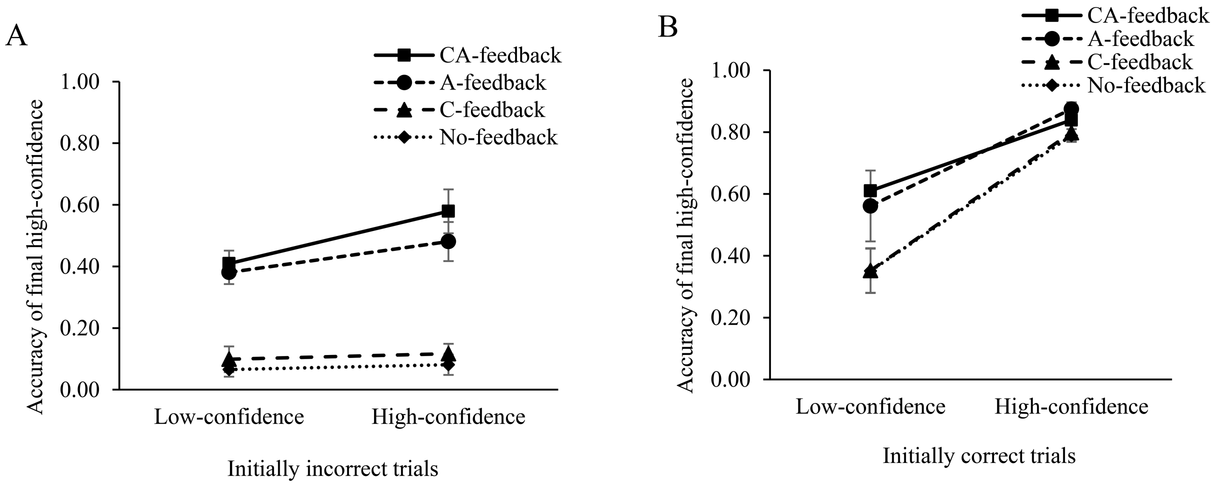

Figure 5. Results of final high-confidence accuracy for initially incorrect (A) and correct (B) trials. The final high-confidence accuracy of initially incorrect trials and initially correct trials improved after CA-feedback and A-feedback. The error bars represent the *SE*s of the means.

significantly higher than that after no feedback, and the accuracy after CA-feedback, $M_{\text{difference}} = 0.04$, $p = .553$, uncorrected $p = .096$, and A-feedback, $M_{\text{difference}} = 0.08$, $p = .006$, was also higher than that after C-feedback (Figure 5B). There was no significant difference between CA-feedback and A-feedback or between C-feedback and no feedback, *p*s = 1.00. Note that although the accuracy after CA-feedback and A-feedback was higher than that after no feedback and C-feedback for both conditions of low and high confidence, the difference was much larger for the high-confidence condition, $M_{\text{difference}} > 0.20$, than for the low-confidence, $M_{\text{difference}} < 0.10$, condition. We further analyzed the accuracy difference between CA-/A-feedback and no feedback, and between CA-/A-feedback and C-feedback in low-confidence and high-confidence conditions by ANOVA (see supplementary materials), and the results showed that the difference with low confidence was significantly larger than that with high confidence (Figure S1). Therefore, the results suggested that CA-feedback and A-feedback both improved the accuracy of initially correct trials, but the improvement with low confidence was better than that with high confidence.

We also analyzed the accuracy of trials with final low confidence. The ANOVA showed that there was no significant effect of feedback type, $F(3, 69) = 1.91$, $p = .136$, $\eta_p^2 = .08$, or three-way interaction, $F(3, 69) = 1.52$, $p = .218$, $\eta_p^2 = .06$. This suggested that the type of feedback had no significant difference on the accuracy of trials with final low confidence, regardless of initial correctness and confidence.

Interaction of feedback type, initial correctness with confidence on final confidence

In addition to memory accuracy, the final confidence was influenced by feedback type and initial memory features. For the initially incorrect trials, about 60% of them was corrected after CA-feedback and A-feedback, and most corrected trials were high confidence in the final test (72–92%). On the other hand, although C-feedback could not correct errors, C-feedback may affect their final confidence, and this effect may be influenced by initial confidence level. Therefore, we performed the ANOVA with two feedback types (C-feedback, no feedback) and initial confidence level (low, high) as factors. The results showed that the initially incorrect trials with high confidence were rated higher in final confidence than those with low confidence, $F(1, 23) = 39.36$, $p < .001$, $\eta_p^2 = .63$, and there was a significant interaction, $F(1, 23) = 5.49$, $p = .028$, $\eta_p^2 = .19$ (Figure 6A). Further analysis showed that for the initially incorrect trials with high confidence, the final confidence after C-feedback was significantly lower than that after no feedback, $p = .050$. For those with low confidence, however, the final confidence between C-feedback and no feedback had no significant difference, $p = .188$. The results suggested that C-feedback decreased the final confidence of the initially incorrect trials with high confidence.

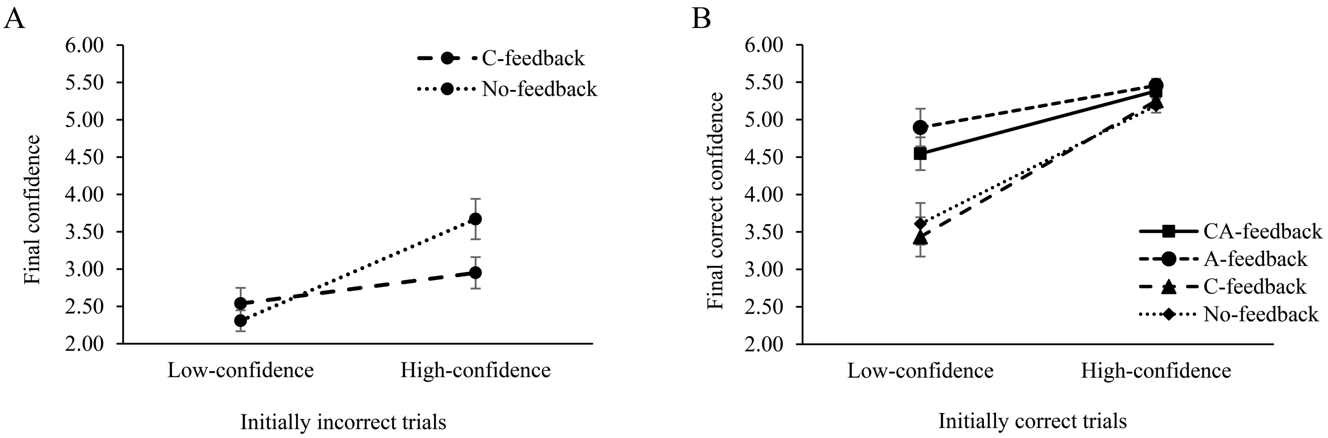

Figure 6. Results of final confidence for initially incorrect (A) and correct (B) trials. The confidence of initially incorrect trials with high confidence was lower after C-feedback than after no feedback. The confidence of initially correct trials with low confidence was higher after CA-feedback and A-feedback. The error bars represent the *SE*s of the means.

For the initially correct trials, only the final correct trials were analyzed. The ANOVA with feedback type and initial confidence level as factors showed that there was a significant main effect of initial confidence level, $F(1, 23) = 90.99$, $p < .001$, $\eta_p^2 = .80$, and feedback type, $F(3, 69) = 13.46$, $p < .001$, $\eta_p^2 = .37$. Their interaction was significant as well, $F(3, 69) = 7.52$, $p < .001$, $\eta_p^2 = .25$ (Figure 6B). Further analysis showed that for the initially correct trials with high confidence, there was no significant difference between either feedback type in the final confidence, *p*s > .050. For those with low confidence, however, the final confidence after CA-feedback and A-feedback was significantly higher than that after C-feedback, *p*s < .011, and no feedback, *p*s < .050. There was no significant difference between CA-feedback and A-feedback, $p = .648$, or between C-feedback and no feedback, $p = 1.00$. The results suggested that only feedback with answers improved the final confidence of initially correct trials with low confidence.

Discussion

In this study, feedback type, retention interval, and initial correctness with confidence were manipulated to explore their influences on subsequent memory performance. Both final memory accuracy and confidence level were analyzed as dependent variables. There were three main findings. First, the feedback containing answers significantly improved memory accuracy, and the improvement was maintained for 7 days. The C-feedback also improved memory performance, but mainly slowed down the forgetting rate. Second, the correctness during the retrieval practice phase interacted with feedback type to influence subsequent memory performance. For those initially incorrect trials, the feedback with answers corrected errors, but the corrected memory was more easily forgotten, especially from 1 day to 7 days. The CA-feedback had an advantage to maintain the corrected memory at shorter intervals. Third, the confidence level during the retrieval practice phase interacted with feedback type to influence subsequent memory performance. The feedback with answers improved the final accuracy, whether the initial confidence was low or high. Furthermore, the feedback with answers improved the final confidence of initially correct trials with low confidence, and C-feedback reduced the final confidence of initially incorrect trials with high confidence. These results highlighted the interaction between feedback type and initial correctness/confidence to influence subsequent memory performance.

Influence of different feedback types on remembering and forgetting

Previous studies have shown that memory advantage of retrieval practice is mainly reflected in the retention of long-term memory, such as 1 week (Roediger & Karpicke, 2006; Rowland, 2014; Wheeler et al., 2003) or even 9 weeks (Spitzer, 1939; Roediger & Butler, 2011). We consistently found that memory enhancement after feedback had a long-term advantage over 7 days. More importantly, we found a significant effect of feedback type as well as the interaction between feedback type and retention interval on memory accuracy and forgetting rate. The memory accuracy was significantly increased after the feedback with answers (i.e., A-feedback, CA-feedback), although the forgetting rate was also higher. According to the reconsolidation theory (Dudai, 2012; for review, see Metcalfe, 2017; Nader, Schafe, & Le Doux, 2000), answer feedback provides an opportunity for additional study (for reviews, see Butler & Woodward, 2018; Jang & Marshall, 2018), which induces stronger reactivation and reconsolidation processes. Participants could correct, update, and further strengthen the memory representations, making the difference between items with and those without feedback larger (Halamish & Bjork, 2011; Kornell et al., 2011). Therefore, the memory enhancement after feedback, especially with answers, could be maintained for a long time (for reviews, see Metcalfe, 2017; Roediger & Butler, 2011).

When only correct/incorrect information (i.e., C-feedback) was provided, the memory enhancement was smaller, but still maintained for 7 days. As shown in previous studies (e.g., Fazio et al., 2010; Marsh et al., 2012; Pashler et al., 2005), memory enhancement was stronger after A-feedback than after C-feedback. Different from these studies, though, we further explored the forgetting rate after different feedback types and found that C-feedback slowed memory forgetting. This suggests that although C-feedback does not provide additional correct information, the reactivation and reconsolidation processes after retrieval practice also strengthen memory traces (Antony et al., 2017; Dudai, 2012; for reviews, see Metcalfe, 2017; Nader et al., 2000).

In this study, the CA-feedback condition was included, during which answers and correct/incorrect information were provided together, to explore whether it had an advantage over A-feedback on memory performance. Although there was no significant difference between the memory accuracy after CA-feedback and A-feedback, the interaction between feedback type and retention interval was significant. The accuracy decreased significantly from 1 day to 7 days after each feedback condition, but it did not decrease significantly from 5 min to 1 day after CA-feedback. In addition, only after A-feedback was the forgetting rate of 1 day higher than chance level. This suggests that compared to feedback with only answers, feedback with additional correct/incorrect information could maintain the memory after 1 day and lead to slower forgetting at shorter intervals.

Initial correctness interacted with feedback type to influence remembering and forgetting

One novel finding of the study was that the memory performance and forgetting rate were influenced by feedback type and initial correctness. There was a significant interaction between the two factors. For the initially correct trials, different types of feedback had similar accuracy (Butler et al., 2008; Fazio et al., 2010; for a review, see Metcalfe, 2017) and forgetting rates. For the initially incorrect trials, feedback with answers significantly corrected these errors, thereby increasing the final accuracy, and this improvement maintained for 7 days. According to the prediction error model (Friston, 2005; for a review, see Metcalfe, 2017), when there is a bias between the incorrect response and correct answer, participants would pay more attention to the feedback information. Thus, the correct information could be re-encoded and updated into the original memory, leading to improved subsequent memory accuracy (Butler et al., 2008; Kornell et al., 2011).

However, the forgetting rate of initially incorrect trials after different feedback types was different. Even if memory for the initially incorrect trials was corrected after A-feedback and CA-feedback, the 7-day forgetting rate was higher than that for the initially correct trials. In addition, the forgetting rate was also higher than that after C-feedback and no feedback. It nicely explained why the forgetting rate after feedback with answers was significantly higher, which suggests that the new memory obtained by feedback correction is different from the originally correct memory (Sadeh, Ozubko, Winocur, & Moscovitch, 2014). The strength of the new memory after feedback may be weaker than the originally correct memory, because the new memory corrected after feedback is only learned once and does not undergo sufficient reconsolidation processes (Antony et al., 2017; Roediger & Butler, 2011). Thus, over time, the memory obtained by the correction of answer feedback is more easily forgotten than the initially correct memory, and the incorrect memory is gradually restored (Metcalfe & Miele, 2014).

Compared to A-feedback, the memory accuracy of initially incorrect trials after CA-feedback did not decrease significantly from 5 min to 1 day. Only after A-feedback was the forgetting rate of initially incorrect trials at 1 day higher than chance level. It was not only higher than that after C-feedback and no feedback, but also higher than that after CA-feedback. Therefore, compared to feedback with only answers, feedback with additional correct/incorrect information could maintain the new corrected memory after 1 day and led to slower forgetting at shorter intervals.

Initial correctness with confidence interacted with feedback type to influence memory confidence

Another novel finding of the study was that the final memory confidence was influenced by the interaction among feedback type, initial correctness, and confidence level. Previous studies have shown that memory accuracy was improved after A-feedback for initially correct trials with low confidence (Butler et al., 2008; Fazio et al., 2010), but few studies have analyzed to what extent the final confidence was influenced by the feedback type and initial memory features. Our study showed that compared to the no-feedback condition, feedback with answers improved the final accuracy of initially correct trials, especially for those with low confidence. In addition, feedback with answers improved the final confidence of initially correct trials with low confidence. Memory confidence reflects metacognition level and is an important memory feature in addition to accuracy (Chua et al., 2008; Dunlosky & Metcalfe, 2009). Even when memory contents are correct, the confidence level may differ. The higher the degree of confidence, the more details and vividness of the memory are, and the more dependence on the recollection process is (Eichenbaum, Yonelinas, & Ranganath, 2007; Yonelinas, 1994). In contrast, the memory with low confidence may indicate fewer details, less vividness, and more dependence on familiarity (Dunlosky & Metcalfe, 2009; Eichenbaum et al., 2007; Yonelinas, 1994). Given answer feedback after retrieval practice, participants could get the correct answer and then add more details to the original memory, which makes the memory more vivid and rely more on the recollection process in the final test. Therefore, the feedback with answer improves the final accuracy and confidence of initially correct trials with low confidence.

In addition, the final confidence of initially incorrect trials with high confidence was reduced after C-feedback. This is the first study that clarifies the effect of C-feedback on the final confidence and suggests that C-feedback also influences memory confidence, but in a different way. After receiving correct/incorrect feedback, participants could realize that their metacognition of the response was biased. By the reactivation processing, the connection between the existing false memory and target cues is inhibited or eliminated (Butler et al., 2008; for a review, see Metcalfe, 2017), which leads to reduced confidence of incorrect trials.

However, contrary to our hypothesis, the confidence of initially correct trials with low confidence was not improved after C-feedback. This may be related to the fact that we did not include retention interval as a factor in the analysis of final confidence. It is possible that the vividness of the participants' responses for the correct trials increases after C-feedback. However, as C-feedback did not provide more details of answers, the confidence improvement after C-feedback may last just for a short period. As there were several missing values for initially correct trials with low confidence at various retention intervals, the data at different intervals were combined. We thus could not perform the repeated measures ANOVA including retention interval as a factor to clarify the effect of retention interval in final confidence.

Limitations and future directions

This study has some limitations for future investigations to consider. First, there were missing values in the analyses including memory confidence. To explore the influence of initial correctness and confidence level, we divided the trials into four conditions. There were 48 sentences used in this study and thus 16 descriptions in each condition (for each interval). As also shown in previous studies (e.g., Butler et al., 2008; Fazio et al., 2010), some participants had no responses for some conditions, especially for the conditions of initially correct trials with low confidence and initially incorrect trials with high confidence. Some repeated ANOVA measurements could not be performed, such as the interaction of feedback type and retention interval on the final confidence of initially correct trials with low confidence. Thus, we used expectation maximization imputation to replace missing values in some analyses (Rashid & Gupta, 2019; Schafer & Graham, 2002), and the results were similar to those when the participants with missing data were excluded. In future studies, researchers could also consider using other types of materials and increasing the number of materials to collect sufficient data for complete analyses. Second, the results of interaction between feedback type and initial confidence for the final high-confidence accuracy (Figure 5B) had some non-significant *p* values after Bonferroni multiple comparisons, although the direct *t* tests were significant (e.g., the comparisons between CA-feedback and no feedback/C-feedback). This happens when repeated measures ANOVAs include two factors (e.g., feedback and retention interval/correctness/confidence level), and each factor has at least two levels (Perneger, 1998). Nevertheless, we are cautious about this part of our results, and further studies with more participants and materials are needed to verify this finding. Third, our study did not include the condition of restudy, which is typically the condition in studies of testing effect (Kornell & Vaughn, 2016; Roediger & Butler, 2011; Roediger & Karpicke, 2006; Rowland, 2014). As the answer feedback corresponds to the restudy after retrieval practice, future studies could add the restudy condition to explore the different stages of feedback.

Conclusions

By setting up four feedback types and three retention intervals combined with initial correctness and confidence level during retrieval practice, the results showed that memory improvement after feedback was influenced by the interactions of these factors. The CA-feedback and A-feedback enhanced memory performance by correcting initial errors and increasing the confidence of correct trials, but the corrected memory representations were more easily forgotten after 7 days. Compared to A-feedback, the CA-feedback maintained the corrected memory after 1 day. The C-feedback did not correct initial errors, but slowed the forgetting rate and reduced the confidence of incorrect trials. The results emphasized the interaction between feedback type and initial memory features (correctness, confidence) to influence subsequent memory performance. Although feedback is a strong factor to improve subsequent memory performance, its effect was influenced by feedback type and initial memory features. Thus, in educational practice, improving initial memory accuracy and confidence is important even when the feedback is applied.

Acknowledgments

This research was supported by Grant 32071027 (J. Yang) from the National Natural Science Foundation of China.

Disclosure of conflict of interest

The authors declare no conflict of interest.

References

Antony, J. W., Ferreira, C. S., Norman, K. A., & Wimber, M. (2017). Retrieval as a fast route to memory consolidation. *Trends in Cognitive Sciences*, *21*(8), 573–576. https://doi.org/10.1016/j.tics.2017.05.001

Butler, A. C., Fazio, L. K., & Marsh, E. J. (2011). The hypercorrection effect persists over a week, but high-confidence errors return. *Psychonomic Bulletin & Review*, *18*(6), 1238–1244. https://doi.org/10.3758/s13423-011-0173-y

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*(4), 918–928. https://doi.org/10.1037/0278-7393.34.4.918

Butler, A. C., & Roediger, H. L. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & Cognition*, *36*(3), 604–616. https://doi.org/10.3758/Mc.36.3.604

Butler, A. C., & Woodward, N. R. (2018). Toward consilience in the use of task-level feedback to promote learning. In K. D. Federmeier (Ed.), *Psychology of learning and motivation* (pp. 1–38). Academic Press. https://doi.org/10.1016/bs.plm.2018.09.001

Butterfield, B., & Metcalfe, J. (2001). Errors committed with high confidence are hypercorrected. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(6), 1491–1494. https://doi.org/10.1037//0278-7393.27.6.1491

Butterfield, B., & Metcalfe, J. (2006). The correction of errors committed with high confidence. *Metacognition and Learning*, *1*(1), 69–84. https://doi.org/10.1007/s11409-006-6894-z

Carpenter, S. K. (2011). Semantic information activated during retrieval contributes to later retention: Support for the mediator effectiveness hypothesis of the testing effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(6), 1547–1552. https://doi.org/10.1037/a0024140

Chen, H., Ning, X., Wang, L., & Yang, J. (2018). Acquiring new factual information: Effect of prior knowledge. *Frontiers in Psychology*, *9*, 1734. http://doi.org/10.3389/fpsyg.2018.01734

Chua, E. F., Schacter, D. L., & Sperling, R. A. (2008). Neural correlates of metamemory: A comparison of feeling-of-knowing and retrospective confidence judgments. *Journal of Cognitive Neuroscience*, *21*(9), 1751–1765. https://doi.org/10.1162/jocn.2009.21123

Dudai, Y. (2004). The neurobiology of consolidations, or, how stable is the engram? *Annual Review of Psychology*, *55*(1), 51–86. https://doi.org/10.1146/annurev.psych.55.090902.142050

Dudai, Y. (2012). The restless engram: Consolidations never end. *Annual Review of Neuroscience*, *35*(1), 227–247. https://doi.org/10.1146/annurev-neuro-062111-150500

Dudai, Y., Karni, A., & Born, J. (2015). The consolidation and transformation of memory. *Neuron*, *88*(1), 20–32. http://doi.org/10.1016/j.neuron.2015.09.004

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

Eichenbaum, H., Yonelinas, A. P., & Ranganath, C. (2007). The medial temporal lobe and recognition memory. *Annual Review of Neuroscience*, *30*(1), 123–152. https://doi.org/10.1146/annurev.neuro.30.051606.094328

Fazio, L. K., Huelser, B. J., Johnson, A., & Marsh, E. J. (2010). Receiving right/wrong feedback: Consequences for learning. *Memory*, *18*(3), 335–350. https://doi.org/10.1080/09658211003652491

Friederic, A. D., & Frisch, S. (2000). Verb argument structure processing: The role of verb-specific and argument-specific information. *Journal of Memory and Language*, *43*(3), 476–507. https://doi.org/10.1006/jmla.2000.2709

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions: Biological Sciences*, *360*, 815–836. https://doi.org/10.1098/rstb.2005.1622

Halamish, V., & Bjork, R. A. (2011). When does testing enhance retention? A distribution-based interpretation of retrieval as a memory modifier. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(4), 801–812. https://doi.org/10.1037/a0023219

Jang, Y., & Marshall, E. (2018). The effect of type of feedback in multiple-choice testing on long-term retention. *The Journal of General Psychology*, *145*(2), 107–119. https://doi.org/10.1080/00221309.2018.1437021

Karpicke, J. D., Lehman, M., & Aue, W. R. (2014). Retrieval-based learning: An episodic context account. In B. H. Ross (Ed.), *Psychology of learning and motivation* (Vol. *61*, pp. 237–284). San Diego, CA: Elsevier Academic Press Inc.

Kornell, N., Bjork, R. A., & Garcia, M. A. (2011). Why tests appear to prevent forgetting: A distribution-based bifurcation model. *Journal of Memory and Language*, *65*(2), 85–97. https://doi.org/10.1016/j.jml.2011.04.002

Kornell, N., & Vaughn, K. E. (2016). How retrieval attempts affect learning. *Psychology of Learning and Motivation*, *65*, 183–215. https://doi.org/10.1016/bs.plm.2016.03.003

Marsh, E. J., Lozito, J. P., Umanath, S., Bjork, E. L., & Bjork, R. A. (2012). Using verification feedback to correct errors made on a multiple-choice test. *Memory*, *20*(6), 645–653. https://doi.org/10.1080/09658211.2012.684882

Metcalfe, J. (2017). Learning from errors. *Annual Review of Psychology*, *68*(1), 465–489. https://doi.org/10.1146/annurev-psych-010416-044022

Metcalfe, J., & Finn, B. (2012). Hypercorrection of high confidence errors in children. *Learning and Instruction*, *22*, 253–261. https://doi.org/10.1016/j.learninstruc.2011.10.004

Metcalfe, J., & Miele, D. B. (2014). Hypercorrection of high confidence errors: Prior testing both enhances delayed performance and blocks the return of the errors. *Journal of Applied Research in Memory and Cognition*, *3*(3), 189–197. http://doi.org/10.1016/j.jarmac.2014.04.001

McRae, K., Cree, G. S., Seidenberg, M. S., & Mcnorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, *37*(4), 547–559. http://doi.org/10.3758/bf03192726

Moscovitch, M., Cabeza, R., Winocur, G., & Nadel, L. (2016). Episodic memory and beyond: The hippocampus and neocortex in transformation. *Annual Review of Psychology*, *67*(1), 105–134. https://doi.org/10.1146/annurev-psych-113011-143733

Nader, K., Schafe, G. E., & Le Doux, J. E. (2000). Fear memories require protein synthesis in the amygdala for reconsolidation after retrieval. *Nature*, *406*(6797), 722–726. https://doi.org/10.1038/35021052

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*, 3–8. https://doi.org/10.1037/0278-7393.31.1.3

Perneger, T. V. (1998). What's wrong with Bonferroni adjustments. *British Medical Journal*, *316*, 1236–1238. https://doi.org/10.1136/bmj.316.7139.1236

Rashid, W., & Gupta, M. K. (2019). A perspective of missing value imputation approaches. *Advances in Computational Intelligence and Communication Technology. Proceedings of CICT 2019*, 307–315. https://doi.org/10.1007/978-981-15-1275-9_25

Roediger, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, *15*(1), 20–27. https://doi.org/10.1016/j.tics.2010.09.003

Roediger, H. L., & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, *1*(3), 181–210. https://doi.org/10.1111/j.1745-6916.2006.00012.x

Roediger, H. L., Wheeler, M. A., & Rajaram, S. (1993). Remembering, knowing and reconstructing the past. *Psychology of Learning and Motivation*, *30*, 97–134. https://doi.org/10.1016/S0079-7421(08)60295-9

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin*, *140*(6), 1432–1463. https://doi.org/10.1037/a0037559

Sadeh, T., Ozubko, J. D., Winocur, G., & Moscovitch, M. (2014). How we forget may depend on how we remember. *Trends in Cognitive Sciences*, *18*(1), 26–36. https://doi.org/10.1016/j.tics.2013.10.008

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*(2), 147–177. https://doi.org/10.1037//1082-989X.7.2.147

Spitzer, H. F. (1939). Studies in retention. *Journal of Educational Psychology*, *30*(9), 641–656. http://doi.org/10.1037/h0063404

Walker, M. P., Brakefield, T., Hobson, J. A., & Stickgold, R. (2003). Dissociable stages of human memory consolidation and reconsolidation. *Nature*, *425*(6958), 616–620. http://doi.org/10.1038/nature01930

Wheeler, M. A., Ewers, M., & Buonanno, J. F. (2003). Different rates of forgetting following study versus test trials. *Memory*, *11*(6), 571–580. https://doi.org/10.1080/09658210244000414

Yonelinas, A. P. (1994). Receiver-operating characteristics in recognition memory: Evidence for a dual-process model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(6), 1341–1354. https://doi.org/10.1037//0278-7393.20.6.1341

Supporting information

Additional Supporting Information may be found in the online version of this article at the publisher's web-site: http://onlinelibrary.wiley.com/doi//suppinfo.

Appendix S1. Supporting Information