**The Dark Side of Corrective Feedback:**

**Controlled and Automatic Influences of Retrieval Practice**

Aeshah A. Alamri and Philip A. Higham

School of Psychology

University of Southampton

**© 2022, American Psychological Association. This paper is not the copy of record and may not exactly replicate the final, authoritative version of the article. Please do not copy or cite without authors' permission. The final article will be available, upon publication, via its DOI: 10.1037/xlm0001138**

**Author Notes**

Aeshah A. Alamri page75image53703056 https://orcid.org/0000-0001-5212-7909

Philip A. Higham page75image53703056 https://orcid.org/0000-0001-6087-7224

We have no known conflict of interest to disclose.

Supplemental materials are available at: https://osf.io/2jqd6/?view\_only=e4e2e621dc5145afa274649c062ce91e

Correspondence concerning this article should be addressed to Aeshah A. Alamri, School of Psychology, University of Southampton, Highfield, Southampton, UK, SO17 1BJ. Email: A.Alamri@soton.ac.uk.

**Abstract**

Corrective feedback is often touted as a critical benefit to learning, boosting testing effects when retrieval is poor and reducing negative testing effects. Here, we explore the dark side of corrective feedback. In 3 experiments, we found that corrective feedback on multiple-choice practice questions is later endorsed as the answer to related second-test questions, even though it is no longer correct. We describe this effect as an automatic influence of memory for feedback which participants fail to control. We explored how this influence is affected by the depth of retrieval during practice by successively increasing the retrieval demands of the multiple-choice practice test across the 3 experiments: Experiment 1 used a standard (select a single favorite option) format; Experiment 2 used ranking (rank order the options); and Experiment 3 used elimination testing (provide reasons for rejecting unchosen options). Increasing retrieval depth enhanced controlled influences on a cued-recall second test, evidenced by better accuracy on related versus new questions. However, it did not reduce the automatic influence on accuracy when the second test was multiple choice, partly because repeating the options between practice and test likely led to false recognition of related questions. Together, the results suggest that multiple-choice practice tests produce both automatic and controlled memory influences on related second-test questions, with retrieval depth at practice being an important determinant of the controlled influence. However, whether that controlled influence will override the automatic influence of memory for the corrective feedback also depends on the second test format.

*Keywords:* multiple-choice, cued recall, testing effect, controlled and automatic memory influences, familiarity, recollection.

**The Dark Side of Corrective Feedback:**

**Controlled and Automatic Influences of Retrieval Practice**

Multiple-choice (MC) questions are a test format widely used to assess learners’ performance. The beginning of MC development can be traced back to Fredrick J. Kelly in 1916 who created questions that needed to be answered by selecting the answers from a set of alternatives rather than generating them (Gierl et al., 2017). Since then, researchers have been investigating the positive and negative aspects of utilizing this test format for assessment. These studies have found many advantages such as the ease of marking and the ability to assess a wide range of topics. However, there are disadvantages as well, such as the time needed to develop good questions for even a single exam (Fellenz, 2004; see Butler, 2018 for a review).

**Positive and Negative Testing Effects**

The usefulness of MC tests extends beyond assessment; they can also be used as a learning tool. Specifically, research has shown that taking MC practice tests can improve later test performance. For instance, McDaniel et al. (2012) had college students take a weekly MC or cued-recall (CR) test with corrective feedback or read the information covered in the test as statements. Another group of participants were not involved in any weekly activities to act as a control group. After three weeks, participants took a final MC test. Results showed that the tested groups outperformed the study group (reading statements) as well as the control group. Moreover, the initial MC test was as effective as the CR test in terms of promoting retention in the final test. Similar benefits of MC testing on retention have been reported in many studies, which collectively are referred to the *testing effect* (e.g., see Cantor et al., 2014; Marsh et al., 2012; McDaniel et al., 2013; McDermott et al., 2014; Nungester & Duchastel, 1982; Rowland, 2014; Yang et al., 2021).

Other research, however, has demonstrated that under certain circumstances, such as cases where there is no feedback provision or there are many lures[[1]](#footnote-2), MC testing can be harmful to learning performance. For example, Roediger and Marsh (2005; see also Roediger et al., 2010) had participants take an initial MC test after a reading session where some of the questions were relevant to the reading passage (studied questions) whereas others were taken from another passage that had not been read (unstudied questions). The number of alternatives presented with the questions ranged from two to six alternatives. Also, no feedback was provided during the initial test. After completing a distracter task, participants took a final CR test which consisted of some previously tested questions as well as some new ones (control). Roediger and Marsh observed an overall positive testing effect of initial MC testing; that is, tested questions were more likely to be answered correctly than non-tested questions on the final CR test. However, closer examination of CR test performance revealed a negative effect as well. Specifically, performance on the previously tested questions deteriorated as the number of lures increased because lures intruded on the final CR test. This study was replicated by Butler and Roediger (2008; see also Marsh et al., 2007, 2009) who found that providing feedback and spending more time on studying the material reduced the negative testing effect of MC questions and increased the positive testing effect.

**Positive Effects of MC Testing on Related Items**

Thus, although some research showed the negative effect with MC testing, the positive testing effect typically outweighed the negative effect. This result indicates that even with the negative testing effect, taking initial MC tests remains advantageous. This advantage has not only been observed with previously tested questions but also with untested questions that are related to previously tested questions. For example, Little, Bjork, and colleagues (e.g., see Little & Bjork, 2016; Little et al., 2012) investigated benefits of practice testing using pairs of questions where a lure from a practice MC question was the correct answer for a CR question on a similar topic. For example, a related question on the initial MC test might be “*Where was Lope de Vega born? a. Valencia,* ***b. Madrid****, c. Genoa, d. Oxford*” (boldface indicates the correct answer), while the second related question on the final CR test was “*Where was Christopher Columbus born?* ***Genoa***”.

Little and Bjork (2015) demonstrated that the benefits of practice testing using materials like these depended on the plausibility of the lures. They had participants first read two expository texts (e.g., information about ferrets), take a practice MC test without feedback, and then take a final CR test. For the practice test, there were two types of MC questions: questions with competitive lures versus questions with uncompetitive lures. For example, one question with competitive lures was “*How many inches long is an average ferret tail? a. 7-10, b. 20,* ***c.******5***” (boldface indicates the correct answer). Uncompetitive lures were less plausible given the question; for example, to create the uncompetitive version of question on ferrets’ tail length in inches, the first two lures were replaced with “*1500*” and “*3500*”, respectively. For the final CR test, there were three types of questions: previously tested questions, related untested questions, and new questions that were generated from a passage that participants had read but not been tested on during practice. The findings from the final CR test showed best performance on previously tested questions – the standard testing effect. More critically, there was also better performance on the related untested items compared to the new questions. However, this advantage only occurred if the questions were accompanied by competitive lures, not if the lures were uncompetitive. Little and Bjork (2015) argued that competitive lures on the first MC test prompted elaboration and retrieval of information about the question, which facilitated participants’ later ability to answer related but untested questions on the CR test. Because the questions with uncompetitive lures were so easy, the same information retrieval was not necessary during practice, and so no facilitation was observed on the final test.

Subsequent studies have shown similar facilitation on related untested questions if the MC practice involves intensive retrieval processes or *lure elaboration*. For example, Sparck et al. (2016) compared the effect of two different MC practice test formats on performance with related untested CR questions. Specifically, they compared standard MC format (choose a single option) with *confidence-weighted* MC format. The latter format requires participants to rate their confidence in each of the options relative to the others, thereby encouraging more lure elaboration. Compared to new questions, they later found significantly better CR performance for related untested items regardless of the initial test format. However, the confidence-weighted MC format produced better performance on related items than the standard MC format.

**Theoretical Mechanisms**

The prior literature suggests that there are both positive and negative effects of MC practice testing. In our view, these positive and negative effects correspond roughly to controlled and automatic influences of memory, respectively (see Chan & McDermott, 2007 for a similar dual-process perspective on testing effects). Controlled memory processes are thought to be slow and involve conscious retrieval which demands attentional resources (e.g., Jacoby, 1991; Jacoby et al., 1993). On the other hand, automatic processes are faster, less conscious, and require fewer attentional resources (Neely, 1977). Jacoby et al. (1989) argued that automatic influences of memory can sometimes oppose people’s intentions causing them to make errors. In this vein, the negative testing effect occurs when familiarity with the lures encountered during practice intrude on later CR tests in ways that participants cannot control. The automatic effect increases if there are multiple lures on the test creating interference (e.g., Roediger & Marsh, 2005) or if knowledge is lacking that could counter the false familiarity (e.g., Marsh et al., 2009). Conversely, if hard practice MC questions are used which can only be answered with intensive retrieval (Little & Bjork, 2015) or if the practice MC format requires lure elaboration (Sparck et al., 2016), the result is a controlled influence of memory which facilitates later CR performance. In this instance, the controlled influence takes the form of participants consciously utilizing information retrieved during practice to answer the later related question in a controlled manner.

In both cases above showing automatic and controlled effects of retrieval practice, the test was CR. However, in Chan and McDermott (2007) research, a testing effect was always observed in final CR tests but not when recognition tests, such as MC, were used as the final test. If both the practice and final tests were MC – tests that are often associated with familiarity-based responding (e.g., Ozuru et al., 2013) ­– would controlled influences from lure elaboration during practice as observed by Little and colleagues be overshadowed by automatic influences?

Higham et al. (2016) tested this possibility in several experiments using untested related items like those used in Little and colleagues’ paradigm. After a reading session, participants took an MC initial test. However, unlike Little and Bjork (2015), participants were provided with corrective feedback after each question in the practice test. Moreover, instead of using a CR test format in the final stage of these experiments, a format likely to encourage effortful retrieval, they used a standard MC test. Untested related questions on the final MC test were analogous to those for the final CR test used in prior research; that is, they queried the same topic, but a lure from the practice test was now the correct option. The main difference was that the same set of options used on the practice MC test were explicitly presented as options on the final MC test. Thus, in most experiments, the options on the final MC test consisted of the previous correct answer (which was now a lure), a new correct answer (which was previously a lure), and two additional lures. The question that Higham et al. (2016) addressed was whether participants would continue to show facilitated performance due to controlled memory influences, as in Little and colleagues’ research. They reasoned that false familiarity derived from explicit presentation of the previously correct option, whose processing would have been reinforced by feedback during practice (regardless of whether it was chosen), might override any controlled influences of memory. If so, then performance on untested related questions would show worse, not better, performance compared to new items.

Indeed, Higham et al. (2016) found that prior MC testing was harmful when performance on related untested items was compared to new questions (i.e., previously untested questions that queried the topic at hand, but which were not related to any previous questions). This interference of MC testing on related items remained even when only the feedback option was repeated between the related items. That is, participants still erroneously chose the feedback option on the second test, which was correct on the previous test but incorrect on the later test, at a greater rate for related untested questions than new ones. The poor performance for related untested questions even persisted when participants were asked to read aloud each question on the final test before answering it. This procedure ensured that the question stem, which changed between the two versions of the related untested questions, was not being ignored. These results differed completely from the findings obtained by Little and colleagues as they always observed facilitation on the final CR test even when corrective feedback was provided during the initial test (e.g., Little et al.,2012). Together, these results suggest that if the final test format allows for quick familiarity-based responding, any controlled influences attributable to lure elaboration during practice might be overshadowed by the false familiarity of the corrective feedback.

Research on the negative testing effect (Roediger & Marsh, 2005) demonstrated that retrieval practice can produce automatic memory influences in the form of intrusions or false recollections (e.g., Jacoby et al., 1993) on CR tests. However, Little and Bjork (2015; see also Little et al., 2019) examined the issue of false recollections/intrusions in their research and concluded that, although they occasionally occurred, they did not overshadow the controlled memory influence. However, given the robustness of Higham et al.’s (2016) false feedback effect with MC final tests, and the fact that lure intrusions occur with some regularity in research on the negative testing effect, we believe further research on automatic and controlled memory influences in both MC and CR is warranted.

For example, in what way does increasing the level of information retrieval during the first test influence those processes? Greater lure elaboration during practice testing has been shown to enhance controlled memory influences (e.g., Little & Bjork, 2015; Sparck et al., 2016), which boosts final test performance. However, what effect might deeper processing during practice have on automatic influences? On the one hand, it might increase the familiarity of the corrective feedback (and other lures), thereby increasing automatic influences on the second test and worsening final-test performance, particularly if the second test is MC. On the other hand, the concomitant increase in controlled influences caused by deeper processing at practice might oppose those enhanced automatic influences, thereby facilitating final-test performance. Thus, examining the impact of practice MC tests on later MC and CR tests might help to understand the interplay of opposing controlled and automatic memory influences in an educational context.

**Current Study**

The main difference between prior studies that investigated the MC influence on related untested items (e.g., Little & Bjork, 2015, 2016; Little et al., 2012, 2019) and the study by Higham et al. (2016) was the format of the final test (CR vs. MC, respectively). To explore the interplay between controlled and automatic influences of retrieval practice, we included both MC and CR final tests in all experiments and compared the patterns of performance (for an example of related items in both test formats see Figure 1). Also, across three experiments, we successively incremented the likelihood of lure elaboration and retrieval during the MC practice test by varying the test format (Sparck et al., 2016). We expected that increasing lure elaboration would enhance the performance of the related questions in CR, in line with Sparck et al. (2016). However, deep retrieval processes during practice might also enhance the familiarity gained for options that are incorrect responses on the final test, unless that familiarity is offset by controlled processes. If so, then compared to shallow retrieval, we might also observe more erroneous selections on the final MC test and/or more intrusions on the final CR test.

All experiments reported in this paper were granted ethical approval by the Ethics Committee at the University of Southampton.

**Experiment 1**

The first experiment explored the consequences of using different final test formats on the learners’ outcomes. Participants first answered a set of MC general knowledge questions on an initial test. On a second test taken immediately afterwards, some questions were repeated, some were related but untested, and some were new. Like Higham et al. (2016), we hypothesized that performance on the MC related items will be impaired compared to new questions due to the false familiarity of the corrective feedback option that is repeated from the practice phase. In contrast, because no option is repeated between phases if the final test is CR, any misleading effects of the corrective feedback will be limited. Instead, we predicted that controlled memory influences would prevail if the second test is in CR format, and performance on related items will be facilitated compared to new items.

**Method**

***Participants***

Since this study was a first attempt to investigate the consequences of employing different final test formats on the performance of related items, we conducted an a priori power analysis assuming a medium effect size of Cohen’s *f* = .25, alpha = .05 and power = .80. It indicated that a minimum of 128 participants was needed. We exceeded that criterion by testing 185 participants. However, 13 participants were excluded after reviewing their performance and responses to attention-check questions (see below). For example, some participants responded to the cued-recall questions with random words such as “Yes” or “No,” or with punctuation marks. A few others seemingly looked up answers to the questions on the internet because they provided very long, detailed answers on the CR test that were taken from websites, despite our warning not to look up the answers. Our final sample included the remaining 172 participants (female = 51), with ages ranging between 20 to 60 years (*M* =34.08, *SD* = 9.20) from the general population. They were recruited through Amazon Mechanical Turk (MTurk) and were given $2 in return for their participation in the study. The experiment involved two groups, with 85 participants in the MC group and 87 in the CR group, and 12 counterbalancing formats with 6-10 participants each as explained in more detail in the next section.

***Design***

The experiment employed a 2 x 3 mixed factorial design with final test type (CR, MC) manipulated between subjects, and question type on the final test (new, repeated, related-but-untested) manipulated within subjects. The primary dependent variable was the participants’ mean accuracy on the final test. To ensure that each question served in each experimental condition equally often, we created 12 surveys on Qualtrics, the software used to present the questions. The 12 surveys rotated the questions through the experimental conditions across participants to eliminate item effects. Twenty-two questions were presented in the initial test and 33 questions in the final test. The 33 final test questions consisted of 11 related (untested) questions, 11 repeated (tested) questions, and 11 new (control) questions. Questions were presented in a random order, but the MC alternatives were always presented in the same order.

***Materials and Procedure***

Thirty-two pairs of trivia questions were acquired from Little et al. (2019). They covered a wide range of topics (e.g., mythology, literature, science, history, geography). To meet the demands of our design, we added one more pair to total 33 question pairs.

The experiment was conducted online using Qualtrics survey software, and two groups of participants were fully instructed on the procedure prior to taking the survey. After reading the instructions, participants answered two attention-check questions about the instructions to ensure that they read and understood them. The experiment started with taking the initial test, consisting of 22 standard MC questions for about seven minutes. For each question in the initial test, a question stem was presented with four alternatives below it. Participants selected a single alternative by clicking the radio button that appeared next to it and then clicked “Next” to receive corrective feedback. The feedback was provided by Qualtrics software which highlighted the correct answer with a green color and a red color for the three incorrect alternatives. Clicking “Next” again advanced to the next question. After completing the initial test, participants engaged in a filler task of answering basic mathematics questions for about four minutes. Until this point, the procedure was identical for both groups.

After finishing the filler task, each group took either the MC or CR version of the final test for about 10 minutes. Both versions contained 33 questions which were divided into three categories: (a) 11 new questions – which acted as the control questions – that were untested and unrelated to the questions from the initial test; (b) 11 repeated questions that had been tested already in the initial test, which was half of the 22 questions that were presented in the initial test; and (c) 11 related untested questions that had not been tested themselves, but were related to previously tested questions (related to the other, non-repeated half of 22 questions that appeared in the initial test). The MC version of the final test was in the same format as the initial MC test. For the CR version of the final test, the question stem was presented with no alternatives and participants typed their response into a text box that appeared below the question stem. No feedback was provided for either version of the final test. The experiment was self-paced and took each participant approximately 20-25 minutes to complete.

For the scoring, the initial MC test as well as the final MC test were scored automatically via Qualtrics (1 = correct; 0 = incorrect). For the CR final test, answers were scored as either fully correct (1), partially correct (0.5; e.g., providing the name of the country rather than the city or a first name rather than a complete name), or wrong (0).

**Results**

***Initial Test Performance***

The initial test was in the same MC format for both the MC and CR groups. (The group names refer to the format of the final test.) Although the two groups were treated the same on the initial test, we compared performance to check for sampling error. For the MC group, the mean proportion of items which participants answered correctly on the first MC test was .62 (*SD* = .17), while it was .64 (SD = .19) for the CR group. An independent samples *t*-test showed that this difference was not significant *t*(168) = -0.59, *p* = .56, *d =* .09*,* suggesting the samples making up the MC and CR groups were comparable.

***Final Test Performance***

**Final-Test Accuracy.** We conducted a 2 (final test type: MC, CR) x 3 (question type: new, repeated, related) mixed-factor Analysis of Variance (ANOVA) with final test accuracy as the dependent variable (see Figure 2). We found a significant main effect of final test type, *F*(1, 170) = 26.39, *p* < .001, ηp2 = .13, such that accuracy was significantly higher on the MC test (*M* = .69; *SD* = .16) compared to the CR test (*M* = .55; *SD* = .18). We also found a significant main effect of question type, *F*(2, 340) = 204.68, *p* < .001, ηp2 = .55. Paired-sample *t*-tests revealed that accuracy was comparable across the new (*M* = .53, *SD* = .24) and related (*M* = .51, *SD* = .24) items, *t*(171) = -0.81, *p* = .42, *d* = .06, but significantly higher on the repeated items (*M* = .82, *SD* = .18), compared to the new, *t*(171) = -16.82, *p* < .001, *d* = 1.36, and related items, *t*(171) = -17.73, *p* < .001, *d* = 1.45. Finally, we found a significant interaction between final test type and question type, *F*(2, 340) = 4.22, *p* = .02, ηp2 = .02. For the CR group, paired-sample *t*-tests showed higher performance on the repeated items (*M* = .78, *SD* = .19) compared to related items (*M* = .44, *SD* = .22), *t*(86) = -13.48, *p* < .001, *d* = 1.60, and new items (*M* = .44, *SD* = .24), *t*(86) = -13.63, *p* < .001, *d* = 1.55, while the performance was comparable between new and related items, *t*(86) = 0.16, *p* = .87, *d* = .02. The same general pattern was observed in the MC group, but the differences between performance on the repeated items and the other two item types were smaller, resulting in the interaction. Specifically, participants performed better on the repeated items (*M* = .86, *SD* = .16) versus related items (*M* = .58, *SD* = .23), *t*(84) = -11.07, *p* < .001, *d* = 1.39, and new items (*M* = .62, *SD* = .21), *t*(84) = -10.53, *p* < .001, *d* = 1.30, while performance was comparable between new and related items *t*(84) = -1.37, *p* = .17, *d* = .15.

**Final-Test Answer Types**. Higham et al. (2016) found that when participants answered the first of the related question pairs incorrectly on the first MC test, they tended to select the corrective feedback for the second pair on the second MC test. In contrast, they tended to avoid their original incorrect selection, despite its high familiarity. To investigate whether a similar pattern was observed in our current data, we computed the probability of different answer types conditioned on their initial responses being incorrect (see Figure 3 for an illustration). The analysis was limited to incorrect answers on the initial test so that we would be able to compare participants’ tendency to select their previous answers on the second test versus the corrective feedback. Those two alternative possibilities corresponded to the same option if the initial test response was correct.

This analysis produced five mutually exclusive second test possibilities (see Table 1): (a) a correct answer on the second test that matched the previous answer on the initial test (correct/previous answer); (b) a correct answer that was neither the previous answer nor the corrective feedback on the first test (correct/other); (c) an incorrect answer that matched the previous answer on the initial test (incorrect/previous answer) (d) an incorrect answer that matched the corrective feedback that was provided in the first test (incorrect/corrective feedback); and (e) an incorrect answer that was neither the previous answer nor the corrective feedback on the first test (incorrect/other). To compare the response type patterns between the MC and CR groups, we limited the analysis to responses that were incorrect on both tests, that is, the bottom three rows in Table 1. By doing so, we were able to compare endorsements of previous answers and corrective feedback separately, while holding constant the level of accuracy on both tests. In the CR group, some answers were assigned part marks (0.5). To enable classification of the answers as correct or incorrect, we converted all the part-mark scores to full scores (1) before running our analysis

We conducted a 2 (final test type: MC, CR) x 3 (final-test answer type: previous answer, corrective feedback, other) mixed-factor ANOVA with the probability of answering with each type of answers as the dependent variable. We found no significant main effect of final test type, *F* < 1. However, we found a significant main effect of final-test answer type *F*(2, 340) = 30.66, *p* < .001, ηp2 = .15. Paired-sample *t*-tests showed a comparable probability of answering with the corrective feedback (*M* = .23, *SD* = .30) and other (*M* = .29, *SD* = .36), *t*(171) = 1.64, *p* = .10, *d* = .19, but both probabilities were higher than that for the previous answer (*M* = .08, *SD* = .16), *t*(171) = 6.43, *p* < .001, *d* = .65, and *t*(171) = 7.24, *p* < .001, *d* = .78, respectively. Finally, we found a significant interaction between test and question types *F*(2, 340) = 32.71, *p* < .001, ηp2 = .16. For the MC group, paired-sample *t*-tests revealed that participants overwhelmingly selected the corrective feedback from the initial test (*M* = .30, *SD* = .32) compared to other answer (*M* = .15, *SD* = .19), *t*(84) = -3.67, *p* < .001, *d* = .56, and previous answer (*M* = .10, *SD* = .17), *t*(84) = 5.62, *p* < .001, *d* = .80. However, in contrast to the MC group, participants in the CR group were more likely to produce other answers (*M* = .46, *SD* = .45) compared to corrective feedback (*M* = .15, *SD* = .23), *t*(86) = -5.63, *p* < .001, *d* = .85, and previous answer (*M* = .05, *SD* = .15), *t*(86) = -7.92, *p* < .001, *d* = 1.20.

**False Endorsements of Corrective Feedback.**The previous analysis demonstrated thatwhen participants made errors on both related questions on the two tests, they tended to avoid their previous answers. In contrast, the corrective feedback was falsely endorsed more frequently, particularly in the MC group. These data are consistent with Higham et al. (2016). To explore the pattern of in more detail, we conducted a second analysis that examined the rates of corrective feedback endorsements regardless of whether the response was correct on the first test.

To control for answer plausibility, we compared the feedback endorsement rates for related items to the analogous rate for new items for both the MC and CR groups. As questions were counterbalanced across conditions, a given final test question served as a related question for some participants but as a new question for other participants. When the question was assigned to the related condition, one option served as the corrective feedback on the first test. When the question was assigned to the new condition, no feedback was given, but it was still possible to determine the endorsement rate of the option that would have served as the corrective feedback if that question was assigned to the related condition. Thus, when we refer to the “corrective feedback” option in the following analyses, we are referring to the option that would have served as the corrective feedback in the related condition for that particular question, even though no feedback was provided when that question was assigned to the new condition. The endorsement rates are shown in Figure 4.

A 2 (final test type: MC, CR) x 2 (question type: new, related) mixed-factor ANOVA, with probability of answering with the corrective feedback as the dependent variable, yielded a significant main effect of final test type, *F*(1, 170) = 56.57, *p* < .001, ηp2 = .25. The probability of answering with the corrective feedback was higher in the MC (*M* = .20, *SD* = .12) compared to the CR group (*M* = .09, *SD* = .07). There was also a significant main effect of question type, *F*(1, 170) = 45.56, *p* < .001, ηp2 = .21, such that the probability of answering with the corrective feedback was higher for the related items (*M* = .19, *SD* = .17) compared to new items (*M* = .10, *SD* = .12). The interaction between final test type and question type was not significant, *F*(1, 170) = 2.38, *p* = .12, ηp2 = .01.

**Discussion**

For the repeated items, initial MC testing significantly enhanced accuracy compared to new items on both the MC and CR tests.However,no significant differences in accuracy were observed between the related and new items on either type of test. These findings did not replicate the results of Higham et al.’s (2016) study which showed impaired accuracy on the related items relative to new items if an MC test format was used in both the initial and final tests. Moreover, recall accuracy on our CR test did not replicate the results of Little et al.’s (2019) study which demonstrated better performance on related items compared to new ones, despite using the same materials. However, it should be noted that our procedure did not exactly replicate the procedures in either of these two studies. For instance, Higham et al. (2016) employed reading passages rather than general knowledge questions whereas Little et al. (2019) used an elimination MC format that encouraged high lure elaboration.

Despite the failure to replicate those aspects of the results, when we examined the probability of endorsing the corrective feedback between the related and new items, differences were observed. Specifically, participants erroneously endorsed the corrective feedback for related items on the second test more than for new items. Furthermore, this pattern was as evident on the CR test as well as the MC test. Thus, although there was no overall difference in related versus new item accuracy, there was still evidence of an automatic memory influence leading to errors on both the MC and CR tests when the data were analyzed in greater depth.

The analysis of final-test response types (given an incorrect initial answer) shed some light on the nature of these automatic processes. First, it is noteworthy that not all familiar options were endorsed at the same rate on the second MC test. For example, previous wrong answers were largely avoided, a result that replicates Higham et al.’s (2016) finding. Thus, it was not the case that all options on the first test were equally primed or activated leading to undifferentiated feelings of familiarity. Instead, participants appeared to be biased to selecting the option that was previously correct rather than incorrect. In contrast, participants tended to avoid *both* the previously incorrect answer and the corrective feedback on the CR test. Instead, they responded with a new answer.[[2]](#footnote-3)

There are several possible reasons for these different patterns of responding between the two test types. One possibility is that the presence of the options on the MC test, which always matched those presented during practice, acted as strong retrieval cues for the earlier presentation of the related questions. The strong match between the related pairs may have falsely led participants to believe that the question was repeated, which would warrant choosing the corrective feedback from the first test. The absence of the matching options on the CR test may have led participants to notice that the question stem had changed, and that a new answer was required. Importantly, though, if participants detected at a high rate that the question stem had changed on the CR test, it was not enough to control the automatic influence of memory. As noted above, CR participants still chose the corrective feedback for related questions at a higher rate than for the new questions.

If participants were falsely endorsing corrective feedback with related items on the second test because of automatic memory influences, then why was there no corresponding difference in accuracy on related versus new questions with either final test type? One possibility is that controlled elaborative retrieval processes occurred with some items on the first test, which facilitated related-item accuracy on the second test. In other words, the presence of facilitative controlled memory influences may have prevented the deleterious automatic memory influences from being clearly expressed in the accuracy measure.

**Experiment 2**

The results of Experiment 1 showed clear evidence of automatic memory influences: For related questions on the second test, participants endorsed the corrective feedback from the initial test even though the question had changed and the feedback was no longer the correct answer. However, we did not find direct evidence of controlled processes in Experiment 1; accuracy did not differ between the related and new items on the CR test, which would have been evidence that elaborative retrieval processes during the initial test facilitated later test performance in a controlled manner (e.g., Little et al., 2012, 2019). Although we argued that controlled processes may have opposed the automatic influences, yielding a null net related-new accuracy difference, this evidence is speculative and weak. The aim of Experiment 2 was to find more direct evidence of controlled memory processes that might be working in conjunction with automatic influences in this paradigm.

To achieve this aim, the second experiment was designed to promote more elaborative retrieval processes during the initial test. We hypothesized that the use of standard MC questions in the initial test in Experiment 1 may not have provoked enough elaborative retrieval to enhance later test performance. Following Sparck et al. (2016), we reasoned that more elaborative retrieval might be encouraged by changing the answer format of the first test. Specifically, we required participants to rank order the four alternatives on the first test instead of merely choosing a favorite option as in Experiment 1 (standard testing format). The ranking format would potentially encourage participants to consider each option, retrieving evidence for or against it, to determine its place in the ranking. To further encourage deeper processing of the questions, we presented each question on the first test for a minimum duration instead of allowing participants to progress through the questions at their own pace.

In line with Sparck et al. (2016), we expected that more elaborative retrieval on the initial test would improve recall accuracy for related items on the CR tests. It is also possible that the new test format would enhance MC test accuracy as well. The enhanced elaborative retrieval processes might assist participants in noticing that related questions between the two experimental phases were different and required different answers. They may also encourage participants to engage in more controlled processing on the second test instead of relying on (false) automatic influences.

**Method**

***Participants***

Based on the power analysis reported in Experiment 1, we aimed to recruit a minimum of 128 participants. We actually tested 198 participants, but after reviewing participants’ performance and responses to the attention checks questions, 26 participants were excluded. The exclusion criteria included not engaging in the task by avoiding ranking any question during the first test, responding to the CR questions with random words such as “Yes” or “No,” looking up the answers by providing some detailed information that was taken from the internet in the CR test (although they were warned not to look up the answers). Thus, the final analysis included 172 participants (female= 68) with ages ranging between 20 to 60 years (*M* = 32.48, *SD* = 8.73) from the general population. They were recruited through Amazon Mechanical Turk (MTurk) and were given $2 in return for their participation in the study. The experiment involved two groups with 85 participants in the MC group and 87 in the CR group and 12 versions of the surveys (for counterbalancing purposes) with 6-9 participants each.

***Design and Materials***

This experiment employed the same design as Experiment 1. However, for the materials, in Experiment 1 we noticed that some questions were very easy and mostly answered correctly by all the participants across the different conditions. Therefore, we replaced nine general-knowledge questions (either both members of a pair or only one) with new ones that covered a variety of different general knowledge topics as well.

***Procedure***

The procedure was identical to that in Experiment 1 with some exceptions. First, we modified the standard MC format in the initial test to use ranking. Rather than selecting the correct answer, participants were asked to rank the four alternatives provided for each question from most to least plausible. Specifically, the question stem was presented with four alternatives below it and participants were required to drag each alternative and drop it to the desired rank position with answers in positions 1 versus 4 being the most versus least likely to be correct. Second, to discourage participants from completing the initial test too quickly, participants were not permitted to advance to the next question until 15 s had elapsed. After 15 s, the “Next” button was shown which allowed participants to move on to the corrective feedback. The feedback was provided regardless of whether the top ranked answer was correct or not (e.g., “*The correct answer is Supernova*”). Clicking “Next” again advanced to the next question. The filler task and the final test were identical to those used in Experiment 1.

For the scoring, the initial ranking test was scored as either correct (1) if the correct answer was placed in the top position, or wrong (0) if the correct answer was ranked in the other three positions (2, 3, 4). For scoring the final MC and CR tests, the scoring method was identical to that in Experiment 1.

**Results**

***Initial Test Performance***

All participants completed the initial MC ranking test and were treated the same at this stage of the experiment. Nonetheless, we compared the response accuracy of the two experimental groups to test for sampling error. The mean proportion of questions answered correctly on the first MC test in the MC and CR groups was .59 (*SD* = .20) and .61 (*SD* = .19), respectively. An independent sample *t*-test indicated that the difference was not significant, *t*(169) = -0.45, *p* = .65, *d = .*07*,* suggesting that the CR and MC groups were comparable.

***Final Test Performance***

**Final-Test Accuracy**. We conducted a 2 (final test type: MC, CR) x 3 (question type: new, repeated, related) mixed-factor ANOVA with final test accuracy as the dependent variable (see Figure 5). The analysis revealed a non-significant effect of final-test type *F*(1, 170) = 3.08, *p* = .08, ηp2 = .02, such that accuracy was comparable between the MC group (*M =* .58, *SD =* .15) and the CR group (*M =* .54, *SD =* .17). There was a significant main effect of question type, *F*(2, 340) = 285.30, *p* < .001, ηp2 = .63. Paired-sample *t*-tests revealed that accuracy was comparable between the new (*M* = .44, *SD* = .24) and related items (*M* = .41, *SD* = .24), *t*(171) = -1.74, *p* = .08, *d* = .15, but significantly higher on the repeated items (*M* = .82, *SD =* .19) compared to the new *t*(171) = -20.38, *p* < .001, *d* = 1.80, and related items *t*(171) = -19.32, *p* < .001, *d* = 1.95.

The main effect of question type was qualified by a significant interaction with final test type, *F*(2, 154) = 19.59, *p* < .001, ηp2 = .10. Paired-sample *t*-tests revealed that the interaction was due to the CR group performing better on the related items (*M* = .45, *SD* = .23) than the new items (*M* = .39, *SD* = .22), *t*(86) = 2.57, *p* = .01, *d* = .29, and significantly better on the repeated items (*M* = .77, *SD* = .19) compared to new items, *t*(86) = -14.71, *p* < .001, *d* = 1.82, and related items, *t*(86) = -11.89, *p* < .001, *d* = 1.49. In contrast, participants in the MC group performed better on the new items (*M =* .50, *SD =* .23) than the related items (*M =* .36, *SD =* .24), *t*(84) = -4.77, *p* < .001, *d* = .59, and significantly better on the repeated items (*M =* .88, *SD =* .16) compared to new items, *t*(84) = -14.06, *p* < .001, *d* = 1.92, and related items, *t*(84) = -17.28, *p* < .001, *d* = 2.56.

**Final-Test Answer Types**. We conducted the same analysis as in Experiment 1, conditioning the final test answers on incorrect initial test answers, producing the same five mutually exclusive possibilities (see Table 2). As before, we converted part marks in the CR group to full marks so that final test responses could be classified as either correct or incorrect. Then we conducted a 2 (final test type: MC, CR) x 3 (final-test answer type: previous answer, corrective feedback, other) mixed-factor ANOVA with the probability of answering with each type of answers as the dependent variable. The main effect of final test type was not significant, *F* < 1. However, we found a significant main effect of final-test answer type *F*(2, 340) = 40.20, *p* < .001, ηp2 = .19. Paired-sample *t*-tests showed a comparable probability of answering with the corrective feedback (*M* = .32, *SD* = .39) and other option (*M* = .28, *SD* = .36), *t*(171) = -0.86, *p* = .38, *d* = .10, but both probabilities were higher than that for the previous answer (*M* = .07, *SD* = .15), *t*(171) = 8.13, *p* < .001, *d* = .85, and *t*(171) = 7.08, *p* < .001, *d* = .78, respectively. Also, we found a significant interaction between test type and answer type *F*(2, 340) = 67.62, *p* < .001, ηp2 = .28. For the MC group, paired-sample *t*-tests revealed that participants overwhelmingly selected the corrective feedback from the initial test (*M* = .50, *SD* = .45) compared to other answer (*M* = .11, *SD* = .17), *t*(84) = -7.39, *p* < .001, *d* = 1.14, and previous answer (*M* = .11, *SD* = .17), *t*(84) = -7.37, *p* < .001, *d* = 1.14. For the CR group, participants were more likely to produce other answers (*M* = .46, *SD* = .41) compared to corrective feedback (*M* = .15, *SD* = .19), *t*(86) = 6.60, *p* < .001, *d* = .97, and previous answers (*M* = .03, *SD* = .11), *t*(86) = -8.90, *p* < .001, *d* = 1.40.

**False Endorsements of Corrective Feedback.** As in Experiment 1, we analyzed theprobability of endorsing the corrective feedback between the related and new items. We conducted a 2 (final test type: MC, CR) x 2 (question type: new, related) mixed-factor ANOVA with the probability of answering with the corrective feedback as the dependent variable (see Figure 6). The ANOVA revealed a significant main effect of final test type, *F*(1, 170) = 147.7, *p* < .001, ηp2 = .46. The probability of answering with the corrective feedback was higher in the MC group (*M* = .34, *SD* = .15) compared to the CR group (*M* = .11, *SD* = .09). We also found a significant main effect of question type, *F*(1, 170) = 89.34, *p* < .001, ηp2 = .34. The probability of answering with the corrective feedback was higher for the related items (*M* = .32, *SD* = .28) compared to new (*M* = .12, *SD* = .13). Finally, we found a significant interaction between final test type and question type, *F*(1, 170) = 12.04, *p* < .001, ηp2 = .07.[[3]](#footnote-4) Paired-sample *t*-tests revealed that participants in both groups were more likely to endorse the corrective feedback for related items compared to new items, but the difference was larger in the MC group (related: *M* = .47, *SD* = .29; new: *M* = .21, *SD* = .14), *t*(84) = 7.29, *p* < .001, *d* = 1.16, compared to the CR group (related: *M* = .17, *SD* = .16; new: *M* = .04, *SD* = .07), *t*(86) = 6.39, *p* < .001, *d* = .99.

**Discussion**

Accuracy with the repeated items was significantly enhanced compared to the related and new items for both the MC and CR final tests. The MC results in this experiment replicated the poor performance on related versus new items observed by Higham et al. (2016); that is, initial MC testing harmed later test accuracy for the related items in the MC group. In contrast, instead of impaired performance, the CR group showed facilitation for related versus new items. The CR results, therefore, were consistent with Little et al.’s (2019) findings. Consistent with Sparck et al.’s (2016) account, using the more intensive ranking MC format in Experiment 2, instead of standard MC testing in Experiment 1, helped participants to retrieve more information about the lures. This retrieval enhanced participants’ overall performance with related items on the second test.

Like Experiment 1, there was a higher probability of answering with the corrective feedback on related questions compared to new ones. This effect was evident on both the final MC test and the final CR test. However, a significant interaction indicated that the tendency to endorse corrective feedback over baseline was greater on the MC final test than the CR final test, which did not occur in Experiment 1. Moreover, the analysis of final-test answer types showed that if participants answered incorrectly on both tests, MC responses on the second test tended to be the corrective feedback from the first test. Conversely, participants in the CR group tended to produce an incorrect response on the second test that was different from both their previous answer and the corrective feedback on the first test. These results replicate those obtained in Experiment 1.

Our attempt at increasing elaborative retrieval during the initial test by using a ranking format rather than a standard one produced clear benefits to controlled processing as the results showed facilitation for related versus new items on the CR test. However, the results suggest that the CR performance was affected as well by automatic influences, as evidenced by corrective feedback intrusions in CR. On the other hand, the difference in accuracy between related and new items on the MC test was larger (and statistically significant) in Experiment 2 compared to Experiment 1 (where the difference was not significant). Also, the tendency to respond with corrective feedback on related versus new questions on the MC test was greater in Experiment 2 (*M*diff = .26) than Experiment 1 (*M*diff = .07). Thus, rather than the ranking format increasing controlled influences on the final MC test – influences that we expected would counter automatic familiarity influences – it appears to have enhanced automatic influences instead.

**Experiment 3**

Although the ranking format used on the first test in Experiment 2 enhanced CR accuracy of the related versus new items on the final test, the effect size was relatively small (Cohen’s *d* = 0.29). Thus, perhaps using a more intensive format on the first test than the ranking format used in Experiment 2 would increase retrieval sufficiently to enhance the related versus new facilitation. Therefore, in Experiment 3, we boosted elaborative retrieval further by using elimination testing as in the initial test of Little et al.’s (2019) study. That is, participants were required to identify the chosen option and provide reasons for rejecting the unchosen options. Unlike ranking, which could be achieved by focusing mostly on just the chosen option, an elimination format should encourage participants to also retrieve some information about all the unchosen options. We therefore expected to observe a sizable difference in accuracy between related and new items on the final CR test. A similar finding might be observed in the MC group if the increase in controlled processing counters the automatic influences. However, given the results of Experiment 2, it may be that more elaborative retrieval during the first test backfires and results in greater automatic influences and an even larger difference between related and new items.

**Method**

***Participants***

Based on the power analysis reported in Experiment 1, we aimed to recruit at least 128 participants. We actually tested 246 participants, but due to the demanding elimination format used in this experiment, 84 participants were excluded.[[4]](#footnote-5) Asking participants to write out a reason to reject each lure encouraged more participants in this experiment to use shortcuts instead of engaging in the task. For example, several participants wrote random words or marks that did not relate to the task in the first or final test (e.g., “Yes,” “No,” and punctuation marks). Some participants did not engage in the task at all by keeping the text boxes empty in the first test. Others looked up the answers and provided very long detailed explanations that were taken from the internet, despite our warning not to look up the answers. Thus, the final sample comprised 162 participants (female = 60) with ages ranging between 23 to 59 years (*M* = 32.96, *SD* = 9.66) from the general population. Eighty-one participants were randomly assigned to each of the MC and CR final test groups. Each group had 6-9 participants in each of 12 versions of the survey (for counterbalancing purposes). Participants were recruited through Amazon Mechanical Turk (MTurk) and were given $2 in return for their participation in the study.

***Design and Material******s***

This experiment employed the same design as Experiment 1 and used the same general knowledge questions as Experiment 2.

***Procedure***

The procedure was mostly identical to Experiment 2 with a few exceptions. First, instead of using the ranking format for the initial test, we replicated Little et al.’s (2019, Experiment 1) design and used an elimination format. On the initial test, each question stem came with four alternatives and below that, we added four empty boxes where participants were asked to type the chosen option in the first box with no need to provide any reason for this option. However, participants were asked to write the three unchosen options and their reasons for rejecting each one in the three remaining boxes. Prior to the initial test, participants were given an example question of how they should respond to the questions in the initial test. Like Experiment 2, each question in the initial test was displayed for 15 seconds before the “Next” button was displayed, which allowed participants to advance to the next question. Participants were not forced to provide reasons for rejecting each lure, so they could advance after the 15 seconds had elapsed without the need to provide a reason for each rejected lure. The rest of the procedure was identical to that of Experiments 1 and 2.

For the scoring, the initial test was scored as either correct (1) if the correct option was provided in the first box or incorrect (0) if the chosen option was wrong. For scoring the final MC and CR tests, the scoring method was identical to the scoring used in Experiments 1 and 2.

**Results**

***Initial Test Performance***

For participants who were included in the final analysis, we found that participants’ responses on the first test indicated that they differed in their conformity to the instructions. Sixty percent of the participants achieved high engagement in the task by providing reasons to reject most of the lures and even when they could not provide any reasons, they illustrated their lack of knowledge about the question (e.g., “*I do not know anything about this question*”). Conversely, 40% of the participants engaged less in the elimination process by merely identifying the correct and incorrect answers and only occasionally providing some reasons.

To test for sampling error, we again compared the two groups on initial test accuracy even though they were treated the same at that point in the experiment. The mean proportion of initial MC test questions answered correctly was .69 (*SD* = .19) and .66 (SD = .20) for the MC and CR groups, respectively. An independent sample *t*-test showed that the difference was not significant *t*(160) = 1.18, *p* = .24, *d = .*19*.* Thus, as in Experiments 1 and 2, the MC and CR groups were comparable.

***Final Test Performance***

**Final-Test Accuracy.** We conducted a 2 (final test type: MC, CR) x 3 (question type: new, repeated, related) mixed-factor ANOVA with final-test accuracy as the dependent variable (see Figure 7). The main effect of final test type was not significant, *F*(1, 160) = 1.09, *p* = .30, ηp2 = .01; accuracy was comparable across the MC (*M =* .57, *SD =* .19) and CR (*M =* .54, *SD =* .18) tests. However, there was a significant main effect of question type, *F*(2, 320) = 227.8, *p* < .001, ηp2 = .59. Paired-sample *t*-tests revealed that accuracy was comparable between the new (*M =* .43, *SD =* .25) and related items (*M =* .43, *SD =* .26), *t*(161) = -0.30, *p* = .76, *d = .*03, but significantly higher on the repeated items (*M =* .81, *SD =* .22) compared to the related items, *t*(161) = 16.76, *p* < .001, *d =* 1*.*60, and new items, *t*(161) = 17.66, *p* < .001, *d =*1 *.*61. The main effect of question type was qualified by a significant interaction between final test type and question type, *F*(2, 320) = 21.7, *p* < .001, ηp2 = .21. Paired-sample *t*-tests revealed that the interaction was due to the CR group performing better on the related items (*M =* .49, *SD =* .22) than the new items (*M =* .37, *SD =* .23), *t*(80) = 5.16, *p* < .001, *d = .*51, and significantly better on the repeated items (*M =* .76, *SD =* .21) compared to the new items, *t*(80) = -12.13, *p* < .001, *d =*1 *.*75, and the related items, *t*(80) = -9.97, *p* < .001, *d =*1 *.*25. In contrast, participants in the MC group performed better on the new items (*M =* .49, *SD* = .25) than the related items (*M =* .36, *SD* = .28), *t*(80) = -4.22, *p* < .001, *d = .*48, and significantly better on the repeated items (*M =* .86, *SD =* .20) compared to the new items, *t*(80) = 12.89, *p* < .001, *d =*1 *.*60, and the related items *t*(80) = 15.23, *p* < .001, *d =* 2*.*01.

**Final-Test Answer Types**. We conducted the same analysis on final-test answer types as in Experiments 1 and 2, producing the same five mutually exclusive second test response rates for items incorrectly answered on the initial test. As before, part marks in the CR group were converted to full marks so that responses could be classified as either correct or incorrect. We then analyzed the three possibilities where the answers were incorrect on both the first and final tests (see Table 3). A 2 (final test type: MC, CR) x 3 (final-test answer type: previous answer, corrective feedback, other) mixed-factor ANOVA, with the probability of answering with each type of answer as the dependent variable, revealed that the main effect of final test type was not significant, *F* < 1*.* However, there was a significant main effect of final-test answer type *F*(2, 320) = 26.39, *p* < .001, ηp2 = .14. Paired-sample *t*-tests showed a comparable probability of answering with the corrective feedback (*M* = .28, *SD* = .40) and other (*M* = .28, *SD* = .37), *t*(161) = 0.04, *p* = .97, *d = .*00, but both probabilities were higher than that for the previous answer (*M* = .07, *SD* = .16), *t*(161) = 6.31, *p* < .001, *d = .*68, and *t*(161) = 6.61, *p* < .001, *d = .*73, respectively. Moreover, we found a significant interaction between test type and answer type *F*(2, 320) = 39.11, *p* < .001, ηp2 = .20. For the MC group, paired-sample *t*-tests revealed that participants overwhelmingly selected the corrective feedback from the initial test (*M* = .43, *SD* = .48) compared to other answer (*M* = .14, *SD* = .21), *t*(80) = -5.43, *p* < .001, *d = .*81, and previous answers (*M* = .11, *SD* = .21), *t*(80) = -5.62, *p* < .001, *d = .*89. For the CR group, participants were more likely to produce other answer (*M* = .42, *SD* = .42) compared to corrective feedback (*M* = .14, *SD* = .24), *t*(80) = 5.24, *p* < .001, *d = .*82, and previous answer (*M* = .04, *SD* = .10), *t*(80) = -7.73, *p* < .001, *d =* 1.24.

**False Endorsements of Corrective Feedback.** As in Experiments 1 and 2, we analyzed theprobability of endorsing the corrective feedback on related and new questions with a 2 (final test type: MC, CR) x 2 (question type: new, related) mixed-factor ANOVA (see Figure 8). It revealed a significant main effect of final-test type, *F*(1, 160) = 142.5, *p* < .001, ηp2 = .47. The probability of answering with the corrective feedback was higher in the MC group (*M* = .35, *SD* = .18) compared to the CR group (*M* = .08, *SD* = .07). It also revealed a significant main effect of question type, *F*(1, 160) = 76.01, *p* < .001, ηp2 = .32, such that the probability of answering with the corrective feedback was higher for the related items (*M* = .30, *SD* = .28) compared to the new ones (*M* = .13, *SD* = .16). Both of these main effects were qualified by a significant interaction, *F*(1, 160) = 5.55, *p* = .02, ηp2 = .03.[[5]](#footnote-6) Paired-sample *t*-tests revealed that participants in both groups were more likely to endorse the corrective feedback on related questions compared to new, but the difference was larger in the MC group (related: *M* = .45, *SD* = .30; new: *M* = .24, *SD* = .17), *t*(80) = -6.04, *p* < .001, *d = .*89, compared to the CR group (related: *M* = .14, *SD* = .14; new: *M* = .02, *SD* = .04), *t*(80) = 7.97, *p* < .001, *d =* 1.25.

**Discussion**

As in both previous experiments, accuracy for the repeated items was better than that for either the related or new items on both the MC and CR tests. However, participants in the CR group had significantly higher accuracy on the related items compared to the new items, replicating Experiment 2 and Little et al.’s (2019) results. Boosting the level of lure elaboration that occurred during the first test with the elimination format in Experiment 3 not only result in enhanced performance on the related versus new items but a larger effect size as well (Experiment 2: *d* = 0.29; Experiment 3: *d* = 0.51). These results provide solid evidence of controlled processing. However, there was also evidence for automatic influences of memory on the CR test as well. Specifically, the corrective feedback intruded significantly more often with related items compared to new items. Thus, the data suggest that both types of influence are present on the CR test at the same time, but if the test format encourages elaborative retrieval during the initial test, the controlled influences override the automatic ones.

In contrast, the enhanced lure elaboration associated with elimination testing did nothing to temper automatic influences on the final MC test. In fact, the accuracy difference between the related and new questions was as great (or greater) in this experiment (*M*diff = .13) as it was in the previous two experiments (Experiment 1: *M*diff = .04; Experiment 2: *M*diff = .14). The feedback analysis also showed that the enhanced likelihood of endorsing the corrective feedback for related questions compared to new ones was again comparable for MC participants in this experiment (*M*diff = .21) as in the previous ones (Experiment 1: *M*diff = .07; Experiment 2: *M*diff = .26). Moreover, the final-test answer type analysis showed that MC participants still showed a tendency to respond with corrective feedback when answers on the earlier test were wrong. Together, these results suggest that even if it is ensured that participants are processing practice test questions at a deep level, it does not guarantee that there will be positive transfer of controlled processes to a later test. If the later test is MC, lure elaboration can even worsen the deleterious automatic effects of false familiarity (cf. related vs. new difference in Experiment 1 vs. Experiment 3).

**General Discussion**

Over three experiments, we examined the positive and negative consequences of taking an MC practice test on a later final test that was either in MC or CR format. In all three experiments, repeated items showed significantly better performance compared to the related and new items regardless of the type of final test. This finding adds to a large literature demonstrating that retrieval practice is an effective learning tool (e.g., see reviews by Rowland, 2014; Yang et al., 2021).

**Related Versus New Questions**

Of greater interest to the current research was how participants performed on the final test with questions related to ones answered (with feedback) from the practice phase versus new questions. Here, the effects of retrieval practice were more equivocal, and depended on the format of both the initial and final tests. In Experiment 1, which used the standard “choose the one best option” MC format during practice, there were no differences in accuracy between the related and new items on the CR final test. Following Little and colleagues (e.g., Little & Bjork, 2015; Little et al., 2012; Sparck et al., 2016), we attributed the failure to observe a difference in CR to low levels of lure elaboration during the practice test, which would have limited any controlled memory influences on the later test. Unless the practice test encourages retrieval of information about the lures, either by the lures being competitive (Little & Bjork, 2015; Little et al., 2012) or by the practice test format encouraging deep retrieval (Sparck et al., 2016), then accuracy on related test items has not typically been enhanced relative to new items. Therefore, a standard initial MC may not be enough to promote retrieval and produce controlled influences, especially if it is self-paced as in Experiment 1, which might have encouraged participants to finish the task very quickly without really taking the time to consider each of the lures. Indeed, in an analysis of the time participants spent completing each experiment, we found that the average time increased monotonically across experiments, from 18 min in Experiment 1, to 27 min in Experiment 2, to 34 min in Experiment 3. Independent sample *t*-tests showed that each successive time increase across experiments was significant (Experiment 1 vs. 2: *t*(329) = -6.52, *p* < .001, *d* = .72; Experiment 2 vs. 3: *t*(313) = -5.35, *p* < .001, *d* = .60).

On the other hand, there was also evidence of automatic influences on the final CR test. Specifically, CR participants were more likely to intrude with the corrective feedback when answering related versus new questions, but this tendency was not strong enough to cause a difference in overall accuracy. Indeed, if both related questions were answered incorrectly, CR participants had a general tendency to intrude with novel words rather than their previous answer or the corrective feedback.

For the final MC test in Experiment 1, we also found no difference in accuracy between the related items compared to new items. This result was somewhat surprising given Higham et al.’s (2016) results. They found across a variety of experimental contexts that accuracy on related questions suffered because participants tended to erroneously endorse the option on the second test that corresponded to the first test’s corrective feedback. However, a closer analysis of the data suggested that this tendency did exist in the MC group. That is, the probability of specifically selecting the corrective feedback was higher for related questions than new questions. Moreover, an analysis of the types of second test responses participants made after answering the practice question incorrectly showed that corrective feedback was the favorite choice. However, this tendency was not strong enough to manifest itself in accuracy differences between related and new items.

In Experiment 2 we exchanged the standard MC question format that we used on the initial test in Experiment 1 for a ranking format more likely to evoke lure elaboration. Indeed, Sparck et al. (2016) found that using a confidence weighted format instead of standard MC format increased the overall performance not only for the repeated questions but for the related untested questions as well. Also, participants in Experiment 2 were forced to wait for 15 s before they could advance to the next question to encourage deep retrieval. As expected, using this question format enhanced the performance for the related CR items. That is, related questions were answered more accurately than new questions which reflects the importance of increasing the level of lure elaboration during the initial test.

For the MC test, however, the switch to a ranking format for the first test in Experiment 2 rendered answers to related items being significantly lower compared to new items, replicating Higham et al. (2016). Furthermore, analyzing the probability of selecting the corrective feedback showed a large increase in selecting the corrective feedback as an answer for the related items in Experiment 2 (*M =* .47, *SD =* .29) compared to Experiment 1 (*M =* .24, *SD =* .20), *t*(147) = 6.00, *p* < .001, *d* = .93. As in Experiment 1, MC participants overwhelmingly preferred the corrective feedback when responses to both related questions were incorrect, even more in Experiment 2 (*M =* .50, *SD =* .45) than Experiment 1 (*M =* .30, *SD =* .32), t(145) = 3.61, *p* < .001, d = .50. These findings are quite surprising in that the ranking format likely fostered at least somewhat more elaborative retrieval during the practice test than standard testing, which should have allowed participants to used controlled processes to counter automatic influences on the second test. One possible reason for this result might be that ranking caused participants to focus on encoding the corrective feedback, perhaps because of enhanced curiosity (e.g., see Potts et al., 2019), whilst mostly ignoring the question stem. Consequently, when the related questions were shown during the second test, the familiar corrective feedback was highly available in memory, but the fact that questions had changed was not noticed, resulting in multiple selections of the corrective feedback option.

In Experiment 3, we increased the level of lure elaboration further by using the elimination technique on the first test (e.g., see Little et al., 2019). Sure enough, by using the elimination format, we replicated Experiment 2 and Little et al.’s (2019) results as there was better accuracy on related items versus new items in CR. Also, by using this question format, we increased the effect size of the related versus new difference from a small effect in Experiment 2 (Cohen’s *d* = .29) to a medium effect in Experiment 3 (Cohens *d* = .51).

However, the greater lure elaboration during practice did not produce controlled influences that tempered the automatic influences on the MC final test. Participants still erroneously endorsed the corrective feedback at a rate high enough to produce a related versus new item accuracy deficit comparable to that observed in Experiment 2. Indeed, both ranking and elimination practice testing *reduced* accuracy on related items on the second MC test (*M =* .36, *SD =* .24; and *M =* .36, *SD =* .28, respectively) compared to standard MC practice format (*M* = .58, *SD =* .23), *F*(2, 248) = 22.21, *p* < .001, ηp2 = .15 (cf. Figures 2, 5, and 7).

**Theoretical Account of the Results**

One might have expected that the elaborative retrieval during practice, particularly in Experiments 2 and 3 where it was enough to facilitate related versus new question accuracy on the later CR test, might also have produced controlled influences that would have assisted MC participants. For example, participants who retrieved deeply during practice might have encoded the details of the question more thoroughly and been more likely to notice that the that related second-test question was not a repetition of the earlier related question. Alternatively, retrieval of information about the options during the first test might have facilitated retrieval about those options during the second test, allowing participants to avoid the corrective feedback and home in on the new correct answer. Indeed, following Little and colleagues (e.g., Little et al., 2012), a controlled retrieval process like this is our explanation for why related item accuracy was better than new item accuracy for CR participants in Experiments 2 and 3. Instead, however, it seems that standard MC testing, such as that used on the second test, does not draw on earlier elaborative processing in the same way as CR testing. Future research might focus on exploring controlled influences with different second-test MC formats. It may be that standard MC testing formats invite quick, familiarity-based responding even in cases where prior testing has activated prior knowledge that would oppose the familiar response. For example, if elimination format or confidence weighting format (Sparck et al., 2016) was used on both MC tests, performance with related questions might improve and exceed that with new questions, as observed in CR.

On the other hand, it may be that no MC testing format would yield good related-question performance on the second test. Some researchers have argued (e.g., Ozuru et al., 2013) that recollection tests such as CR present very limited information and therefore promote generation of several internal cues related to the desired information in a controlled multi-step retrieval process. On the other hand, recognition tests such as MC present the candidate answers as part of the question. Thus, in many cases, the retrieval process is shifted from generation of internal cues to assist in retrieving the to-be-remembered information to reliance on the relative familiarity of the options.

Although lure elaboration during practice facilitated controlled influences, it had no effect on automatic influences. This was true not just when the second test was MC, but also when it is CR; that is, participants’ tendency to intrude with corrective feedback on related versus new questions in CR remained invariant across the three experiments (*M*diffs = .12, .13, .12 for Experiments 1, 2 and 3, respectively). Such a finding is consistent with reports of dissociations between controlled and automatic influences in a variety of contexts including memory (e.g., Jacoby, 1991, 1996, 1999; Jacoby et al., 1989), social psychology (e.g., Payne, 2001, 2008), education (e.g., Ozuru et al., 2013), and classification tasks (e.g., Higham & Vokey, 2000; Higham et al., 2000). Often (but not always) these dissociations take the form of some factor affecting the controlled process, but leaving the automatic one invariant (e.g., Jacoby, 1998; Stolz & Merikle, 2000; Toth et al., 1994), which is the type of dissociation we observed here.

However, although the automatic influences were invariant, we noted earlier that not all familiar options were treated the same in MC testing. Previous answers on incorrectly answered practice questions were as familiar as corrective feedback, but they were generally avoided if answers on both tests were incorrect. This result suggests that the automatic memory influence in the MC group did not take the form of general, undifferentiated activation of all repeated options. Instead, the pattern is more consistent with participants sometimes believing that related questions were repeated ones (i.e., false recognition), which if true, would warrant selecting the corrective feedback.

However, it’s important to emphasize that false recognition of MC questions is not the only form that automatic influences took in these experiments. CR participants in Experiment 1, for example, also tended to erroneously produce the corrective feedback more for related questions than new questions and this tendency was greater than for the MC test (*M*diff = .12 vs. .07, respectively; see Figure 4). This tendency to endorse the feedback for related questions on the CR test occurred even though participants tended to avoid the corrective feedback when they had answered the related question incorrectly during practice. Avoiding the corrective feedback suggests that participants did not believe the question was repeated, mostly likely because they were not misled into this assumption by the presence of repeated response options on the CR test. Additionally, in cases where the second test was MC, Higham et al. (2016) removed repeated items from one of their experiments altogether, and more recently, Alamri and Higham (2022) manipulated the presence/absence of repeated items in a multi-group experiment. In both cases, the researchers observed a slight residual tendency for participants to select the corrective feedback more often for related versus new questions even when there were no items repeated between the tests.

**Relationship to the Negative Testing Effect**

On the surface, readers might conclude that the poor performance we observed with related questions that we have attributed to automatic memory influences is just another case of the negative testing effect, first reported by Roediger and Marsh (2005). However, although the original negative testing effect can be classified as a type of automatic memory error similar to what we observed in our research (i.e., false recollection on the CR test), there are some important differences. The original negative testing effect occurred when participants erroneously selected lures during practice, received no feedback on their incorrect selections, and then were required to answer the same question again in CR format. By selecting the lure, it gained familiarity, and selected lures sometimes intruded on a later CR test, particular in cases where the practice test questions had many lures. Importantly, Butler and Roediger (2008) showed that providing feedback during practice reduced the negative testing effect (and increased the positive testing effect), leading them to advise educators to provide feedback during MC practice tests. In contrast, the automatic memory influence on test performance investigated in our research is *caused* by corrective feedback, not reduced by it. If the questions on the final test were repeated, as in the original negative testing effect scenario, then repeated responding with the feedback would have enhanced performance. However, if related questions are used that have different correct answers, then responding with the corrective feedback on the final test harms performance rather than enhancing it. Thus, with the original negative testing effect, the problem is repeating uncorrected errors. In contrast, the effect we have investigated is caused by responding with feedback when it is no longer appropriate to do so, which reveals a dark side to corrective feedback not previously explored.

**Conclusions**

Practice answering MC questions can produce robust testing effects. Indeed, Yang et al.’s (2021) recent meta-analysis suggests that MC practice tests produce larger testing effects than practice tests involving recall, at least in real educational contexts. Therefore, if used properly, MC tests can be an effective learning tool. However, there are also pitfalls to avoid. Roediger and Marsh (2005) have reported that in retrieval practice scenarios where question stems are repeated between a MC practice test and a CR final test, previously selected lures can intrude on the CR test (negative testing effect). Our research demonstrates that additional problems can arise when both tests are MC and related questions are used. If MC tests with related questions (that have different correct answers) are to be used for both practice and final testing, we suggest ensuring that no option is repeated across the tests, particularly if feedback is provided during practice. Using a CR final test will also limit the problem, particularly if deep retrieval is required during the MC practice. Indeed, deep retrieval during practice can sometimes facilitate controlled processing and lead to positive testing effects, as both Little and colleagues (e.g., Little et al., 2012) and our Experiments 2 and 3 have shown. However, even under those circumstances, automatic influences can cause corrective feedback to intrude on the CR final test. Thus, designing related questions should be managed very carefully. We hope that our research will help in this regard by highlighting the kinds of situations that give rise to problems so that they can be avoided.

**References**

Alamri, A. A., & Higham, P. A. (2022). *Multiple-choice testing: Controlled and automatic influences of retrieval practice in an educational context.* Manuscript in preparation.

Butler, A. C. (2018). Multiple-choice testing in education: Are the best practices for assessment also good for learning? *Journal of Applied Research in Memory and Cognition, 7*(3), 323–331. https://doi.org/10.1016/j.jarmac.2018.07.002

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: 10.3758/mc.36.3.604

Cantor, A. D., Eslick, A. N., Marsh, E. J., Bjork, R. A., & Bjork, E. L. (2014). Multiple-choice tests stabilize access to marginal knowledge. *Memory & Cognition, 43(2)*, 193–205. https://doi.org/10.3758/s13421-014-0462-6

Chan, J. C. K., & McDermott, K. B. (2007). The testing effect in recognition memory: A dual process account. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33(2),* 431–437. https://doi.org/10.1037/0278-7393.33.2.431

Fellenz, M. R. (2004). Using assessment to support higher level learning: The multiple-choice item development assignment. *Assessment & Evaluation in Higher Education, 29(6)*, 703–719. https://doi: 10.1080/0260293042000227245

Gierl, M., Bulut, O., Guo, Q., & Zhang, X. (2017). Developing, analyzing, and using distractors for multiple-choice tests in education: A comprehensive review. *Review of Educational Research, 87(6)*, 1082-1116. https://doi: 10.3102/0034654317726529

Higham, P. A., & Vokey, J. R. (2000). The controlled application of a strategy can still produce automatic effects: Reply to Redington (2000). *Journal of Experimental Psychology: General, 129(4),* 476–480. https://doi.org/10.1037//0096-3445.129.4.476

Higham, P. A., Vokey, J. R., & Pritchard, J. L. (2000). Beyond dissociation logic: Evidence for controlled and automatic influences in artificial grammar learning. *Journal of Experimental Psychology: General, 129(4)*, 457–470. https://doi.org/10.1037/0096-3445.129.4.457

Higham, P. A., Griffiths, L. & Rackstraw, H. (2016, November 17-20). *How can it be wrong when it feels so right? Responding correctly on multiple-choice practice tests can negatively transfer to later tests* [Conference presentation]. 57th Annual Meeting of the Psychonomic Society. Boston: MA, United States.

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language, 30(5),* 513–541. https://doi.org/10.1016/0749-596X(91)90025-F

Jacoby, L. L. (1996). Dissociating automatic and consciously controlled effects of study/test compatibility. *Journal of Memory and Language, 35(1),* 32–52. https://doi.org/10.1006/jmla.1996.0002

Jacoby, L. L. (1998). Invariance in automatic influences of memory: Toward a user's guide for the process-dissociation procedure. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24(1)*, 3–26. https://doi.org/10.1037/0278-7393.24.1.3

Jacoby, L. L. (1999). Ironic effects of repetition: Measuring age-related differences in memory. *Journal of Experimental Psychology: Learning Memory and Cognition*, *25*(1), 3–22. https://doi.org/10.1037/0278-7393.25.1.3

Jacoby, L. L., Kelley, C., Brown, J., & Jasechko, J. (1989). Becoming famous overnight: Limits on the ability to avoid unconscious influences of the past. *Journal of Personality and Social Psychology*, *56*(3), 326–338. https://doi.org/10.1037/0022-3514.56.3.326

Jacoby, L. L., Toth, J. P., & Yonelinas, A. P. (1993). Separating conscious and unconscious influences of memory: Measuring recollection. *Journal of Experimental Psychology: General*, *122*(2), 139–154. https://doi.org/10.1037/0096-3445.122.2.139

Little, J. L., & Bjork, E. L. (2015). Optimizing multiple-choice tests as tools for learning. *Memory & Cognition, 43(1)*, 14–26. https://doi.org/10.3758/s13421-014-0452-8

Little, J. L., & Bjork, E. L. (2016). Multiple-choice pretesting potentiates learning of related information. *Memory and Cognition*, *44*(7), 1085–1101. https://doi.org/10.3758/s13421-016-0621-z

Little, J. L., Bjork, E. L., Bjork, R. A., & Angello, G. (2012). Multiple-choice tests exonerated, at least of some charges. *Psychological Science, 23(11)*, 1337–1344. https://doi: 10.1177/0956797612443370

Little, J. L., Frickey, E. A., & Fung, A. K. (2019). The role of retrieval in answering multiple-choice questions. *Journal of Experimental Psychology: Learning, Memory, and Cognition. 45(8),* 1473–1485. http://dx.doi.org/10.1037/xlm0000638

Marsh, E. J., Agarwal, P. K., & Roediger, H. L. (2009). Memorial consequences of answering SAT II questions. *Journal of Experimental Psychology: Applied, 15(1)*, 1–11. https://doi.org/10.1037/a0014721

Marsh, E. J., Fazio, L. K., & Goswick, A. E. (2012). Memorial consequences of testing school-aged children. *Memory, 20(8),* 899–906. https://doi.org/10.1080/09658211.2012.708757

Marsh, E. J., Roediger, H. L., Bjork, R. A., & Bjork, E. L. (2007). The memorial consequences of multiple-choice testing. *Psychonomic Bulletin & Review, 14(2)*, 194–199. https://doi.org/10.3758/BF03194051

McDaniel, M. A., Thomas, R. C., Agarwal, P. K., McDermott, K. B., & Roediger, H. L. (2013). Quizzing in middle-school science: Successful transfer performance on classroom exams. *Applied Cognitive Psychology, 27(3)*, 360–372. https://doi.org/10.1002/acp.2914

‌McDaniel, M. A., Wildman, K. M., & Anderson, J. L. (2012). Using quizzes to enhance summative-assessment performance in a web-based class: An experimental study. *Journal of Applied Research in Memory and Cognition*, *1*(1), 18–26. https://doi.org/10.1016/j.jarmac.2011.10.001

McDermott, K. B., Agarwal, P. K., D’Antonio, L., Roediger, H. L., & McDaniel, M. A. (2014). Both multiple-choice and short-answer quizzes enhance later exam performance in middle and high school classes. *Journal of Experimental Psychology: Applied, 20(1)*, 3–21. https://doi.org/10.1037/xap0000004

Neely, J. H. (1977). Semantic priming and retrieval from lexical memory: Roles of inhibitionless spreading activation and limited-capacity attention. *Journal of Experimental Psychology: General, 106*(3), 226–254. https://doi.org/10.1037/0096-3445.106.3.226

Nungester, R. J., & Duchastel, P. C. (1982). Testing versus review: Effects on retention. *Journal of Educational Psychology, 74(1)*, 18–22. https://doi.org/10.1037/0022-0663.74.1.18

Ozuru, Y., Briner, S., Kurby, C. A., & McNamara, D. S. (2013). Comparing comprehension measured by multiple-choice and open-ended questions. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale, 67(3)*, 215–227. https://doi.org/10.1037/a0032918

Payne, B. K. (2001). Prejudice and perception: The role of automatic and controlled processes in misperceiving a weapon. *Journal of Personality and Social Psychology, 81(2)*, 181–192. https://doi.org/10.1037/0022-3514.81.2.181

Payne, B. K. (2008). What mistakes disclose: A process dissociation approach to automatic and controlled processes in social psychology. *Social and Personality Psychology Compass, 2(2)*, 1073–1092. https://doi.org/10.1111/j.1751-9004.2008.00091.x

Potts, R., Davies, G., & Shanks, D. R. (2019). The benefit of generating errors during learning: What is the locus of the effect? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 45(6)*, 1023–1041. https://doi.org/10.1037/xlm0000637

Roediger, H. L., Agarwal, P. K., Kang, S. H. K., & Marsh, E. J. (2010). Benefits of testing memory: Best practices and boundary conditions. In G. M. Davies & D. B. Wright (Eds.), *New frontiers in applied memory* (pp. 13– 49). Brighton, UK: Psychology Press.

Roediger, H. L., & Marsh, E. J. (2005). The positive and negative consequences of multiple-choice testing. *Journal of Experimental Psychology:* *Learning, Memory, and Cognition, 31(5)*, 1155–1159. https://doi: 10.1037/0278-7393.31.5.1155

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/a003755

Sparck, E. M., Bjork, E. L., & Bjork, R. A. (2016). On the learning benefits of confidence-weighted testing. *Cognitive Research: Principles and Implications, 1(1).* https://doi.org/10.1186/s41235-016-0003-x

Stolz, J. A., & Merikle, P. M. (2000). Conscious and unconscious influences of memory: Temporal dynamics. *Memory*, *8*(5), 333–343. https://doi.org/10.1080/09658210050117753

Toth, J. P., Reingold, E. M., & Jacoby, L. L. (1994). Toward a redefinition of implicit memory: process dissociations following elaborative processing and self-generation. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *20*(2), 290–303. https://doi.org/10.1037//0278-7393.20.2.290

Yang, C., Luo, L., Vadillo, M. A., Yu, R., & Shanks, D. R. (2021). Testing (quizzing) boosts classroom learning: A systematic and meta-analytic review. *Psychological Bulletin*, *147*(4), 399–435. https://doi.org/10.1037/bul0000309

**Table 1**

*Mean (SD) Proportion of Test-2 Answer Types to Related Questions Conditioned on Being Answered Incorrectly in The First Test in Experiment 1*

|  |  |  |
| --- | --- | --- |
| Answer type on the final test | MC | CR |
| Correct |  |  |
| Previous answer | 0.19 (.20) | 0.14 (.20) |
| Other | 0.25 (.23) | 0.19 (.22) |
| Incorrect |  |  |
| Previous answer | 0.10 (.17) | 0.05 (.15) |
| Corrective feedback | 0.30 (.32) | 0.15 (.23) |
| Other | 0.15 (.19) | 0.46 (.45) |

*Note.* MC and CR represent multiple-choice and cued recall respectively.

**Table 2**

*Mean (SD) Proportion of Test-2 Answer Types Conditioned on Being Answered Incorrectly on The First Test in Experiment 2*

|  |  |  |
| --- | --- | --- |
| Answer type on the final test | MC | CR |
| Correct |  |  |
| Previous answer | 0.12 (.14) | 0.16 (.20) |
| Other | 0.16 (.21) | 0.21 (.26) |
| Incorrect |  |  |
| Previous answer | 0.11 (.17) | 0.03 (.11) |
| Corrective feedback | 0.50 (.45) | 0.15 (.19) |
| Other | 0.11 (.17) | 0.46 (.41) |

*Note.* MC and CR represent multiple-choice and cued recall respectively.

**Table 3**

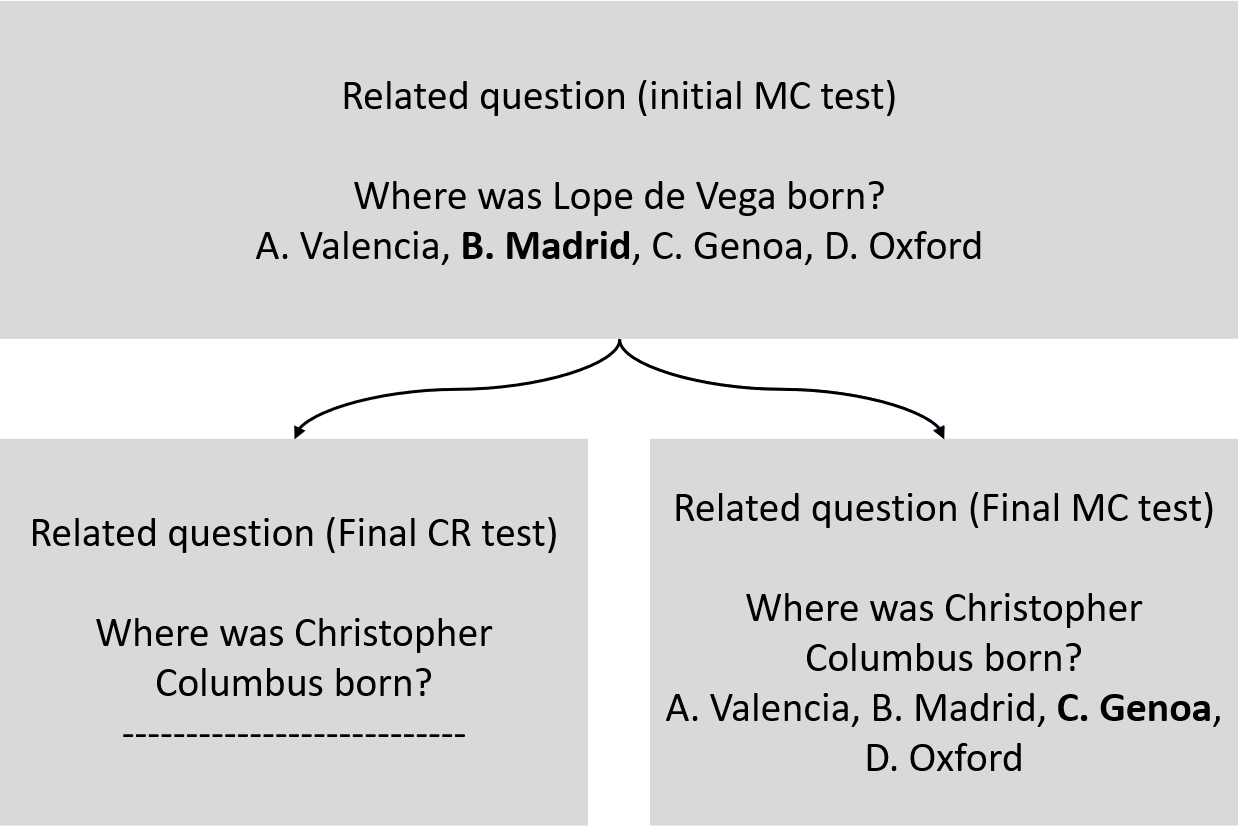
*Mean (SD) Proportion of Test-2 Answer Types Conditioned on Being Answered Incorrectly in The First Test in Experiment 3.*

|  |  |  |
| --- | --- | --- |
| Answer type in the final test | MC | CR |
| Correct |  |  |
| Previous answer | 0.18 (.20) | 0.21 (.31) |
| Other | 0.15 (.24) | 0.19 (.24) |
| Incorrect |  |  |
| Previous answer | 0.11 (.21) | 0.04 (.10) |
| Corrective feedback | 0.43 (.48) | 0.14 (.24) |
| Other | 0.14 (.21) | 0.42 (.42) |

*Note.* MC and CR represent multiple-choice and cued recall respectively.

**Figure 1**

*Example of Related Items Presented in Two Final Test Formats.*



*Note*. Correct answers in boldface. CR = cued recall; MC = multiple choice

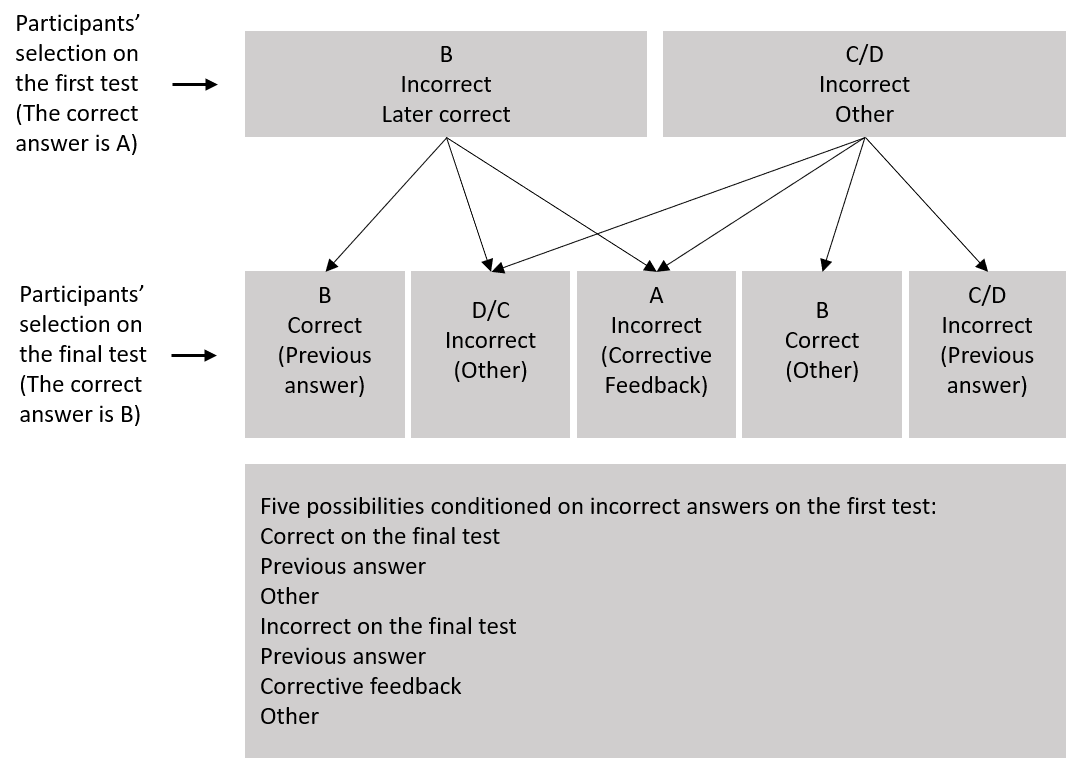
**Figure 2**

*Mean Recall Accuracy for Each Question Type Broken Down by Final Test Type in Experiment 1.*

*Note.* MC and CR represent multiple-choice and cued recall respectively. Bars indicate standard errors of the mean.

**Figure 3**

*Schematic Illustrating the Five Final-Test Answer Types (Related Items)*



**Figure 4**

*Mean Proportion of Endorsing the Corrective Feedback for Each Question Type Broken Down by Final Test Type in Experiment 1.*

*Note.* MC and CR represent multiple-choice and cued recall respectively. Error bars indicate standard errors of the mean.

**Figure 5**

*Mean Proportion Correct for Each Question Type Broken Down by Final Test Type in Experiment 2.*

*Note.* MC and CR represent multiple-choice and cued recall respectively. Error bars indicate standard errors of the mean.

**Figure 6**

*Mean Proportion of Endorsing the Corrective Feedback for Each Question Type Broken Down by Final Test Type in Experiment 2.*

*Note.* MC and CR represent multiple-choice and cued recall, respectively. Error bars indicate standard errors of the mean.

**Figure 7**

*Mean Proportion Correct for Each Question Type Broken Down by Final Test Type in Experiment 3.*

*Note.* MC and CR represent multiple-choice and cued recall respectively. Error bars indicate standard errors of the mean.

**Figure 8**

*Mean Proportion of Endorsing the Corrective Feedback for Each Question Type Broken Down by Final Test Type in Experiment 3.*

*Note.* MC and CR represent multiple-choice and cued recall respectively. Error bars indicate standard errors of the mean.

1. “Lures” is another term for incorrect alternatives. [↑](#footnote-ref-2)
2. Obviously, given the results of the previous analysis and the CR data in Figure 4, participants did intrude with the corrective feedback more often with related items than new items. However, compared to the MC group, their tendency to endorse the corrective feedback was less, at least when the practice test question was answered incorrectly (Table 1). [↑](#footnote-ref-3)
3. This interaction needs to be treated with caution given that feedback intrusions for new questions in the CR group were near floor. [↑](#footnote-ref-4)
4. Little et al. (2019) did not report eliminating any participants in their study, although we employed the same design on the first test and used the same platform (MTurk). [↑](#footnote-ref-5)
5. This interaction needs to be treated with caution given that feedback intrusions for new questions in the CR group were near floor. [↑](#footnote-ref-6)