

**Multiple-Choice Testing: Controlled and Automatic Influences of Retrieval Practice in
an Educational Context**

Aeshah Alamri¹ and Philip A. Higham²


¹ Department of Psychology, King Saud University

² School of Psychology, University of Southampton

IN PRESS

QUARTERLY JOURNAL OF EXPERIMENTAL PSYCHOLOGY

Author Notes

Aeshah Alamri  <https://orcid.org/0000-0001-5212-7909>

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

We have no known conflict of interest to disclose.

Supplemental materials are available at:

https://osf.io/ry6ug/?view_only=3a7f962636bf4b6c96e4d2537f89e719

Correspondence concerning this article should be addressed to Aeshah Alamri,
Department of Psychology, King Saud University, Riyadh, Kingdom of Saudi Arabia. Email:
aealamri@ksu.edu.sa

Abstract

Previous studies have shown that taking an initial multiple-choice (MC) test produced both automatic influences (i.e., those that operate quickly, without effort, and requiring few attentional resources) and controlled influences (influences that are slower, applied more deliberately, sometimes oppose automatic processes, and require more attentional resources) on performance in a subsequent test. In this study, we examined the involvement of automatic and controlled processes on performance with MC questions that are related to earlier practice questions, but which have different correct answers. In Experiment 1, which was conducted online with MTurk, automatic influences tended to dominate responding despite using educational materials (SAT questions). Including repeated items in the final test (Experiment 1, 4) and increasing the time lags between questions (Experiment 2, 4) increased the automatic influence. However, in a genuine educational environment (university classroom), controlled influences tended to dominate responding instead, similar to what has been observed with cued recall (CR) final tests, but only when there are no repeated items. These controlled influences were enhanced by presenting the related questions back-to-back in the testing sequence (Experiment 2) but were unaffected by feedback on the initial test (Experiment 3). We conclude that performance on both MC and CR tests are affected by both automatic and controlled influences of retrieval practice, but that one type of influence will override the other depending on the presence of repeated items, the specific testing format, and examinees' investment in scoring well.

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

Multiple-Choice Testing: Controlled and Automatic Influence of Retrieval Practice in an Educational Context

Tests have been utilized as the main method to assess students' performance for centuries. Many educational institutions around the world use testing as a reliable and feasible way to evaluate what students have learned after engaging in a learning experience. Although there are a variety of testing formats that can serve different purposes, multiple-choice (MC) questions can be considered the most common testing format used worldwide. For example, almost all students in the United States encounter MC questions during their school years or when taking standardized tests such as the SAT and the GRE¹ (Rauschert et al., 2019). This is due to the advantages of using this test format with large groups of students, including the ease of marking.

Positive and Negative Testing Effects

The prior literature has demonstrated that MC testing is useful not only to assess learners' performance but also to enhance it. For example, taking an MC practice test can enhance later retention on a final test taken some time later, an enhancement known as the *testing effect* (see Yang et al., 2021 for a meta-analysis). For instance, McDermott et al. (2014) gave high-school students two initial quizzes that contained MC and cued-recall (CR) questions which were presented after the material was taught, but before a final test. The final test that followed included both MC and CR questions that did not necessarily match the format of the questions in the initial tests. The findings indicated an enhancement in students' performance on the tested items compared to the control items, regardless of the initial testing format. This research is consistent with other studies that have shown the usefulness of taking

¹ The SAT and GRE (Graduate Record Examination) are standardized tests that are used traditionally in the United States for university admission. The SAT is taken by high school students for admission into undergraduate programs, while the GRE is taken by undergraduate students for admission into postgraduate programs.

a practice MC test on later performance for students at different educational levels, including college students (McDaniel et al., 2012), middle school students (McDaniel et al., 2013), and elementary school students (Marsh et al., 2012).

Despite the positive effects of MC testing, some researchers (e.g., Roediger & Marsh, 2005) have reported a *negative testing effect* of taking an MC practice test. The negative testing effect occurs when lures that learners are exposed to while completing a practice MC test are encoded in memory and intrude on later CR versions of those questions. The problem is exacerbated if there is no corrective feedback during practice or if practice questions are used that have many lures. For example, Marsh et al. (2009) examined the positive and negative effects of taking an initial MC test using educational material. Participants answered initial MC questions without feedback that were acquired from SAT subjects tests before taking the final test. The findings from the final CR test demonstrated both positive and negative effects of taking an initial MC test. For the positive effect, participants performed better on the tested items compared to the untested items. For the negative effect, however, taking an initial MC test resulted in lure intrusions on the final CR, particularly if participants did not receive corrective feedback during the initial test (cf. Butler & Roediger, 2008).

Positive Effects of MC Testing on Related Items

The positive effect of MC testing extends beyond the tested items; it is found even with certain types of untested but related items. For example, Little, Bjork, and colleagues (e.g., Little et al., 2019) have investigated pairs of questions for which a lure from an MC question presented on the initial MC test is the correct answer to a second, related CR question on the final test. For example, a related pair of questions might be, “*What is the capital of Norway? A. Helsinki, B. Leningrad, C. **Oslo**, D. Stockholm*” presented on a practice test (boldface indicates the correct answer), followed by, “*What is the capital of Finland?*”

(Helsinki) on a final CR test. Note that both questions are related in that they both query a similar topic (capital cities of Scandinavian countries). However, the correct answers are different; specifically, a lure from the practice test (*option A. Helsinki*) is the correct answer to the related question on the second test.

In several studies, Little, Bjork, and colleagues found that taking a practice MC test enhanced later performance on related items in the final CR test. For instance, Little and Bjork (2012; see also, 2015; Little et al., 2012; 2019) had participants read some expository text and then take an initial MC or CR test on the text without providing corrective feedback. All the MC questions presented during the initial test were constructed with competitive lures (i.e., plausible answers for the question). For example, one question with competitive lures was “*The body of Saturn is primarily composed of what element?? a. Oxygen, b. Hydrogen, c. Helium, d. Carbon*”. After either a five min or a 48 hr retention interval, participants took a final CR test which contained some questions repeated from the first test, related but untested questions, and new questions as a control. The results showed that, compared to new items, taking an initial MC test enhanced both repeated- and related-item performance on the final CR test at both retention intervals. Moreover, for repeated items, the enhancement following a MC practice test was greater than that observed following a CR practice test. For related items, there was only enhancement if participants wrote a MC practice test, not if they wrote a CR practice test. The researchers concluded that MC practice testing is not necessarily less effective than practice tests such as CR that require more effortful retrieval. As long as the MC practice questions have competitive lures, learners engage in reasoning and retrieval processes to reject incorrect alternatives, and the products of those processes can be used later to retrieve correct answers to related questions on the final CR test. This explanation is bolstered by other studies demonstrating the importance of increasing the depth of retrieval on the initial MC testing either by using competitive lures (e.g., Little & Bjork, 2015) or by

utilizing MC practice test formats that encourage intensive retrieval (e.g., Alamri & Higham, 2022; Little et al., 2019; Sparck et al., 2016).

Negative Effects of MC Testing on Related Items

Although the literature showing the positive effect of initial MC questions on related items is extensive, these studies used CR questions in the final test. In contrast, Higham et al. (2016) examined whether taking a practice MC test also facilitated performance on related items when the final test format encouraged familiarity-based responding. In their studies, participants first read an expository text and then took an initial MC test with feedback being provided after each question. As in Little and Bjork's (2012) study, they then took a final test comprised of repeated, related but untested, and new questions. However, their final test was in MC format rather than CR format.

To create related MC questions, Higham et al. (2016) added alternatives to the final test questions that were previously in CR format ensuring one of the lures was a correct answer on the initial test. In contrast to the facilitated performance on related items observed in previous studies, Higham et al. observed *impaired* performance regardless of the number of options that matched between the related MC questions. That is, the likelihood of correctly answering the related MC questions on the final test was *less* than for new MC items.

Higham et al. (2016) conditioned participants' final test responses on their practice test responses to determine the cause of the impairment on related items. They found that participants overwhelmingly selected the corrective feedback from the first test as the correct answer for the second related question on the final test, even though it was no longer correct. The impairment on the related items persisted even when the repeated items were dropped from the final test, as they might have misled participants to consider the related questions as

repeated as well. These results contrast with those from many prior studies that showed facilitation on the related items in final CR testing (e.g., Little & Bjork, 2012, 2015).

To further explore the contrasting final test format results with related items, Alamri and Higham (2022) directly compared MC and CR final-test performance in three experiments. Across the three experiments, the depth of retrieval was increased by using different MC formats on the first test (Sparck et al., 2016). In Experiment 1, the researchers used standard MC format (i.e., select a single answer). In Experiment 2, participants were asked to rank the alternatives from most to least favourite. In Experiment 3, participants were required to provide reasons to reject alternatives (i.e., elimination testing; Little et al., 2019). They found that utilising a practice test format that elicited a low level of retrieval (i.e., standard MC) resulted in no MC impairment or CR facilitation. However, when retrieval depth was increased in Experiments 2 and 3, task dissociations were observed. Specifically, MC performance on the final test was impaired (related < new), replicating Higham et al.'s (2016) results, whereas CR final test performance was enhanced (related > new), replicating Little and colleagues (e.g., Little et al., 2012, 2019). As in Higham et al.'s (2016) research, the problem with the MC final test was that participants tended to erroneously select the option that was the corrective feedback on the initial test which was no longer correct. Although this tendency to endorse the prior corrective feedback existed on the CR final test as well, it was overshadowed by an opposing tendency to benefit from the earlier retrieval practice, resulting in net facilitation.

Theoretical Mechanisms

Alamri and Higham (2022) argued that their results were consistent with a dual-process model incorporating controlled and automatic influences.² On the one hand, answering difficult MC practice questions with a test format that encourages lure processing elicits reasoning and retrieval processes that can be used in a controlled manner to facilitate performance on later tests. This account is similar to the one forwarded by Little and colleagues (e.g., Little et al., 2012, 2019; Sparck et al., 2016). On the other hand, answering MC questions and receiving feedback creates episodes in memory that generate automatic feelings of familiarity when related questions are encountered on later tests. If the later test is also MC and the options are matched, the options act as a retrieval cue for the earlier episode enhancing automatic familiarity influences. The result is that participants falsely recognize the related final test question, believing that it has been repeated from the practice test. Consequently, participants respond with the corrective feedback, which is no longer the correct answer, causing performance to be impaired relative to new questions.

This dual-process account of the effects of retrieval practice was bolstered by Alamri and Higham's (2024) follow-up study. One potential criticism of the dual-process account is that the automatic influence was not actually out of participants' control but was in fact a strategic effect. That is, when participants encounter difficult final-test questions and they are unsure of the correct answer, they may reason that the option most likely to be correct is the

² Following Alamri and Higham (2022), we adopted the terms "controlled" and "automatic" to distinguish between the different types of influence that occur in this paradigm. On the one hand, automatic influences are unintentional, not under voluntary control, and effortless (e.g., choosing the most familiar MC option without careful analysis). In our paradigm, such influences sometimes seduce participants into making incorrect responses, suggesting that participants are not controlling them. On the other hand, controlled influences are more intentional, under voluntary control, and more effortful (e.g., deliberately applying the products of an earlier retrieval episode). In our paradigm, they allow participants to counter the attractiveness of a familiar lure and typically lead to a correct response. We believe this terminology is useful, but we also understand that it carries theoretical baggage. For example, automatic influences are sometimes described as "unconscious" or "implicit" influences (e.g., Jacoby et al., 1989), but we do not believe our data speak to the conscious/unconscious debate. Other dichotomies might also be suitable such as System 1 versus System 2 thinking (Kahneman, 2011) or analytic versus nonanalytic cognition (Jacoby & Brooks, 1984). However, we reserve the automatic/controlled terminology in the current research both for consistency with our prior research (Alamri & Higham, 2022) and because, for the most part, we believe it best captures the distinction we are trying to make.

correct answer to the earlier related question. They may be fully aware that the questions are different and may well have different answers. However, as a best guess, they deliberately choose the option that was correct earlier. To address this alternative account, Alamri and Higham used opposition instructions (Jacoby et al., 1989). Specifically, participants were told that there were questions on the final test that were similar to ones on the practice test. They were also told that should be very careful with these questions because the correct answer to these similar final-test questions was *never the same* as the correct answer to the earlier question. Thus, the controlled, strategic influence and the automatic, false-recognition influence were set in opposition. Any tendency on the part of participants to respond with the same option to the related questions strategically and deliberately to boost accuracy would be undermined by these instructions. Conversely, if the influence was automatic and thereby out of participants' control, then the opposition instructions should make little difference.

In addition to the opposition instructions, Alamri and Higham (2024) also required participants to indicate whether test questions were *old* (answered earlier on the practice test) or *new* (not answered earlier). Overall, the results supported the dual-process account; related-item performance on the MC final test was impaired relative to control items, and the opposition instructions made very little difference to the size of the impairment. Moreover, the impairment was mostly attributable to cases where related questions were falsely recognized (i.e., falsely called "old" for the recognition question).

The automatic, false-recognition account was also given support from the results of another experiment from the same study in which the presence of feedback during the practice test was manipulated. Alamri and Higham (2024) showed in this experiment that if responses were conditioned on an incorrect response on the practice test, removing the feedback during the practice test shifted participants' tendency to respond with corrective feedback to responding with their previous answer or sometimes with an *other* answer. For

some participants, in the absence of feedback, there would be no reason to change their earlier favourite option to something different if participants falsely believed that the related question was repeated. Hence, the previous response was an attractive option when no feedback was provided but was largely avoided if corrective feedback was provided (and their previous answer was wrong).

Current Study

The literature to date on retrieval practice effects with related items suggests that opposing results are obtained depending on the format of the final test. That is, facilitation is typically observed with a CR final test whereas impairment is observed if the final test is MC. However, the story may not be that simple. As Alamri and Higham (2024) noted, this division makes a *process purity* assumption (e.g., Jacoby, 1991) whereby CR versus MC tasks measure solely controlled versus automatic processes, respectively. However, a deeper analysis of Alamri and Higham's (2022) results undermined this interpretation. For example, even in cases where facilitation was found with a CR task, participants still demonstrated a greater tendency to endorse the option that was the corrective feedback for the related questions than to produce that same option when the questions were new.³ Thus, there was evidence for an automatic influence on the CR final test, although it was much smaller than with an MC final test.

Although both controlled and automatic effects have been identified with CR final tests, to date there is no evidence for a controlled influence with MC final tests. Alamri and Higham (2022) reasoned that as the practice test format encouraged deeper retrieval (e.g., as with elimination testing), controlled processes might start to overshadow the automatic influences. However, no evidence of such overshadowing was found; even with elimination

³ Because items were counterbalanced between related and new items, this comparison was possible.

testing, the practice test format most likely to elicit deep retrieval during practice, impairment with related items (vs. new items) was still observed due to the corrective feedback being selected at a high rate.

The central aim of the current research is to investigate whether there are conditions under which controlled influences might be found with MC final tests. This question is important to answer for both theoretical and practical reasons. First, finding evidence of controlled processing in MC tests would bolster Alamri and Higham's (2022, 2024) dual-process account of the influences of practice testing. Specifically, it would demonstrate that both controlled and automatic influences occur with both CR and MC final tests, but to different degrees (i.e., the process purity assumption does not apply). From a practical perspective, identifying the situations under which controlled processes can be made more dominant with MC final tests would allow us to make recommendations to educators who may be using such tests. Obviously, educators want their students' performance to benefit from retrieval practice, not have it undermined. Ideally, we would like to find a scenario that leads to facilitation on the MC final tests rather than impairment, the latter being the case in all experiments investigating this issue with an MC final test to date.

Our rationale was that one reason that automatic processes might dominate MC final-test performance is that the tests were low stakes. For example, following Little et al. (2019), Alamri and Higham (2022) used general-knowledge questions. With such "trivia" materials, participants may not engage in effortful processing unless the final-test format demands it, as with a CR final test. Perhaps with materials that are more education focused, participants will engage in deeper processing of the questions on the final MC final test, allowing controlled influences of retrieval practice to override automatic influences. To test this possibility, we conducted four experiments where participants answered two sets of MC questions, one designated as an *initial* or *practice* test and the other designated as a *final* test. In all

experiments, we attempted to raise the stakes by using educationally relevant materials (details later) and manipulating variables that we thought would moderate the relative influence of automatic and controlled processes. The tests in Experiments 2, 3, and 4 were also administered in a real university classroom to raise the stakes further.

Experiment 1 was conducted online, and the materials were MC SAT exam questions. We hypothesized that SAT questions, which are part of a high-stakes standardized exam taken by high-school seniors as an entrance requirement to university, might encourage more effortful processing and potentially reveal controlled influences on the final test. We also manipulated two variables in Experiment 1 in an attempt to reveal controlled influences. First, we manipulated the presence of repeated items. We reasoned that repeated items may encourage participants to be less cautious and to rely more heavily on automatic influences during the test because repeated items provided a reason for the familiarity of the corrective feedback, the options, and other aspects of the related questions. For repeated items, reliance on automatic influences would lead to the correct response, but it would produce errors for related items. In contrast, the absence of repeated items might make participants more cautious about feelings of familiarity such that they attempt to counter familiarity with controlled processes (e.g., read the question more carefully), allowing controlled influences to be expressed. Second, we manipulated whether the questions appeared to participants as a single test or as two tests. The reasoning here was that two discrete tests might encourage participants to “look back” and search for repeated items, potentially enhancing the role of automatic influences and false recognition. On the other hand, with all questions appearing as a single test, participants might be more inclined to focus on each question individually without trying to retrieve an earlier encounter with that question. In other words, single-test participants might focus on “solving” the question rather than “remembering” answers given

to previous questions (Jacoby, 1978), thereby limiting the effect of automatic influences and false recognition on performance.

The aim of Experiments 2, 3, and 4 was to investigate controlled and automatic influences in a genuine educational context where the stakes are higher than in an online environment. Students in this context may be more interested in learning and enhancing their knowledge compared to online participants, thereby enhancing controlled influences. As in Experiment 1, we also manipulated a variable that we hypothesized would moderate the relative influence of automatic and controlled processes. Specifically, we varied the lag between the related questions. We reasoned that, compared to longer lags, if the related questions were presented back-to-back (lag 0), students may tend to notice that the questions are related, but not the same, and required different answers. Therefore, automatic influences could potentially be overridden by controlled influences.

In Experiment 3, we revisited the role of corrective feedback. Alamri and Higham's (2024) results showed that, compared to the no-feedback condition, final MC test performance on related items was worsened if corrective feedback was provided on the initial test. However, those results were obtained in a context where automatic influences were dominating responding. If presenting the questions in a genuine educational environment causes controlled influences to dominate instead, then the provision of corrective feedback might have a very different effect. For example, Little et al. (2012) found that feedback on the initial test made little difference when the final test was CR and controlled processes were dominating.

In Experiment 4, we partially replicated the conditions in Experiment 2 but with one critical exception: repeated items were included in the final test. If controlled influences dominate performance in a genuine educational environment, then including such items

should have no effect on related items. However, including repeated items could increase participants' tendency to search the test list for repeated items (i.e., there might be a shift from a "solving" strategy to a "remembering" strategy; Jacoby, 1978). Such a search strategy would make participants more vulnerable to automatic familiarity influences when related questions are encountered. The result might be greater false recognition of related items resulting in impaired performance. Table 1 summarises the experimental designs used in each of the four experiments.

Overall, we expected to see more controlled influences in the current experimental series than we have observed previously with MC final tests, which could potentially lead to facilitated performance on related versus new questions. However, in line with previous studies (e.g., Alamri & Higham, 2022), we also expected that automatic influences would be present as well, just to a lesser degree than in our previous research.

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

Experiment 1

In Experiment 1, participants answered SAT MC questions in the first test. After a distractor task, they took a final MC test that included repeated, related but untested, and new questions. As noted earlier, we compared participants' performance in a one-test condition with their performance in a two-test condition. It was important to include a distractor task in both conditions to ensure that the retention interval between the first and final test was held constant. However, the inclusion of a distractor task in the one-test condition might give the appearance of two discrete tests. Consequently, we replaced the distractor task in the one-test condition with filler questions similar to the other questions on the test, making it seem like a longer single test.

In addition to one test/two test manipulation, we also manipulated the presence of repeated items. Higham et al. (2016) found that regardless of including or eliminating the repeated items from the final test, taking a practice MC test impaired performance on the final MC test. However, it is not clear how a combination of utilizing educational material, presenting the questions in a single test, and dropping repeated items would affect the performance on related items. We expected that if we were to observe facilitation with related items (vs. new), it would most likely occur in the single-test, no-repetition condition, where automatic influences would be limited (due to the focus of attention being on the present and less confusion between item types) and controlled influences would be great (due to the serious educational nature of the questions).

Method

Participants

Across Alamri and Higham's (2022) experiments, the lowest effect size for impaired performance with related MC items (i.e., related < new on the final test) was Cohen's $f = .24$. Based on this result, we conducted an a priori power analysis with 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 initially tested 145 participants. However, seven participants were excluded after reviewing their performance and responses to attention-check questions. For example, some participants did not rank the alternatives as instructed (see below). The final analysis involved the remaining 138 participants (female = 74), with ages ranging between 22 and 60 years ($M = 37.59$, $SD = 9.57$) 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 comprised four groups with 35 participants in the two-tests/repetition group, 34 participants in the one-test/repetition group, 35 participants in the two-tests/no-

repetition questions group, and 34 participants in the one-test/no-repetition group.

Participants in each group were randomly assigned to three counterbalancing formats with 10-13 participants each as explained later.

Design

The experiment employed a 2 x 2 x 2 mixed factorial design with test type (two tests, one test) as well as presence of repeated questions on the final test (repetition, no repetition) manipulated between subjects, and question type on the final test (new, related) manipulated within subjects.⁴ The main dependent variable was the participants' mean performance on the final MC test.

Twenty-two questions were presented on the initial test and 33 questions on the final test. To ensure that each question served in each experimental condition equally often, we created three surveys on Qualtrics, the software used to present the questions. In the repetition group, the three surveys rotated the questions through the repeated, related, and new conditions across participants to eliminate item effects. Specifically, we divided the 33 question pairs into three sets (1-11 A/B, 12-22 A/B, & 23-33 A/B, where A refers to the first question of each pair and B to the second). The final test always consisted only of the 33 B items from the 33 pairs.

For the first test in the repeated group, the first counterbalance condition comprised items 1-11 B repeated questions (the first presentation of the repeated questions) and items 12-22 A related questions (the first questions of the related pairs). Items 23-33 B acted as new items that were on the final test only. The second counterbalance version comprised items 12-22 B repeated questions (the first presentation of the repeated questions) and items 23-33 A

⁴ Although repeated items were included on the final test for the repetition groups, these items were not included in the core 2 X 2 X 2 design because there were no repeated items in the no-repetition groups.

related questions (the first questions of the related pairs). Items 1-11 B acted as new items that were on the final test only. Finally, the third counterbalance condition comprised items 23-33 B repeated questions (the first presentation of the repeated questions) and items 1-11 A related questions (the first questions of the related pairs). Items 12-22 B acted as new items that were on the final test only. The same counterbalancing procedure was used in the no-repetition group except 11 filler items replaced the first presentation of the 11 repeated items in each first test. Questions were presented in a random order for each participant, but the MC alternatives were always presented in the same order. Figure 1 summarizes the design.

Materials and Procedure

Sixty-six questions on SAT subjects test were obtained from (CrackSAT, n.d.) and amended to suit the purpose of our study. For example, all the questions were formed into pairs based on the topic (i.e., 33 pairs) which were related conceptually (e.g., both questions in a given pair were about United States' presidents) and each pair shared the same alternatives. Also, we reduced the number of alternatives for each question from five, which is the usual number of alternatives for SAT questions, to four alternatives only by removing one of the lures. This was done to keep the number of alternatives consistent with previous studies that have investigated MC related items (e.g., Alamri & Higham, 2022, 2024). The questions covered a wide range of the topics included in the SAT subjects test (physics, chemistry, US history, biology, world history). Additionally, 11 more (filler) questions were prepared to replace the repeated questions in the no-repetition condition on the final test. The addition of the filler items was done to ensure that all the groups had the same number of questions on the final test (i.e., 33 questions). Those questions were similar to the original repeated questions in terms of the topics covered. All questions are available as Supplementary Materials.

The experiment was conducted online using Qualtrics survey software, and four groups of participants were instructed on the procedure for their group before taking the survey. For the two-tests conditions, participants were instructed to take the first test, answer a few math questions (i.e., distractor task), and then take the final test, whereas participants in the one test condition were instructed to take one single test. After reading the instructions, participants answered two attention-check questions about the instructions to ensure that they read and understood them. For all the groups, the experiment started with taking a 22-item initial MC test. To ensure a reasonable depth of retrieval, the initial test was presented in a ranking-options format similar to that used in Alamri and Higham (2022, Experiment 2). 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 1 as the answer most likely to be correct and 4 as the least likely answer. Also, to discourage participants from completing the test (including the final test) too quickly, participants were not permitted to advance to the next question until 15 s had elapsed. By clicking “Next,” participants received the corrective feedback. The feedback was provided regardless of whether the chosen option was correct (e.g., “*The correct answer is Gamma rays*”). Clicking “Next” again, advanced to the next question. The presentation order of the questions in the initial test was randomized.

After answering the initial 22 questions, the procedure differed based on the experimental condition. For the two-tests condition, the first 22 questions were demarcated as a complete initial test, and so participants engaged in a distractor task which involved answering basic mathematics questions before they were instructed to start a second, final test. For the one-test condition, the initial 22 questions were treated as if they were the first part of a longer single test. To ensure that the retention interval between the initial and final test was the same across the one- and two-tests groups, the mathematics questions that served

as a distracter task in the two-tests groups were replaced with filler questions. These questions were different from the 66 SAT questions that were used in the initial and final tests but covered similar topics and presented in exactly the same format as the SAT questions. Thus, although the task was divided into an initial test, filler questions, and a final test for experimental purposes, the task for the single-test groups seemed like one single test from the participants' point of view.

Following the filler task/questions, each group took the final MC test which comprised 33 questions. For participants in the repetition groups, the questions were divided into three categories presented in a random order: (a) 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, (b) 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), and (c) 11 new questions – which acted as the control questions – that were untested and unrelated to the questions from the initial test. For the no-repetition groups, the repeated questions were replaced with 11 filler questions to equate the groups in terms of the final test length. The repeated questions would have been easy for participants in the repetition groups. Therefore, to balance the overall difficulty level of the final test for the repetition and no-repetition groups, we chose easy questions to use as replacement filler questions in the no-repetition groups. Also, to make all the four groups consistent in terms of the final-test format, and to ensure that the initial and final tests were not demarcated for participants in the one-test groups, the questions on the final MC test were presented in the same format (ranking) and for the same duration (15 s) as for the initial test.

No feedback was provided for the final-test questions for any group. To ensure that participants in the one-test condition would continue to consider all questions as belonging to a single test, they were told that we were testing the benefits of immediate versus delayed

feedback. Thus, part way through the test, immediate feedback would no longer be presented and, instead, the feedback would be presented at the end of the test. Therefore, the feedback was provided immediately after each question for the first 22 questions (i.e., initial test) for all four groups and for the filler (i.e., distractor) questions for the one-test groups. However, feedback was delayed to the end of the final test for the remaining 33 questions (i.e., final test). The experiment took each participant approximately 20-25 minutes to complete.

For the scoring of the initial and final tests in this and subsequent experiments, the answers were marked based on how the correct answer was ranked: a correct answer ranked first, second, third, and fourth received 3, 2, 1, and 0 marks, respectively. These scores were then converted into *accuracy*, our primary dependent variable, which was the proportion of points scored out of three possible points per question (e.g., a correct answer ranked second would produce an accuracy score of $2/3 = .67$ for that question).

Results

Initial Test Performance

Mean accuracy on the first test in the two-tests/repetition, one-test/repetition, two-tests/no-repetition, and one-test/no-repetition groups was .64 ($SD = .11$), .66 ($SD = .13$), .64 ($SD = .12$), and .64 ($SD = .11$), respectively. A 2 (test type: two tests, one test) x 2 (repetition type: repetition, no repetition) between-subjects Analysis of Variance (ANOVA) indicated no significant differences, all $F_s < 1$, suggesting that the four groups were comparable.

Final Test Performance

Final-Test Accuracy. Although there were repeated questions and filler questions depending on group, we focused our analyses on the related and new questions.⁵ We conducted a 2 (test type: two tests, one test) x 2 (repetition type: repetition, no repetition) x 2 (question type: new, related) mixed-factor ANOVA with final test accuracy as the dependent variable (see Figure 2). The main effect of test type was not significant, $F(1, 134) = 1.39, p = .24, \eta_p^2 = .01$; accuracy was comparable across the two-tests ($M = .64, SD = .12$) and one-test ($M = .62, SD = .11$) groups. However, there was a significant main effect of repetition type, $F(1, 134) = 6.63, p = .01, \eta_p^2 = .05$; accuracy was higher on the no-repetition condition ($M = .65, SD = .12$) compared to the repetition condition ($M = .60, SD = .09$). Also, there was a significant main effect of question type, $F(1, 134) = 19.70, p < .001, \eta_p^2 = .13$, as accuracy was higher on the new items ($M = .66, SD = .13$) compared to the related items ($M = .60, SD = .15$). These two main effects were qualified by a significant interaction between repetition type and question type, $F(1, 134) = 4.67, p = .032, \eta_p^2 = .03$. A paired-sample *t*-test revealed that participants in the repetition condition performed better on the new items ($M = .65, SD = .11$) compared to the related items ($M = .56, SD = .13$), $t(68) = -4.94, p < .001, d = .74$. However, for the no-repetition condition, the difference between performance on the new items ($M = .67, SD = .14$), and the related items ($M = .64, SD = .15$) was not significant, $t(68) = -1.53, p = .13, d = .21$. No other interaction was significant, largest $F(1, 134) = 1.69, p = .19, \eta_p^2 = .01$.

Final-Test Answer Types. Previous studies (e.g., Higham et al., 2016; Alamri & Higham, 2022, 2024) 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 member of the pair on the final MC test. To examine whether a similar pattern was

⁵ Although not included in the core analysis, we calculated mean accuracy for the repeated and filler items. In all cases, accuracy was high as expected (one-test repeated: $M = .91, SD = .11$; two-tests repeated: $M = .94, SD = .08$; filler: $M = .87, SD = .12$).

observed here, we analysed the probability of different answer types conditioned on answering incorrectly on the first test (see Figure 3 for an illustration). This analysis produced five mutually exclusive final-test possibilities: (a) a correct answer on the final test that matched the previous incorrect answer on the initial test (correct/previous answer); (b) a correct answer that was neither the incorrect previous answer nor the corrective feedback on the first test (correct/*other*); (c) an incorrect answer that matched the incorrect 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 incorrect previous answer nor the corrective feedback on the first test (incorrect/*other*). We limited the analysis to incorrect answers on the first 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.

The distribution of responses across the five answer types for each experimental group is shown in Table 2. To analyse the data, we focused on responses that were incorrect on both tests (bottom panel of Table 2). Doing so allowed us to compare endorsements of previous answers and corrective feedback separately while holding constant the level of accuracy on both tests. A 2 (test type: two, one) x 2 (repetition type: repetition, no repetition) x 3 (answer type: feedback, *other*, previous answer) mixed-factor ANOVA, with the probability of answering with each type of answer as the dependent variable, showed that the main effect of test type was not significant, $F < 1$. However, there was a significant main effect of repetition type, $F(1, 134) = 4.29, p = .04, \eta_p^2 = .03$; the probability of answering with the three types of incorrect final-test answers was higher in the repetition condition ($M = .24, SD = .10$) than the no-repetition condition ($M = .20, SD = .10$). Also, we found a significant main effect of final-test answer type $F(2, 268) = 36.46, p < .001, \eta_p^2 = .21$. A paired-sample t -test showed that

participants were more likely to endorse the corrective feedback ($M = .33$, $SD = .27$) than an *other* answer ($M = .20$, $SD = .19$), $t(137) = 4.27$, $p < .001$, $d = .58$, or the previous answer ($M = .13$, $SD = .13$), $t(137) = 7.92$, $p < .001$, $d = .98$. Also, they endorsed an *other* answer more than the previous answer $t(137) = 3.75$, $p < .001$, $d = .45$.

The main effects of repetition type and answer type were qualified by a significant interaction, $F(2, 268) = 18.80$, $p < .001$, $\eta_p^2 = .12$. For the repetition condition, a paired-sample t -test revealed that participants overwhelmingly selected the corrective feedback from the initial test ($M = .44$, $SD = .28$) compared to *other* answers ($M = .16$, $SD = .15$), $t(68) = 6.62$, $p < .001$, $d = 1.25$, and previous answers ($M = .12$, $SD = .13$), $t(68) = 8.84$, $p < .001$, $d = 1.49$, whereas no difference was found between *other* answers and previous answers, $t(68) = 1.63$, $p = .11$, $d = .30$. For the no-repetition condition, however, we found no difference between endorsing the corrective feedback ($M = .23$, $SD = .22$), and *other* answers ($M = .24$, $SD = .21$), $t(68) = -0.23$, $p = .82$, $d = .04$, but both probabilities were higher than the probability of choosing the previous answer ($M = .14$, $SD = .13$), $t(68) = 2.88$, $p < .01$, $d = .52$, and $t(68) = 3.58$, $p < .001$, $d = .59$, respectively. No other interaction was significant, largest $F(1, 134) = 1.98$, $p = .16$, $\eta_p^2 = .01$.

False Endorsements of Corrective Feedback. The previous analysis was limited to cases of incorrect responses on both tests. Here, we conducted another 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 rate of endorsing the corrective feedback between the related items and new items. 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 (test type: two, one) x 2 (repetition type: repetition, no repetition) x 2 (question type: new, related) mixed-factor ANOVA, with probability of answering with the corrective feedback option as the dependent variable, yielded a significant main effect of repetition type, $F(1, 134) = 31.30, p < .001, \eta_p^2 = .19$. The probability of answering with the corrective feedback was higher in the repetition condition ($M = .31, SD = .11$) compared to the no-repetition condition ($M = .21, SD = .09$). There was also a significant main effect of question type, $F(1, 134) = 46.53, p < .001, \eta_p^2 = .26$; the probability of answering with the corrective feedback was higher for the related items ($M = .33, SD = .20$) compared to new items ($M = .20, SD = .11$). These main effects were qualified by a significant interaction, $F(1, 134) = 17.44, p < .001, \eta_p^2 = .12$. A paired-sample *t*-test revealed that participants in all the four groups were more likely to endorse the corrective feedback on related questions compared to new, but the difference was larger in the repetition groups, $t(68) = 7.32, p < .001, d = 1.30$, (related: $M = .41, SD = .20$; new: $M = .21, SD = .12$) compared to the no-repetition groups, $t(68) = 2.02, p = .046, d = .37$, (related: $M = .24, SD = .17$; new: $M = .19, SD = .10$). No other main effect or interaction was significant, largest $F < 1$.

Discussion

The final-test accuracy results showed that manipulating the number of tests (i.e., two-tests groups vs. the one-test groups) had no effect on participants' later performance. In contrast, manipulating the presence of repeated items in the final test changed the way

participants performed on the related items. Specifically, for the repetition groups, taking an initial MC test harmed participants' performance on the related items compared to the new items. For the no-repetition groups, however, the difference between participants' performance on the related versus new items was not significant (although numerically, performance remained poorer on the related items than on the new items). These findings were supported by the analysis of response types conditioned on an inaccurate initial test response. That analysis showed that participants predominantly selected the corrective feedback on the final test in the repetition groups, whereas selections were about evenly split between corrective feedback and *other* responses in the no-repetition groups. In short, removing repeated items greatly reduced the allure of the corrective feedback on the final test.

These findings are surprising given Higham et al.'s (2016) results, which showed impairment on the related items even when the repeated items were dropped from the final test. The different results might be attributed to the different methodologies used in Higham et al.'s study versus the current one. For example, Higham et al. presented the final test in standard MC format and the whole study was self-paced. In contrast, our study used a ranking format and presented the questions for at least 15 s, which was true for the whole task including the final test. Conceivably, the ranking format and the requirement to process the questions for a minimum of 15 s in Experiment 1 invoked more controlled processes which tempered automatic influences. Importantly, however, if controlled processing was promoted with the final-test procedure used in Experiment 1, it was not enough to produce facilitation on related items (vs. new) as has been observed with CR final tests (e.g., Alamri & Higham, 2022; Little et al., 2012, 2019).

Although there was no difference in accuracy between the related and new items in the no-repetition condition, a deeper analysis showed that automatic influences were still

present. In particular, the analysis that focused specifically on the tendency to select the corrective feedback regardless of the accuracy of the initial test response showed that both the repetition and no-repetition groups endorsed the corrective feedback for related items more than for new items. This difference was more pronounced when the repeated items were included in the final test, which suggested that presenting the repeated items enhanced automatic influences. Nonetheless, the residual effect in the no-repetition groups was still statistically significant, suggesting that automatic influences still affected performance even when no repeated items were included on the final test.

Overall, contrary to our hypotheses, controlled influences did not dominate responding on the MC final test despite using educational materials (SAT questions), administering a single test, or eliminating repeated items. Although automatic influences were tempered, particularly by the elimination of repeated items, they still tended to overshadow the controlled influences.

One possibility is that online tests that are completed on MTurk for payment are simply not high stakes enough to promote controlled influences if the test is MC. Although controlled influences have been shown with online studies if the final test is CR (e.g., Alamri & Higham, 2022; Little et al., 2019), it may be that the explicit presentation of the corrective feedback amongst the MC options is simply too appealing to be resisted in a testing environment of this sort. Consequently, in Experiment 2, we investigated whether evidence of controlled influences on MC final-test performance might emerge in a real educational context with students who are motivated to learn.

Experiment 2

In Experiment 2, a group of introductory psychology students took a single MC test that consisted of related and control questions and no repeated items were included (as is true

of most educational tests). Thus, the group of participants tested in Experiment 2 were similar to the single-test, no-repetition group in Experiment 1. The main manipulation was the sequencing of the related items which were separated by different lags. Specifically, some related items were separated by several other items whereas some were presented back-to-back.

We used a lag manipulation because of its potential to separate controlled and automatic influences in this paradigm. Our logic was similar to that underpinning research on the *false fame effect*. The false fame effect occurs when people incorrectly judge non-famous names to be famous names because they are made familiar within the experimental context. For example, Jacoby et al. (1989, Experiment 1) had participants read a list of non-famous names and told them that none of the names was famous. After reading the list, participants were shown old non-famous names from the list, some new non-famous names (control), and some new famous names. They were asked to judge the fame of each name. If the fame-judgment task immediately followed reading the list, the old non-famous names were rated as less famous than the new non-famous names. With an immediate test, participants recollected that the old non-famous names were presented in the list of non-famous names and could use that information in a controlled and intentional manner to reject the names as non-famous. However, if the fame judgment task occurred 24 hr after reading the list of non-famous names, the old non-famous names were rated as *more* famous than the new non-famous names.

Jacoby et al. (1989) reasoned that after a 24 hr delay, participants were less likely to recollect encountering the old non-famous names in the list they read earlier than if the fame judgment task was completed immediately. However, they still experienced a feeling of familiarity for the old non-famous names which they incorrectly attributed to the name being famous. To use Jacoby et al.'s phrase, the names "became famous overnight". Jacoby et al.

likened the false fame effect to the *sleeper effect* in persuasion research whereby people reject a persuasive message from a low-reliability source if they are tested immediately after receiving the message but show evidence of being persuaded by the message if tested after a delay (e.g., Hovland & Weiss, 1951). Presumably, if people recollect the source of the message on an immediate test, they can reject its influence in a controlled manner. However, if the source is forgotten but the message is retained after a delay, the message has an automatic influence on attitudes.

Analogously, we reasoned that by presenting related items back-to-back such that there was minimal delay between the related pairs, the role of controlled processes would be maximized. That is, the products of any retrieval processes evoked to reject lures when answering the first question would be readily accessible to help participants to answer the second related question. Also, false recognition of the second question resulting from automatic influences would be low because the short lag would highlight the differences between the questions. Therefore, we expected controlled influences to override automatic influences in the back-to-back condition. However, at longer lags, these controlled influences would taper off, potentially allowing unchecked automatic influences (e.g., misleading familiarity of the feedback) to exert an influence, in much the same way that non-famous names were judged to be famous after a lag. However, because of the high-stakes educational context, we hypothesized that we might observe controlled influences continue to dominate responses and benefit performance even with a lag, but not to the same degree as with no lag.⁶

⁶ One difference between our current experiment on the one hand and the false-fame and sleeper paradigms on the other is the role of source memory. Unlike the false-fame and sleeper paradigms, increased lag in our paradigm can limit the ability of controlled processes to oppose automatic influences in ways that do not necessarily involve memory for source. For example, memory for the question stem, memory for the association between the stem and the options, and memory for the reasoning processes and the information retrieved that leads to acceptance or rejection of the options, are all potentially decreased with greater lag between related

Method

Participants

Participants were students enrolled in an introductory psychology module at the University of Southampton and attended a tutorial at the end of term. We tested 171 students; however, seven students were excluded from the final analysis due to failure to follow instructions such as not ranking the alternatives for most or all of the questions. The final analysis involved the remaining 164 participants (male = 22), with ages ranging between 17 and 39 years ($M = 19.37$, $SD = 3.07$). Also, there were three counterbalancing formats with 54-55 participants each, as explained later.

As the study was completed in a genuine university classroom, we had no control over our sample size and so we did not conduct an a priori power analysis. However, we did conduct a post hoc sensitivity analysis based on $\alpha = .05$, power = .80, and 164 participants. That analysis indicated our sample size was large enough to detect a small (or larger) within-subjects effect of question type (Cohen's $f = 0.10$).

Design

The experiment had one independent variable, question type, with three levels: related-separated, related-back-to-back, and new. The main dependent variable was the participants' mean performance on the final test. Although the test was constructed such that

questions. The reduction of these controlled processes may also make participants more vulnerable to falsely recognizing the related question because those processes help participants to discriminate between the related questions. Thus, whereas increased lag likely lessened the effectiveness of controlled processes via reduced source memory in the false fame and sleeper paradigms, increased lag will likely reduce the effectiveness of controlled processes in a somewhat different manner in our paradigm. Notably, the idea that controlled processes can limit automatic influences by means other than memory for source has been adopted in many other areas of psychology that have investigated controlled and automatic processes including unconscious perception (e.g., Debner & Jacoby, 1994), the Stroop effect (e.g., Jacoby et al., 2003), stereotyping and other social psychological phenomena (e.g., Payne, 2008), and implicit learning of artificial grammars strings (e.g., Higham et al., 2000), to name a few.

it appeared to students as single test, we maintained the “initial test” and “final test” naming system used in Experiment 1 for convenience. Twenty-two questions served as the initial test for the related-separated and related-back-to-back conditions while 33 questions served as the final test. The 33 final-test questions consisted of 11 related (separated) questions, 11 related (back-to-back) questions, and 11 new (control) questions. To control the lags between the questions, all the questions were presented in a fixed order. Also, the MC alternatives were always presented in the same order per question. To ensure that each question served in each condition equally often, we created three surveys on Qualtrics. Following a similar procedure as Experiment 1, the three surveys rotated the questions through the three experimental conditions across students. The final test items were held constant across the three surveys with only the initial test items varied to create the experimental conditions.

The lefthand panel of Figure 5 summarizes the design. The first 11 questions of the test were the first questions of the 11 related-separated pairs (e.g., questions 1-11 A) which counted as the initial test for the related-separated condition. Then the 22 related-back-to-back questions were presented where the first and second questions of the same pair were presented in immediate succession (e.g., 12-22 A/B; question 12 A followed immediately by question 12 B, followed by question 13 A, then 13 B, and so on). The first question in each pair counted toward initial test performance (e.g., 12-22 A) whereas the second question counted toward the final test (e.g., 12-22 B). After that, the 11 control questions were presented (e.g., 23-33 B). Finally, the second set of questions from the 11 related-separated pairs was presented (e.g., 1-11 B). These questions counted as the final test for the related-separated condition. Therefore, the final-test analysis included 11 questions from each of the related-back-to-back, related-separated, and new conditions. Twenty-two questions were not included in the final test analysis as they counted as the first test.

Materials and Procedure

The materials were 33 MC question pairs pertaining to an introductory psychology module which covered most of the topics presented in the lectures. As with the materials in Experiment 1, all the questions were formed into pairs based on the topic as they were related conceptually (e.g., both questions in a given pair were about founders of different psychological approaches) and each pair shared the same alternatives. The questions are available as Supplementary Materials.

The test was administered in a face-to-face environment two weeks before the final exam after all the weekly lectures were finished. Students were divided into small groups (30 or fewer) and tested over seven sessions. They were told that this test was practice for the final exam and would not directly affect their final marks for the module. However, they were also told that they should take the test seriously and try their best to maximize the benefits of the practice. The questions were presented via Qualtrics survey software and students accessed the test individually via a special link using their personal device. All questions were presented in a single test and students gave their answers with an MC ranking format identical to that in Experiment 1. Feedback was presented immediately after each question for all the 55 questions regardless of whether the chosen option was correct or not (e.g., “*The correct answer is classical conditioning*”). The test was self-paced and took each student approximately 20-30 minutes to complete. The scoring method was identical to that used in Experiment 1.

Results

Initial Test Performance

Mean accuracy on the initial test was .77 ($SD = .09$).

Final Test Performance

Final-Test Accuracy. A one-way within-subjects ANOVA on final test accuracy in the related-separated, related-back-to-back, and new conditions revealed a significant main effect of question type, $F(2, 489) = 15.00, p < .001, \eta_p^2 = .06$ (see Figure 6). Paired-sample t -tests showed that participants performed better on the related-back-to-back questions ($M = .81, SD = .13$), compared to the related-separated questions ($M = .77, SD = .13$), $t(163) = 2.97, p < .01, d = .30$, and new questions ($M = .73, SD = .14$), $t(163) = 6.00, p < .001, d = .61$. Also, participants had more accurate answers on the related-separated questions compared to the new questions $t(163) = -3.20, p < .01, d = .30$.

Final-Test Answer Types. As in Experiment 1, we conducted an analysis of the final-test answers to related questions conditioned on incorrect initial test answers, producing the same five mutually exclusive possibilities (see Table 3). We then analysed the three possibilities where the answers were incorrect on both the first and final tests. A 2 (related-item type: back-to-back, separated) 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 related item type was not significant, $F < 1$. However, there was a significant main effect of final-test answer type, $F(2, 652) = 25.20, p < .001, \eta_p^2 = .07$. Paired-sample t -tests showed a comparable probability of endorsing the *other* answers ($M = .17, SD = .20$) and previous answers ($M = .16, SD = .21$), $t(327) = -0.62, p = .53, d = .05$, but both probabilities were higher than that for the corrective feedback ($M = .08, SD = .15$), $t(327) = -6.75, p < .001, d = .52$, and $t(327) = -5.80, p < .001, d = .44$, respectively. Moreover, we found a significant interaction between related-item type and answer type, $F(2, 652) = 14.08, p < .001, \eta_p^2 = .04$. For the related-back-to-back questions, paired-sample t -tests showed that participants were more likely to select an *other* answer ($M = .17, SD = .19$), and previous answer ($M = .19, SD = .24$), than the corrective feedback ($M = .04, SD = .09$), $t(163) = -8.24, p < .001, d = .87$, and $t(163) = -8.02, p < .001, d$

= .82, respectively. No significant difference was found between the probability of selecting an *other* answer and the previous answer $t(163) = -0.88, p = .38, d = .10$. For the related-separated items, participants selected an *other* answer ($M = .17, SD = .20$), more than the corrective feedback ($M = .12, SD = .18$), $t(163) = -2.29, p = .02, d = .25$, and the previous answer ($M = .13, SD = .16$), $t(163) = -2.26, p = .02, d = .23$. However, no difference was found between the probability of selecting the corrective feedback and the previous answer, $t(163) = -0.37, p = .70, d = .04$.

False Endorsements of Corrective Feedback. As in Experiment 1, we analysed the probability of endorsing the corrective feedback between the related items (i.e., separated, back-to-back) and the new items (see Figure 7). A one-way ANOVA revealed a significant effect, $F(2, 489) = 96.1, p < .001, \eta_p^2 = .28$. Paired-sample *t*-tests showed that participants were more likely to endorse the corrective feedback on the new items ($M = .18, SD = .13$), more than the related-back-to-back ($M = .02, SD = .05$), $t(163) = 15.71, p < .001, d = 1.63$, and related-separated items ($M = .10, SD = .11$), $t(163) = -5.76, p < .001, d = .61$. Also, participants endorsed the corrective feedback on the related-separated items more than the related-back-to-back items, $t(163) = -9.13, p < .001, d = .95$.

Discussion

In contrast to the results of Experiment 1, the results of Experiment 2 demonstrated that taking initial MC testing *enhanced* performance on later MC related questions regardless of the different lags separating the questions. That is, participants performed *better* on both related-back-to-back and related-separated questions than on new questions. Moreover, participants were *less* likely to select the corrective feedback on both types of related questions compared to new questions. Also, an analysis of the final-test answer types conditioned on incorrect initial responses produced results that were very different from

Experiment 1. Specifically, it showed that participants were no longer selecting the corrective feedback at a high rate and previous answers at a low rate, particularly for related-back-to-back questions. These data are the first that we know of to demonstrate a benefit rather than a detriment to overall performance with related (vs. new) questions when both the initial and final tests are MC. They are also the first data to show that participants were not seduced by the corrective feedback and avoidant of their previous answers when answering related questions on the final test.

A key to understanding why the reversal occurred can be found by comparing the two types of related questions. As we hypothesized, sequencing the pairs of related questions so that the members of each pair appeared on the test in immediate succession produced the best performance and the greatest benefit of retrieval practice. By our reasoning, by presenting the related questions back-to-back, automatic influences (associated with false recognition of the second, related question) would be kept to a minimum because the retention interval between the questions was negligible allowing controlled processes to counter automatic influences. Also, automatic influences would be kept to a minimum in the context of a single practice test taken in a genuine educational setting because students likely assumed that questions would not be repeated, an assumption that was correct (i.e., there were no repeated questions in the test).

In addition to limited automatic influences, students in Experiment 2 received corrective feedback throughout the whole test which would have helped them to notice changes to the related questions, particularly in the related-back-to-back condition. Noticing changes to related questions may have, in turn, improved students' use of controlled strategies while taking the test, encouraging them to closely read each question. Together, these factors, coupled with the fact that students were motivated to score well on the test,

were enough for controlled influences of retrieval practice to dominate responding, resulting in a benefit rather than an impairment to final-test performance.

Experiment 3

We established for the first time in Experiment 2 that it is possible to produce facilitated performance on related questions when both tests are MC. In Experiment 3, we again tested students in a genuine educational environment in the hopes that once again automatic influences would be kept to a minimum and controlled influences would dominate responding, just as they did in Experiment 2. In addition, we revisited the role of feedback in Experiment 3. Alamri and Higham (2024) found that corrective feedback on the initial test worsened performance on related questions compared to no feedback. However, that difference was observed when automatic influences were dominating responding. Conversely, Little et al. (2012) found that initial-test feedback made little difference to the facilitation observed with related questions when the final test was CR. If facilitation is observed in Experiment 3, providing evidence that controlled influences are dominating responding, then the scenario would be similar to Little et al.'s experiments. Therefore, we expected that feedback would have little effect on the size of the controlled influence, just as Little et al. found.

Feedback on the final test may also have played a role in Experiment 2. As noted earlier, noticing changes between the related questions may have been crucial to adopting strategies during testing that allowed controlled influences to prevail. Conceivably, removing corrective feedback during the final test would reduce the role of change detection and allow automatic influences to dominate responding, just as they have in most experiments with MC final tests. Experiment 3 provides a test of this possibility.

Method

Participants

Participants were students enrolled in the introductory psychology module at the University of Southampton who attended an online tutorial at the end of term. The final analysis involved 223 participants (male = 31), with ages ranging between 18 to 26 years ($M = 18.77$, $SD = 1.02$). The experiment had two groups, with 114 participants in the feedback group and 109 in the no-feedback group, and two counterbalancing formats with 54-57 participants each as explained later.

As in Experiment 2, the study was completed in a real university classroom with a fixed sample size, so we did not conduct an a priori power analysis. However, a post hoc sensitivity analysis based on $\alpha = .05$, $\text{power} = .80$, and 223 participants revealed that our sample size was large enough to detect a small (or larger) within-subjects effect of question type (Cohen's $f = 0.09$).

Design

The experiment employed a 2 x 2 mixed factorial design with feedback type on the initial test (feedback, no-feedback) manipulated between subjects, and question type on the final test (new, related) manipulated within subjects. The main dependent variable was the participants' mean performance on the final test. As in Experiment 2, although the test was designed such that it appeared to students as a single test, we retained the "initial test" and "final test" terms for convenience. The first test involved 22 questions which were the first questions of the related pairs (e.g., 1-22 A). The final test involved 44 questions consisting of 22 related questions (the second questions from the related pairs, e.g., 1-22 B) and 22 new (control) questions (e.g., 23-44 B). To rotate questions through the experimental conditions across students, we created two surveys on Qualtrics that varied the items presented during the initial test, leaving the items in the final test unchanged. The initial test was presented in a

fixed random order as well as the related and new questions in the final test which were presented in a fixed random order. Also, the MC alternatives were always presented in the same order. A summary of the design is shown in Figure 8.

Materials and Procedure

Forty-four MC related pairs were generated from the introductory psychology materials which covered most of the topics presented in the weekly lectures. Twenty-nine pairs were identical to those used in Experiment 2, whereas 15 new pairs were generated to accommodate new lecture material. The questions are available as Supplementary Materials.

The test was conducted online via Blackboard as a part of an online tutorial held at the end of the term after all the weekly lectures were completed. Students were told that, although their scores would not count toward their final mark, the test should be considered a substitute for a final exam that was not possible to have during the coronavirus pandemic. Although the scores did not count, there was ample evidence that students were taking the test seriously. For example, several students remarked that they would have liked more forewarning of the test to allow them to prepare for it.

The questions were presented via Qualtrics survey software and students individually accessed the test on their own device via a link sent to them online. Students started the test by answering the 22 questions which counted as the initial test. Half of the students were provided with corrective feedback after each question during the initial test (i.e., feedback group) regardless of whether the chosen option was correct or not (e.g., “*The correct answer is classical conditioning*”), whereas the other half were not (i.e., no-feedback group). Then students completed the final test that contained the 22 related untested questions and the 22 new questions which were presented in a fixed random order. No corrective feedback was provided during the final test for any student. However, all students were told that the

questions and answers to all the questions would be made available on Blackboard after the test was completed. The same ranking test format and scoring method were used as in Experiment 2.

Results

Initial Test Performance

For the feedback group, the mean accuracy was .75 ($SD = .09$), whereas it was .77 ($SD = .10$) for the no-feedback group. An independent samples t -test showed that this difference was significant $t(217) = -2.01, p = .045, d = .27$. However, the difference between the groups was small in terms of both magnitude and effect size.

Final Test Performance

Final-Test Accuracy. We conducted a 2 (feedback type: feedback, no feedback) x 2 (question type: new, related) mixed-factor ANOVA with final-test accuracy as the dependent variable (see Figure 9). The main effect of feedback type was not significant, $F < 1$; accuracy was comparable between the feedback group ($M = .71; SD = .10$) and the no-feedback group ($M = .72; SD = .10$). However, there was a significant main effect of question type, $F(1, 221) = 17.87, p < .001, \eta_p^2 = .07$. Accuracy was higher on the related items ($M = .73, SD = .13$) compared to the new items ($M = .69, SD = .12$). The interaction was not significant, $F < 1$.

Final-Test Answer Types. We conducted the same analysis as in Experiments 1 and 2 on final-test answer types conditioned on incorrect initial-test answers. The analysis produced the same five mutually exclusive final-test response rates (see Table 4). We then analysed the three rates where the answers were incorrect on both the first and final tests for the same reasons as in previous experiments. A 2 (feedback type: feedback, no feedback) 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 feedback type was not significant, $F < 1$. However, there was a significant main effect of final-test answer type, $F(2, 442) = 10.73, p < .001, \eta_p^2 = .05$. A paired-sample t -test showed that participants were more likely to select an *other* answer ($M = .21, SD = .17$) than either the corrective feedback ($M = .16, SD = .16$), $t(222) = -3.23, p < .001, d = .29$, or the previous answer ($M = .15, SD = .16$), $t(222) = 4.55, p < .001, d = .40$. There was no difference between the probability of selecting the corrective feedback and the previous answer $t(222) = 1.11, p = .26, d = .10$. Finally, we found marginal interaction between feedback type and answer type $F(2, 442) = 2.36, p = .09, \eta_p^2 = .01$.

False Endorsements of Corrective Feedback. As in Experiments 1 and 2, we analysed the probability of endorsing the corrective feedback on related and new questions (see Figure 10). A 2 (feedback type: feedback, no feedback) x 2 (question type: new, related) mixed-factor ANOVA showed that the main effect of feedback type was not significant $F < 1$. However, there was a significant main effect of question type, $F(1, 221) = 78.86, p < .001, \eta_p^2 = .26$, such that the probability of answering with the corrective feedback was higher for the new items ($M = .19, SD = .09$) compared to the related ones ($M = .12, SD = .10$). The interaction between feedback type and question type was not significant, $F < 1$.

Discussion

The results of this experiment were largely consistent with those in Experiment 2 in that participants performed better on the related items than on the new items. Also, participants endorsed the corrective feedback on the related items less than for new items, and when participants answered incorrectly on both the first and final tests, participants selected an *other* answer on the final test more often than they selected the corrective feedback or the previous answer. In other words, the corrective feedback was not an appealing choice, just as it was not in Experiment 2, but in stark contrast to Experiment 1. These results provide

evidence of controlled influences dominating participants' performance in this experiment, just as in Experiment 2.

Experiment 3 also showed that the provision of corrective feedback on the initial test had little effect on performance. Participants benefited from answering related questions to the same extent regardless of feedback. These results are broadly consistent with those obtained by Little and colleagues who have also found that feedback had little effect on performance (e.g., Little et al., 2012). Thus, when automatic influences prevail, feedback has a deleterious effect on performance (e.g., Alamri & Higham, 2024). On the other hand, when controlled influences prevail, initial-test feedback has little effect.

Experiment 3 also demonstrated that feedback during the final test, which would have facilitated change detection between the related questions, was not necessary for controlled influences to dominate responding. Thus, writing an MC test with no repeated questions in a genuine educational environment with students who are keen to perform well appears to be more important than final-test feedback in reaping the benefits of retrieval practice with related questions. However, an important question might be asked: would controlled influences remain dominant when including repeated items in a final MC test conducted in a genuine educational environment? The results of Experiment 1 suggested that adding repeated items worsened performance on related items (vs. new) where the test was conducted in an online platform (MTurk). Hence, in Experiment 4 we examined the role of adding repeated items but where the test was conducted in a genuine educational environment.

Experiment 4

The results of Experiment 1 showed impaired performance with related items in the repetition groups compared to the no-repetition groups. However, these results were observed when automatic influences were dominating responses. Therefore, in Experiment 4 we

explored whether a similar effect would be observed in a real classroom where controlled influences are dominating as we observed in Experiments 2 and 3. Specifically, in Experiment 4, we partially replicated the procedure of Experiment 2 except that we included repeated items in the final test. If facilitation is observed in Experiment 4, then that would provide more evidence of the controlled influences dominating performance in a genuine educational environment. However, adding repeated items in the final test may make participants less cautious about checking automatic feelings of familiarity, thereby increasing false recognition of related items. A reduction in attempts to counter automatic influences would lead to worse performance with related items (vs. new) and show evidence of automatic influences dominating responding in an educational context for the first time.

In addition to the inclusion of repeated items in the final test list, the other significant change in Experiment 4 in comparison to Experiment 2 was to present most of the items in a randomized order rather than in blocks. That is, apart from a block of related-back-to-back items presented near the middle of the test, the items presented before and after that block were presented in randomized order. This change was made to reduce the potential for carry-over effects from one block to the next.

Method

Participants

As in Experiments 2 and 3, participants were students enrolled in an introductory psychology module at the University of Southampton who attended a tutorial at the end of term during which the initial and final tests were administered. The final analysis involved 224 participants (male = 39), with ages ranging between 18 to 30 years ($M = 18.96$, $SD = 1.47$). The experiment had four counterbalancing formats with 52-60 participants each.

As in Experiments 2 and 3, the sample size of this experiment was fixed, so we did not conduct an a priori power analysis. However, a post hoc sensitivity analysis based on $\alpha = .05$, power = .80, and 224 participants revealed that our sample size was large enough to detect a small (or larger) within-subjects effect of question type (Cohen's $f = 0.08$).

Design and Materials

As in Experiment 2, we maintained the initial/final test naming convention even though students wrote a single test from their point of view. For the materials, we used the same 44 related MC pairs that we used in Experiment 3 (88 individual questions in total), but only 77 questions appeared across the initial and final tests for any given student. The design was similar to that used in Experiment 2, except that not all experimental conditions were blocked and we included repeated questions on the final test in this experiment.

A summary of the design is shown on the righthand panel of Figure 5. The initial test in this experiment comprised 33 questions instead of 22 as in Experiment 2. Specifically, the first 22 items of the initial test questions were a random mixture of the first questions from the related-separated pairs (e.g., 1-11 A) and the repeated pairs (e.g., 12-22 B). Following those questions, a block of 22 related-back-to-back questions were presented in pairs (e.g., 23-33 A/B), with the initial- (e.g., 23-33 A) and final-test (e.g., 23-33 B) questions shown in immediate succession. Finally, following the related-back-to-back block, 33 final test questions were presented which consisted of a random intermixture of the second pairs of the related-separated (e.g., 1-11 B) and repeated (e.g., 12-22 B) pairs, plus new control (11) items (e.g., 34-44 B). Thus, the final test comprised 44 items in this experiment instead of 33 as in Experiment 2.

We created four surveys on Qualtrics which rotated the questions through the four experimental conditions (repeated, related-back-to-back, related-separated, and new) across

students. The final test items were the same across the counterbalance versions; only the constitution of initial test items varied to create the four experimental conditions.

Procedure

The procedure was similar to Experiment 2 except for the changes to the constitution and ordering of the items on the test as explained earlier and that the students were tested as a single, large group at the end of the semester instead of in multiple smaller groups. As in Experiment 2, students were told that the test did not count toward their final mark and should be considered a practice test for the final exam which was to follow in a few weeks. Feedback was presented immediately after each question for all questions regardless of whether or not the chosen option was correct. The same ranking test format and scoring method were used as in the previous experiments.

Results

Initial Test Performance

Mean accuracy on the initial test was .76 ($SD = .11$).

Final Test Performance

Final-Test Accuracy. A one-way within-subjects ANOVA on comparing final test accuracy between the repeated, related-separated, related back-to-back, and new conditions revealed a significant main effect of question type $F(3, 892) = 60.12, p < .001, \eta_p^2 = .17$ (see Figure 11). Paired-sample t -tests showed that participants performed better on the repeated questions ($M = .88, SD = .11$), compared to the related-back-to-back questions ($M = .81, SD = .13$), $t(223) = -7.40, p < .001, d = .57$, new questions ($M = .75, SD = .14$), $t(223) = 13.04, p < .001, d = 1.02$, and related-separated questions ($M = .71, SD = .18$), $t(223) = 14.23, p < .001, d = 1.12$. Also, participants provided more accurate answers to the related-back-to-back

questions compared to the new and related-separated questions $t(223) = 6.01, p < .001, d = .46$ and $t(223) = 8.52, p < .001, d = .64$, respectively. Interestingly, despite better accuracy for related questions that were presented back-to-back compared to new questions, participants showed the opposite pattern on related questions that were separated (i.e., new > related-separated), $t(223) = -3.46, p < .001, d = .24$.

Final-Test Answer Types. As in the previous experiments, we conducted an analysis of the final-test answers to related questions conditioned on incorrect initial test answers, producing the same five mutually exclusive possibilities (see Table 5). We then analysed the three possibilities where the answers were incorrect on both the first and final tests. A 2 (related-item type: back-to-back, separated) 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 related-item type was significant, $F(1, 446) = 10.29, p < .01, \eta_p^2 = .02$. The probability of answering with one of the three types of incorrect answer was higher for related-separated items ($M = .19, SD = .16$) than the related-back-to-back items ($M = .14, SD = .12$). Also, there was a significant main effect of final-test answer type, $F(2, 892) = 7.79, p < .001, \eta_p^2 = .02$. Paired-sample *t*-tests showed a comparable probability of endorsing the feedback answers ($M = .15, SD = .26$) and previous answers ($M = .14, SD = .22$), $t(447) = 0.50, p = .61, d = .03$, but both probabilities were lower than that for an *other* answer ($M = .20, SD = .22$), $t(447) = -2.68, p < .01, d = .18$, and $t(447) = 3.66, p < .001, d = .24$, respectively. Moreover, we found a significant interaction between related-item type and answer type, $F(2, 892) = 82.50, p < .001, \eta_p^2 = .16$. For the related-back-to-back questions, paired-sample *t*-tests showed that participants were more likely to select an *other* answer ($M = .22, SD = .23$), and previous answer ($M = .17, SD = .23$), than the corrective feedback ($M = .02, SD = .07$), $t(223) = -12.21, p < .001, d = 1.15$, and $t(223) = -9.65, p < .001, d = .87$, respectively. Also, the

probability of selecting an *other* answer was higher than the previous answer $t(223) = -2.43, p < .01, d = .22$. For the related-separated items, participants selected the feedback answer ($M = .29, SD = .33$), more than an *other* answer ($M = .17, SD = .21$), $t(223) = 4.69, p < .001, d = .43$, and the previous answer ($M = .12, SD = .20$), $t(223) = 7.36, p < .001, d = .63$. Also, there was a significant difference between the probability of selecting an *other* answer and the previous answer, $t(223) = 2.78, p < .01, d = .26$.

False Endorsements of Corrective Feedback. As in Experiment 2, we analysed the probability of endorsing the corrective feedback between the related items (i.e., separated, back-to-back) and the new items (see Figure 12). A one-way ANOVA revealed a significant effect, $F(2, 669) = 163.8, p < .001, \eta_p^2 = .33$. Paired-sample t -tests showed that participants were more likely to endorse the corrective feedback for the new items ($M = .19, SD = .13$), and related-separated items ($M = .20, SD = .16$), than the related-back-to-back items ($M = .02, SD = .04$), $t(223) = 18.72, p < .001, d = 1.75$ and $t(163) = -17.76, p < .001, d = 1.64$, respectively. However, there was no difference between the probability of endorsing the corrective feedback on the related-separated items and the new items, $t(223) = -1.16, p = .24, d = .10$.

Discussion

In contrast to the results of Experiments 2 and 3, the results of Experiment 4 showed for the first time, that taking initial MC testing could sometimes impair performance on later MC related questions in a genuine educational environment if there is a lag between the related question pairs. That is, participants performed better on repeated, related-back-to-back, and new questions than on related-separated questions. Adding repeated questions to the final test appears to have played a major role in the reversed results that we observed here versus Experiments 2 and 3. These findings were supported by the analysis of the final-test answer

types when both the initial- and final-test answers were incorrect. It demonstrated that when answering related-separated questions, participants were more likely to select the corrective feedback than other answer types (i.e., other, previous answer). Analysing the probability of endorsing the corrective feedback on the final test showed no significant difference between related-separated and new questions (although, numerically, the probability was higher on related-separated items than on the new items).

These results indicate that, apart from participants' performance on the related-back-to-back questions which replicated what we observed in Experiment 2, the results on the related-separated (vs. new) questions were very different from Experiment 2. That is participants in Experiment 2 performed better on the related-separated questions (vs. new), selected an *other* answer more than corrective feedback on related-separated questions, and selected the corrective feedback on the new questions significantly more often than on related-separated questions. When repeated items were not included on the final test in Experiment 2, automatic influences would be kept to a minimum because students likely assumed that questions would not be repeated and consequently were not searching the list for repeated items.⁷ In contrast, including repeated questions on the final test likely changed participants' search strategy and seduced participants to falsely recognize more related questions, which resulted in automatic influences dominating responding. The results of the related-separated questions in this experiment were largely consistent with those for the related items in

⁷ Some readers may argue that if the presence of repeated items encourages a deliberate search strategy for previously answered questions, then any negative influence on performance with related-separated items is a controlled rather than automatic influence. However, Mandler's (1980) "butcher on the bus" example might clarify our position. Mandler argued that when people see their butcher on the bus, out of the context in which he is usually encountered, people may experience an automatic feeling of familiarity (i.e., "I've seen that man before, but I don't remember where"). The likelihood of experiencing this automatic familiarity influence would likely be greater if people adopted a search strategy to look for familiar people on the bus than if they did not. Thus, by this analysis, a deliberate search strategy can affect the likelihood of automatic influences.

Experiment 1 in that they emphasised the negative effect of including repeated questions in the final MC test.

General Discussion

Over four experiments, we examined the effects of taking an initial MC practice test on participants' performance with related versus new items on a final MC test. The goal of the research was to determine whether there were conditions under which controlled processes might override automatic ones when both tests were MC. To date, controlled influences have been identified when the final test is CR (e.g., Alamri & Higham, 2022; Little et al., 2012, 2019). In contrast, there is only evidence in the literature for automatic influences when the final test is MC (e.g., Alamri & Higham, 2022, 2024; Higham et al., 2016).

If only automatic influences occurred when testing is entirely MC, then this finding would be important for at least two reasons. First, it potentially severely limits the utility of using MC tests as a learning tool. Instructors may create final MC exams that contain questions related to earlier practice questions (e.g., questions that query the same topic and/or contain similar options), but which are worded differently. If it is not possible to create learning conditions that promote controlled influences of retrieval practice, then automatic influences with such test combinations may undermine assessment results rather than enhance them. Second, from a theoretical perspective, the dual-process theory of retrieval practice that Alamri and Higham (2022) forwarded assumes that controlled and automatic influences occur with both MC and CR tests (i.e., it rejects the *process purity* assumption that equates tasks with processes). Thus, without a clear experimental demonstration of controlled influences dominating responding with MC final tests, Alamri and Higham's (2022) dual-process model may not be a suitable framework for retrieval practice effects in this paradigm.

Overall, our results showed that controlled influences of retrieval practice do dominate responding with related items on MC final tests, but only in specific circumstances.

Experiment 1, which was conducted online and used SAT materials, showed only automatic influences. That is, participants scored worse on related (vs. new) questions and the poor performance with related questions was largely due to selecting the corrective feedback. This finding, coupled with even worse performance on related items when repeated items were included on the final test, suggests that unchecked automatic influences leading to false recognition of related items was the source of the automatic influence, just as it was in Alamri and Higham's (2022, 2024) earlier work. In other words, Experiment 1 replicated the finding that only automatic influences were at play when both tests are MC.

Experiments 2 and 3, which were conducted in a university classroom, revealed a very different pattern of results. In both experiments, evidence of controlled influences was obtained; that is, related questions were answered better than new ones, the same pattern observed repeatedly with CR final tests (e.g., Alamri & Higham, 2022; Little & Bjork, 2015; Little et al., 2012, 2019; Sparck et al., 2016). However, in Experiment 4 which also was conducted in a real classroom, we observed a similar pattern as Experiment 1 with the related-separated pairs: those questions were answered worse than new questions providing evidence once again of automatic influences dominating responding. The fundamental question, then, is what factor(s) caused the difference between Experiments 1 and 4 on the one hand, and Experiments 2 and 3 on the other?⁸

⁸ We rescored the data from Experiments 1-4 using traditional number (or proportion) right scores (i.e., correct answer in top rank position = 1; correct answer in any other rank = 0) instead of ranking scores (scores of 3, 2, 1, 0 for the correct answer in ranks 1, 2, 3, and 4, respectively). Given that number-right scoring is more common in educational contexts, it was a pertinent test to conduct. For the most part, the results were the same. The only exceptions were that the significant interaction between repetition type and question type from the ANOVA in Experiment 1 was no longer significant, and the *p* value from the separated versus new comparison in Experiment 4 increased from .001 to .05. Most likely, these minor differences were attributable to differential sensitivity of the scoring methods; although more common, number/proportion right scoring is less sensitive than rank scoring to partial knowledge (knowledge that might lead participants to rank an option high but not in the top position; e.g., see Ben-Simon et al., 1997).

In our view, the main difference was that Experiments 1 and 4 had repeated items included in the final test whereas the tests in Experiments 2 and 3 did not. The removal of repeated items likely meant that participants were less cautious about countering automatic feelings of familiarity with controlled processes, thereby increasing false recognition of related items. When repeated questions are included, they motivate participants to search for them on the final test. If participants recognize repeated questions and also remember the corrective feedback, it provides a quick and easy method of increasing their test scores. A problem occurs, of course, if related questions are falsely recognized as repeated. In our view, removing repeated items reduced participants' tendency to search for them, thereby encouraging participants to *solve* related questions rather than trying to *remember* their solution (Jacoby, 1978).⁹

One problem with this explanation is that even when there were no repeated items in the test list in Experiment 1 (i.e., in the no-repetition groups), participants were still more likely to endorse the corrective feedback with related questions than with new questions. In Experiments 2 and 3, on the other hand, the absence of repeated items was associated with related-item performance that was superior to that with new items. In our view, these differences can be attributed to the different contexts in which Experiment 1 on the one hand and Experiments 2 and 3 on the other were conducted. Specifically, Experiments 2 and 3 were administered in a formal educational setting whereas the tests in Experiment 1 were conducted online (MTurk), and the formal context likely raised the stakes for achieving high marks. At first blush, critics may question this assessment because the tests in Experiments 2 and 3 were also administered online, just as in Experiment 1, and they were formative, not

⁹ Little, Bjork, and colleagues (e.g., Little et al., 2012) typically ordered their final test items so that repeated items were not presented until after the related items. Naturally, such ordering effectively renders their testing conditions analogous to the no-repetition conditions of the current study. Moreover, such a design may have minimised automatic influences on related items associated with false recognition and may be another reason (in addition to only using CR final tests) that they tended to only observe controlled influences dominating responding in their studies.

counting toward students' final marks. However, there was evidence that students were taking the tests seriously and wanted to score well. For example, there were high participation rates on the tests, and some students clearly wanted more forewarning that the tests were to be administered. Also, there was good reason for students to take the tests seriously. In Experiments 2 and 4, for example, the test was good practice for the upcoming final summative exam, which was also MC. In Experiment 3, the test provided an opportunity for students to test their knowledge of course material in the absence of a formative final exam which was cancelled due to coronavirus.

Thus, we believe that when there are no repeated items included in the tests and students consider the tests high stakes as in Experiments 2 and 3, there is a tendency for controlled processes to overshadow automatic ones. Participants may have read questions more carefully and covertly compared them to earlier questions, more effort may have been expended at retrieving information, and candidate responses to the questions may have been metacognitively monitored more stringently before being offered as answers. Together, these processes likely reduced the allure of the corrective feedback and allowed participants to benefit from, rather than be seduced by the similarity between the related questions.

Benefitting from, rather than being undermined by the similarity of the related questions was particularly evident on related-back-to-back questions in Experiments 2 and 4. These final-test questions were answered well, and the corrective feedback was virtually never endorsed on related final-test questions. The retention interval between the related-back-to-back questions was at a minimum, so participants were in an ideal position to consciously identify discrepancies between the questions and keep automatic influences in check. Change detection also likely promoted deeper understanding by directing students' attention to the key point(s) the two questions were querying. Indeed, one student commented after the tutorial that presenting the related questions back-to-back was a good learning tool

because it helped him understand distinctions that he would have otherwise glossed over. Given these results, future research might investigate the back-to-back method further as a potential way to enhance the benefits of retrieval practice.

Although the back-to-back method was beneficial, it was not a necessary ingredient to obtain controlled influences on MC final tests. In both Experiments 2 and 3, related questions that were separated by several other items also showed controlled influences. Furthermore, it was not necessary to provide feedback on final-test questions either, which would have promoted change detection. No final-test feedback was presented in Experiment 3, and dominance of controlled influences was still observed. The suggestion from these data is that, as long as there are no repeated items included in the final test, the context is right and examinees consider the test to be important, controlled influences are fairly robust, even if both tests are MC.

Base-Rate Considerations

One potential design issue with Experiments 2 and 4 is that students may have been operating with different base rates in the back-to-back and separated conditions. Because the content of the first related question and the response made to it was likely fully accessible in working memory when answering the second related question in the back-to-back condition, participants may have responded to the second question after rejecting the feedback to the first question. That would mean the chance of guessing correctly would effectively increase from .25 (four options) to .33 (three options). Thus, participants' performance on the second related question may have been enhanced in the back-to-back condition relative to the separated condition because the chances of correctly guessing the correct answer were higher.

To address this point, we computed the predicted performance increase in the back-to-back condition and compared it to the separated condition assuming that participants'

knowledge was the same (i.e., controlled influences were not greater in the back-to-back condition), but the number of options was reduced. Specifically, we assumed there were effectively four options in the separated condition compared to only three in the back-to-back condition. We then computed knowledge (k) from the observed score (O) in the separated condition and then used the k value to compute the expected observed score in the back-to-back condition assuming constant knowledge but fewer response options. This computation was achieved using the standard guessing model: $O = k + (1-k)/n$, where n = the number of alternatives (Bereby-Meyer et al., 2002; Diamond & Evans, 1973). Also, instead of using ranking data, we used proportion right (i.e., correct answer in top rank = 1; correct answer in any other position = 0) so that the standard guessing model would apply.

In Experiment 2, the observed final test accuracy in the separated condition was .62, comprised of $k = .49$ plus the probability of guessing the correct answer if the answer was not known (i.e., $[1-k]/n = .51/4 = .13$). If participants had the same level of knowledge in the back-to-back condition ($k = .49$) but there were only three options to consider instead of four, their observed score would be equal to $.49 + (.51/3) = .66$. However, the observed performance in the back-to-back condition was .69. In Experiment 4, the observed final test accuracy in the separated condition was .59, comprised of $k = .45$ plus the probability of guessing the correct answer if the answer was not known (i.e., $[1-k]/n = .55/4 = .14$). If participants had the same level of knowledge in the back-to-back condition as in the separated condition ($k = .45$), their observed score would be equal to $.45 + (.55/3) = .63$. However, the observed performance in the back-to-back condition of Experiment 2 was .69. Thus, in both experiments (but particularly Experiment 4), the increase in back-to-back performance compared to the separated condition could not be fully explained by an increased likelihood of guessing the correct answer due to rejection of the feedback option.

Conclusions

Overall, the current results coupled with previous research show that retrieval practice can produce both automatic and controlled influences on later tests, and that both types of influence occur in both MC and CR tests to varying degrees. CR tests tend to tap the controlled processes better than MC, but automatic influences occur with CR as well (Alamri & Higham, 2022). MC tests are a riskier option because if two questions are similar to each other but have different correct responses, participants may falsely recognize the second question and respond with the corrective feedback (which is now wrong). This problem is particularly evident if some questions are repeated on the test (Experiment 1 and 4) or if corrective feedback is provided on the initial test (Alamri & Higham, 2024). On the other hand, if automatic influences are limited by, for example, presenting related questions back-to-back (Experiment 2 and 4), or removing corrective feedback (Alamri & Higham, 2024), then these problems are less critical. These results support Alamri and Higham's (2022) dual-process framework of retrieval practice effects and allow educators to breathe a sigh of relief.

There is still more work to be done, but overall, research on this topic is suggesting that retrieval practice can be beneficial in many scenarios and with a variety of different tests, but particular caution should be exerted if the final test is MC. Some of the questions that need answering include: (a) How do retention intervals between practice and final tests moderate controlled and automatic influences? Would automatic and/or controlled influences continue to exert effects on performance over longer intervals such as those seen in typical classrooms? (b) What are the similarity dimensions that seduce examinees into believing that related questions are repeated? (c) Are there MC formats that can be used at a test that promote controlled influences? For example, would formats that require participants to thoroughly consider all the options tend to promote controlled influences, such as elimination testing (Little et al., 2019) or the need to assign probabilities to all the alternatives? (d) What is the practical relevance of the effects we have observed in the current experiments (e.g.,

how would students' marks be affected, if at all)? (e) Are some students more inclined toward automatic versus controlled influences and can anything be done to change that? These are avenues for future research. For now, at least, we can rest assured that MC final tests are not always a bad thing as long as precautions are taken.

References

- Alamri, A., & Higham, P. A. (2022). The dark side of corrective feedback: controlled and automatic influences of retrieval practice. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. <https://doi.org/10.1037/xlm0001138>
- Alamri, A., & Higham, P. A. (2024). *Automatic influences of retrieval practice: The role of feedback, false recognition, and opposition instructions*. Manuscript in preparation.
- Ben-Simon, A., Budescu, D. V., & Nevo, B. (1997). A comparative study of measures of partial knowledge in multiple-choice tests. *Applied Psychological Measurement, 21*(1), 65–88. <http://dx.doi.org/10.1177/0146621697211006>
- Bereby-Meyer, Y., Meyer, J., & Flascher, O. M. (2002). Prospect theory analysis of guessing in multiple choice tests. *Journal of Behavioral Decision Making, 15*(4), 313. <https://doi.org/10.1002/bdm.417>
- Butler, A. C., & Roediger, H. L. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory and Cognition, 36*(3), 604–616. <https://doi.org/10.3758/MC.36.3.604>
- Debner, J. A., & Jacoby, L. L. (1994). Unconscious perception: Attention, awareness, and control. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*(2), 304–317. <https://doi.org/10.1037/0278-7393.20.2.304>
- Diamond, J., & Evans, W. (1973). The correction for guessing. *Review of Educational Research, 43*(2), 181–191. <https://doi.org/10.3102/00346543043002181>
- 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.

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>

Hovland, C. I., & Weiss, W. (1951). The influence of source credibility on communication effectiveness. *Public Opinion Quarterly*, *15*(4), 635. <https://doi.org/10.1086/266350>

Jacoby, L. L. (1978). On interpreting the effects of repetition: Solving a problem versus remembering a solution. *Journal of Verbal Learning and Verbal Behavior*, *17*(6), 649–667. [https://doi.org/10.1016/S0022-5371\(78\)90393-6](https://doi.org/10.1016/S0022-5371(78)90393-6)

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](https://doi.org/10.1016/0749-596X(91)90025-F)

Jacoby, L. L., & Brooks, L. R. (1984). Nonanalytic cognition: Memory, perception, and concept learning. In *Psychology of Learning and Motivation* (Vol. 18, pp. 1–47). Elsevier. [https://doi.org/10.1016/S0079-7421\(08\)60358-8](https://doi.org/10.1016/S0079-7421(08)60358-8)

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., Lindsay, D. S., & Hessels, S. (2003). Item-specific control of automatic processes: Stroop process dissociations. *Psychonomic Bulletin & Review*, *10*(3), 638–644. <https://doi.org/10.3758/BF03196526>

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.

- Little, J. L., & Bjork, E. L. (2012). *The persisting benefits of using multiple-choice tests as learning events*. [Conference presentation]. 34th Annual Conference of the Cognitive Science Society. Sapporo, Japan.
- Little, J. L., & Bjork, E. L. (2015). Optimizing multiple-choice tests as tools for learning. *Memory and Cognition*, *43*(1), 14–26. <https://doi.org/10.3758/s13421-014-0452-8>
- Little, J. L., Bjork, E. L., Bjork, R. A., & Angello, G. (2012). Multiple-choice tests exonerated, at least of some charges: Fostering test-induced learning and avoiding test-induced forgetting. *Psychological Science*, *23*(11), 1337–1344.
<https://doi.org/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. <https://doi.org/10.1037/xlm0000638>
- Mandler, G. (1980). Recognizing: The judgment of previous occurrence. *Psychological Review*, *87*(3), 252–271. <https://doi.org/10.1037/0033-295X.87.3.252>
- 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>
- 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. I., & 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>
- 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>
- Rauschert, E. S. J., Yang, S., & Pigg, R. M. (2019). Which of the following is true: We can write better multiple choice questions. *Bulletin of the Ecological Society of America, 100*(1), 1–7.
- 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.org/10.1037/0278-7393.31.5.1155>
- 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), 1–10. <https://doi.org/10.1186/s41235-016-0003-x>
- 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*Summary of the Experimental Details in Experiments 1-4*

	Experiment Number			
	1	2	3	4
Context	MTurk	Educational	Educational	Educational
Participants	138 online Participants	164 students	223 students	224 students
Material	SAT	Introductory psychology module	Introductory psychology module	Introductory psychology module
Design	2 x 2 x 2 mixed factorial	One independent variable (3 levels)	2 x 2 mixed factorial	One independent variable (4 levels)
Independent variables	Test type (two tests, one test) x repetition type (repetition, no repetition) x question type (new, related)	Question type (related- separated, related-back-to- back, new)	Feedback type (feedback, no feedback) x question type (new, related)	Question type (repeated, related- separated, related-back- to-back, new)

Table 2

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

Answer type on the final test	Two tests	One test	Two tests	One test
	Repetition	Repetition	No-Repetition	No-Repetition
Correct				
Previous answer	0.10 (.09)	0.09 (.10)	0.19 (.19)	0.14 (.13)
Other	0.19 (.18)	0.19 (.18)	0.22 (.17)	0.24 (.15)
Incorrect				
Previous answer	0.10 (.12)	0.14 (.13)	0.12 (.13)	0.15 (.13)
Corrective feedback	0.46 (.25)	0.42 (.30)	0.23 (.25)	0.23 (.17)
Other	0.16 (.14)	0.16 (.17)	0.24 (.23)	0.23 (.19)

Table 3

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

Answer type on the final test	Back-to-back	Separated
Correct		
Previous answer	0.31 (.24)	0.28 (.23)
Other	0.29 (.27)	0.31 (.26)
Incorrect		
Previous answer	0.19 (.24)	0.13 (.16)
Corrective feedback	0.04 (.09)	0.12 (.18)
Other	0.17 (.19)	0.17 (.20)

Table 4

Mean (SD) Proportion of Final-Test Answer Types to Related Questions Conditioned on Being Answered Incorrectly in The First Test in Experiment 3

Answer type on the final test	Feedback	No feedback
Correct		
Previous answer	0.26 (.18)	0.23 (.16)
Other	0.24 (.18)	0.24 (.19)
Incorrect		
Previous answer	0.16 (.15)	0.13 (.16)
Corrective feedback	0.14 (.17)	0.18 (.16)
Other	0.20 (.16)	0.23 (.19)

Table 5

Mean (SD) Proportion of Final-Test Answer Types to Related Questions Conditioned on Being Answered Incorrectly in The First Test in Experiment 4

Answer type on the final test	Back-to-back	Separated
Correct		
Previous answer	0.35 (.28)	0.23 (.24)
Other	0.23 (.23)	0.19 (.21)
Incorrect		
Previous answer	0.17 (.23)	0.12 (.20)
Corrective feedback	0.02 (.07)	0.29 (.33)
Other	0.22 (.23)	0.17 (.21)

Figure 1

Schematic Illustrating the Design Used in Experiment 1

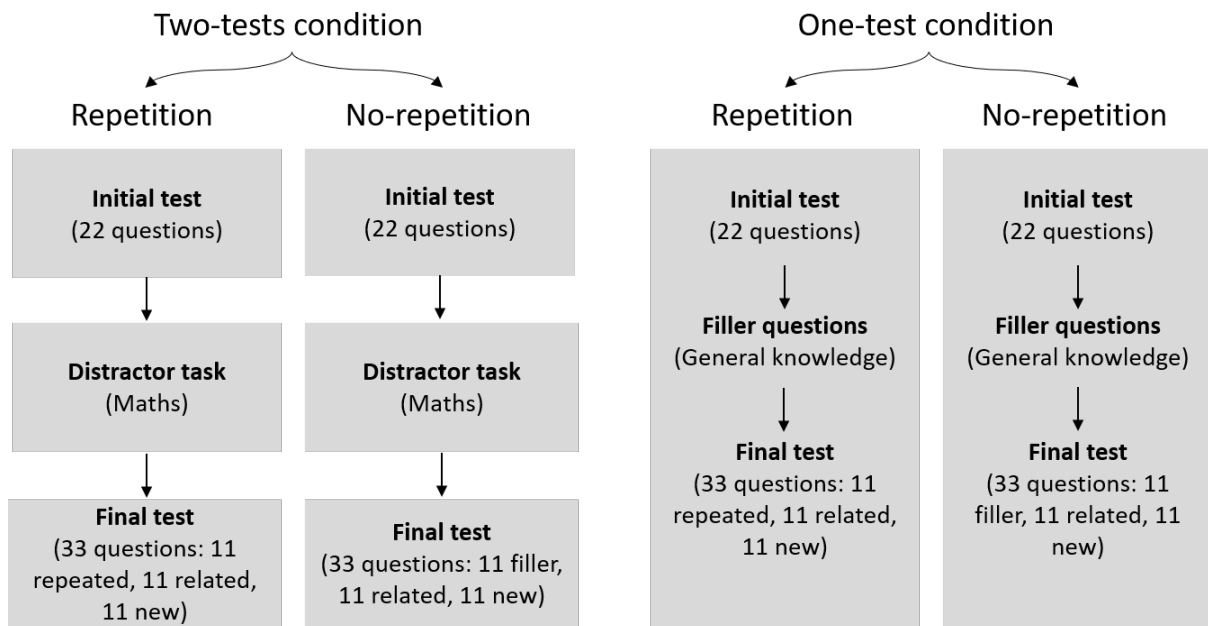
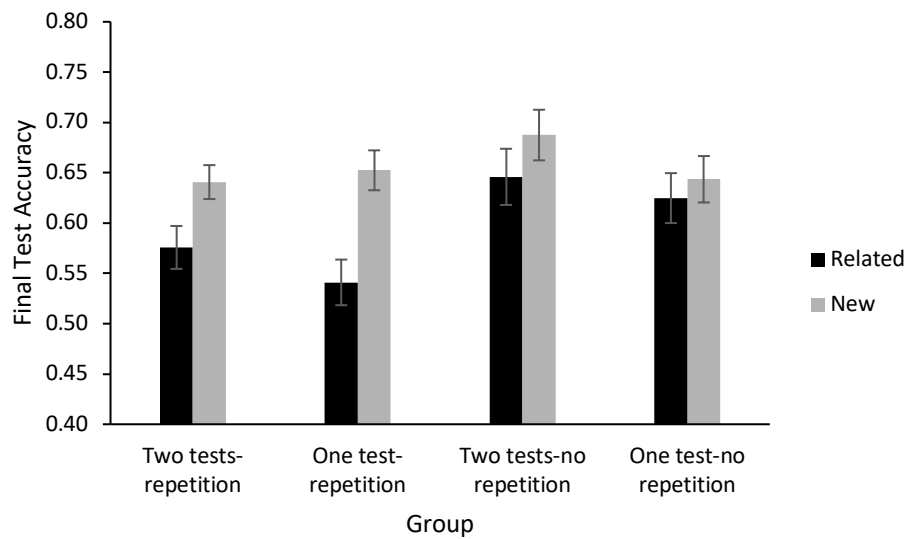


Figure 2

Mean Final-Test Accuracy for Each Question Type Broken Down by Test Type and Repetition Type in Experiment 1



Note: Final-test accuracy was defined as the mean proportion scored out of three possible points per question. Error bars indicate standard errors of the mean.

Figure 3

Schematic Illustrating the Five Final-Test Answer Types (Related Items) Conditioned on an Incorrect Initial-Test Response

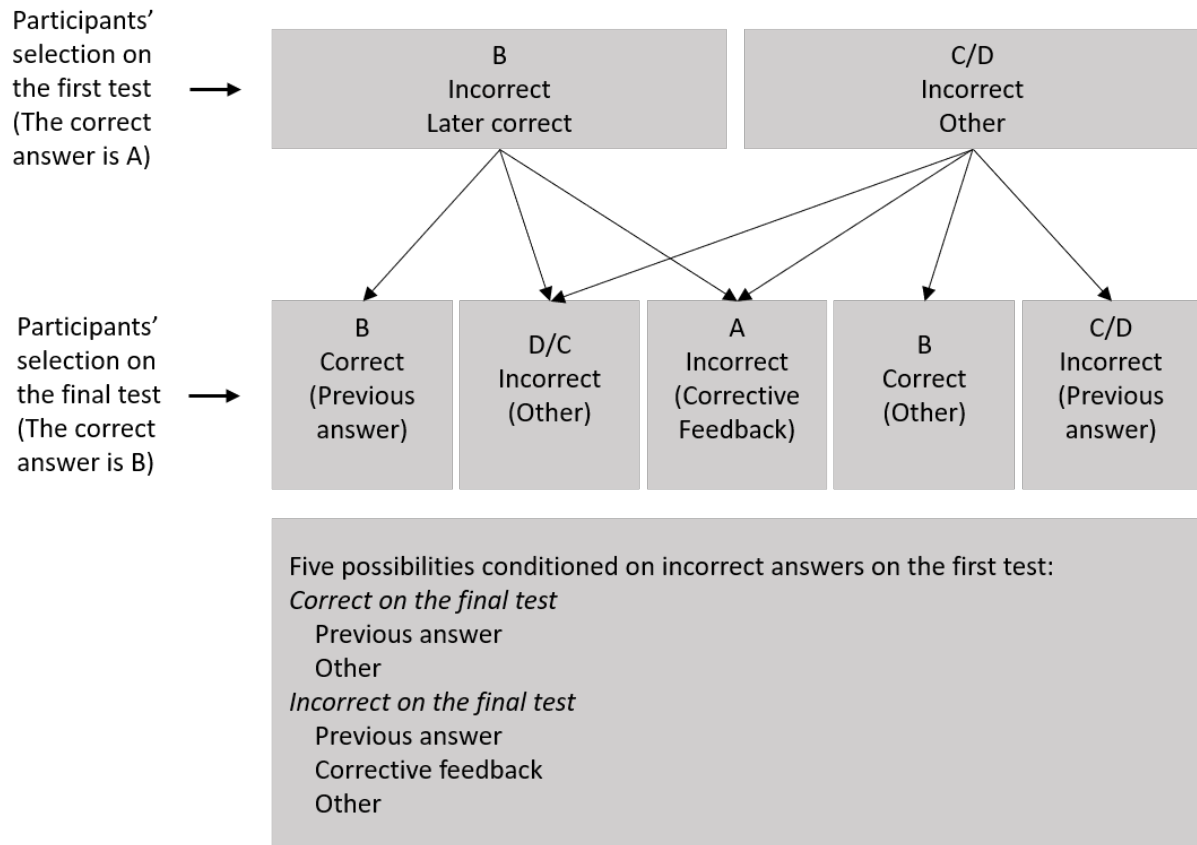
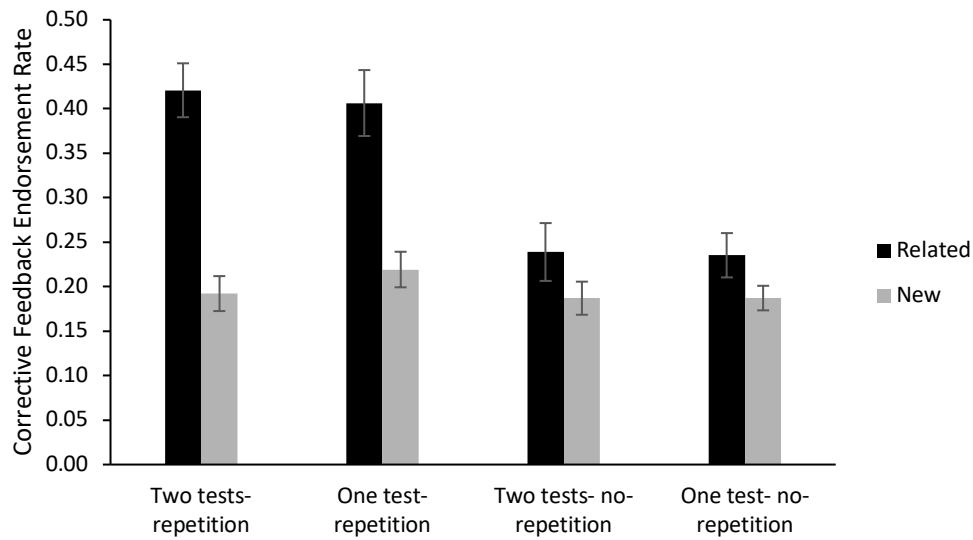


Figure 4

Mean Corrective Feedback Endorsement Rate for Each Question Type Broken Down by Test Type and Repetition Type in Experiment 1



Note: Error bars indicate standard errors of the mean.

Figure 5

Schematics Illustrating the Designs Used in Experiments 2 and 4

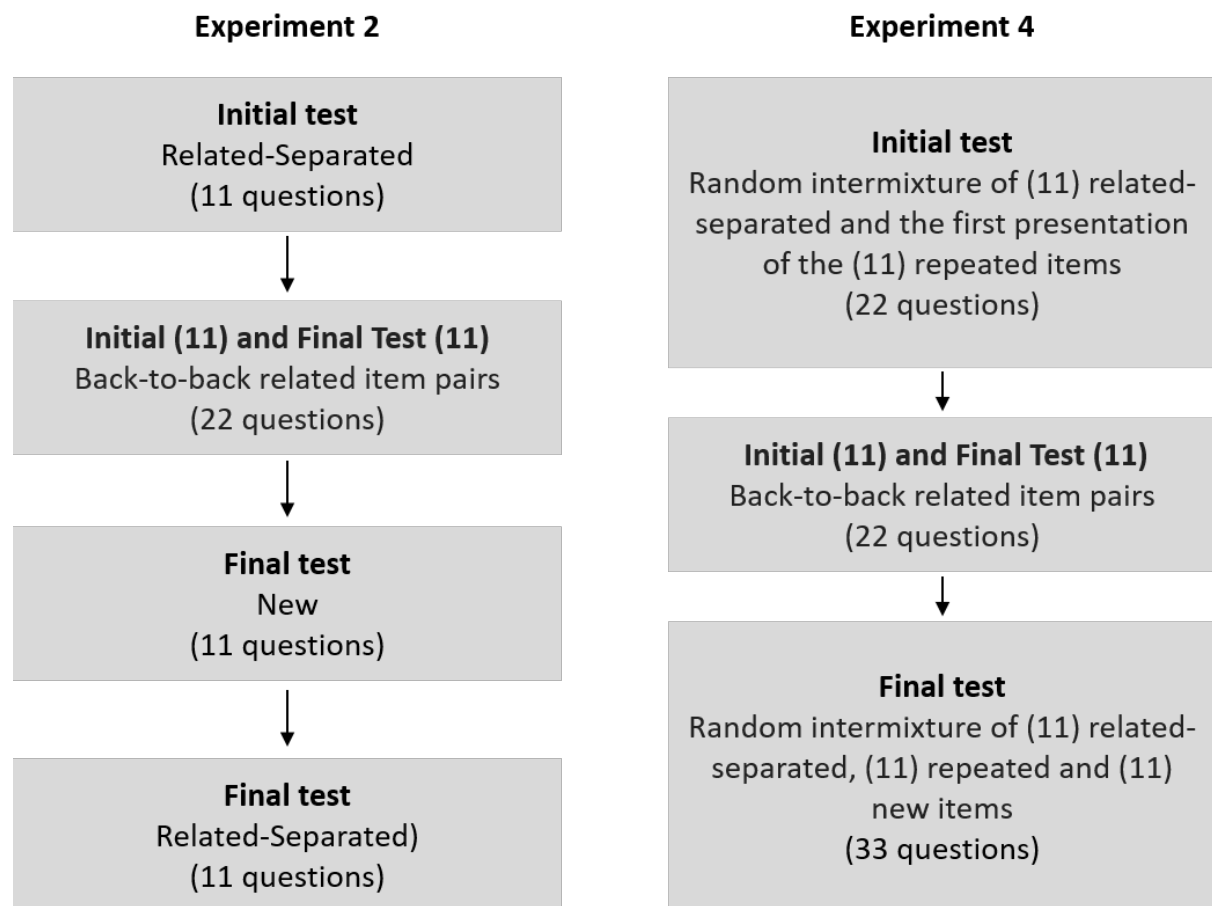
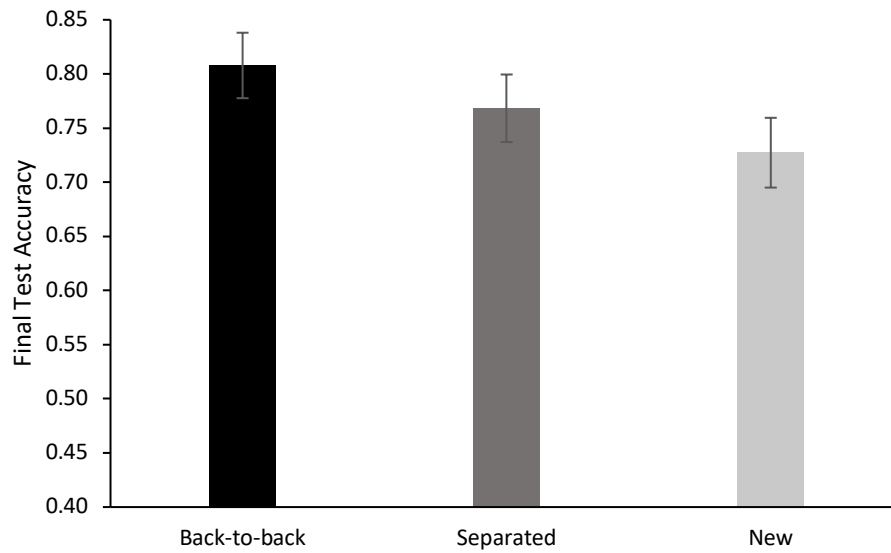


Figure 6

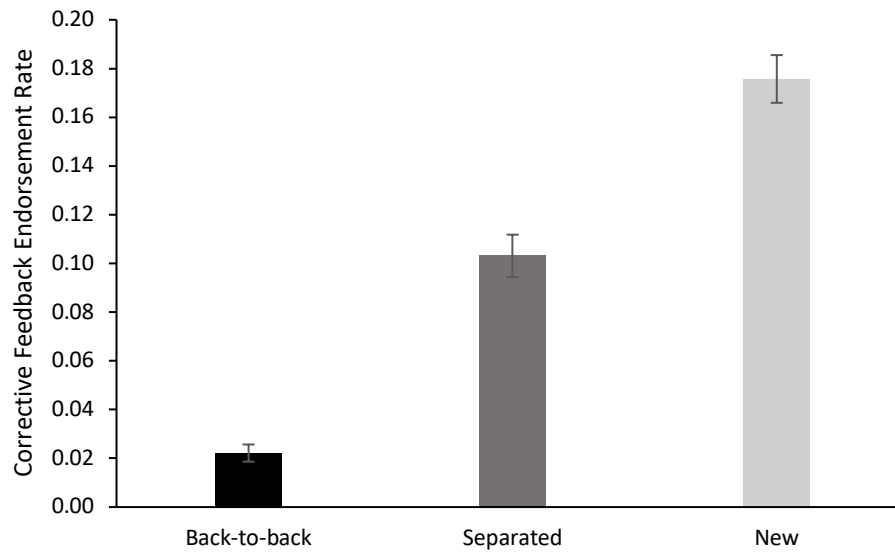
Mean Final-Test Accuracy for the Related-Back-to-Back, Related-Separated, and New Question Types in Experiment 2



Note: Final-test accuracy was defined as the mean proportion scored out of three possible points per question. Error bars indicate standard errors of the mean. All items in this Experiment were related or new items.

Figure 7

Mean Corrective Feedback Endorsement Rate for the Related-Back-to-Back, Related-Separated, and New Question Types on the Final Test in Experiment 2



Note: Error bars indicate standard errors of the mean.

Figure 8

Schematic Illustrating the Design Used in Experiment 3

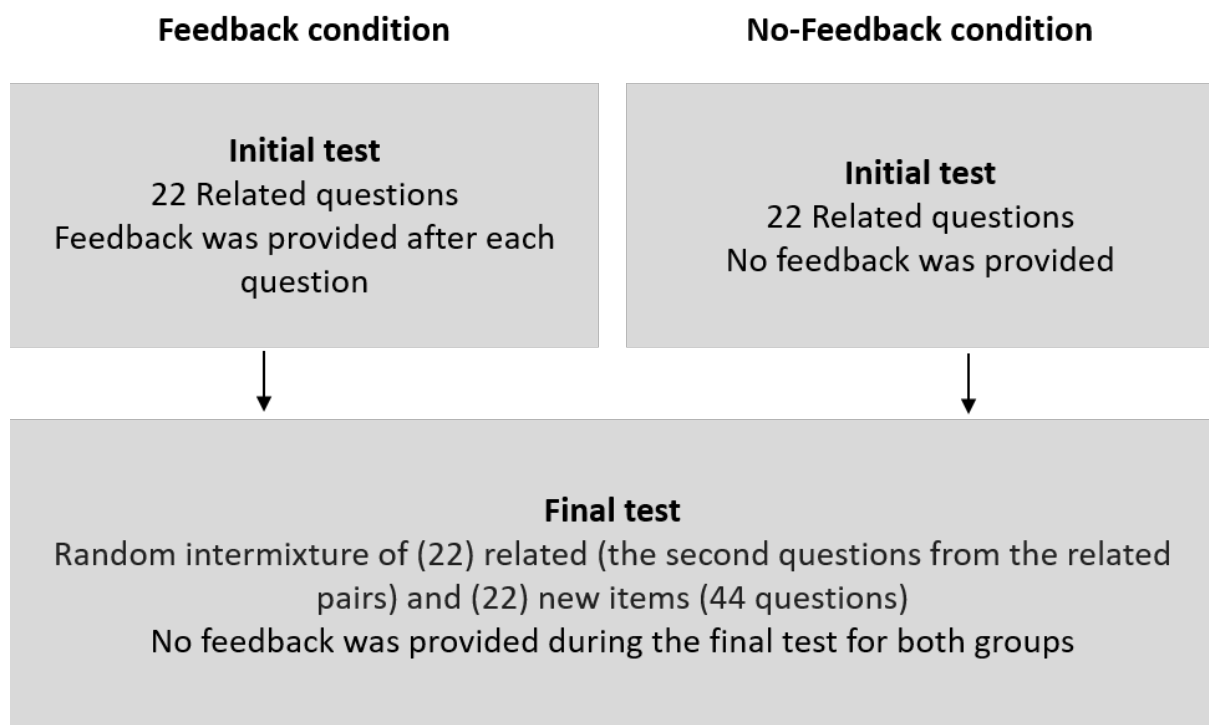
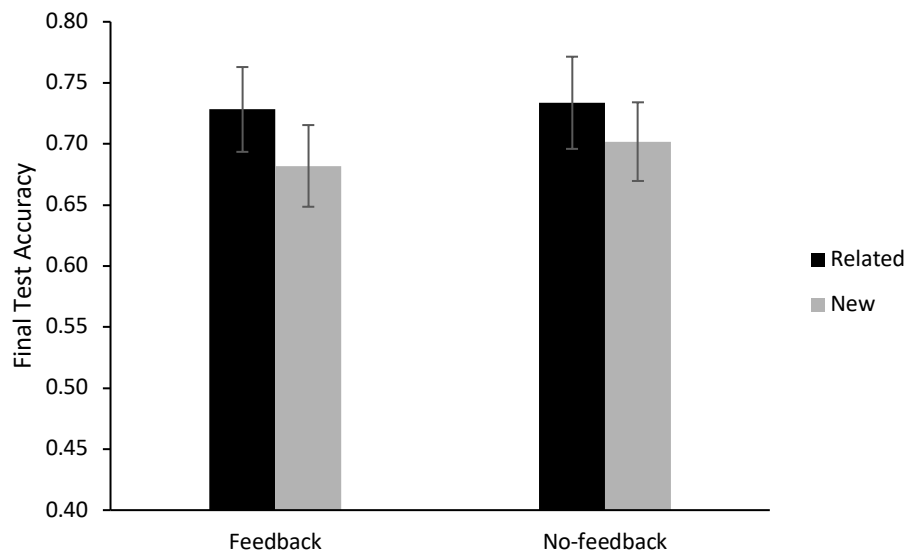


Figure 9

Mean Final-Test Accuracy for Each Question Type Broken Down by Feedback Type in Experiment 3

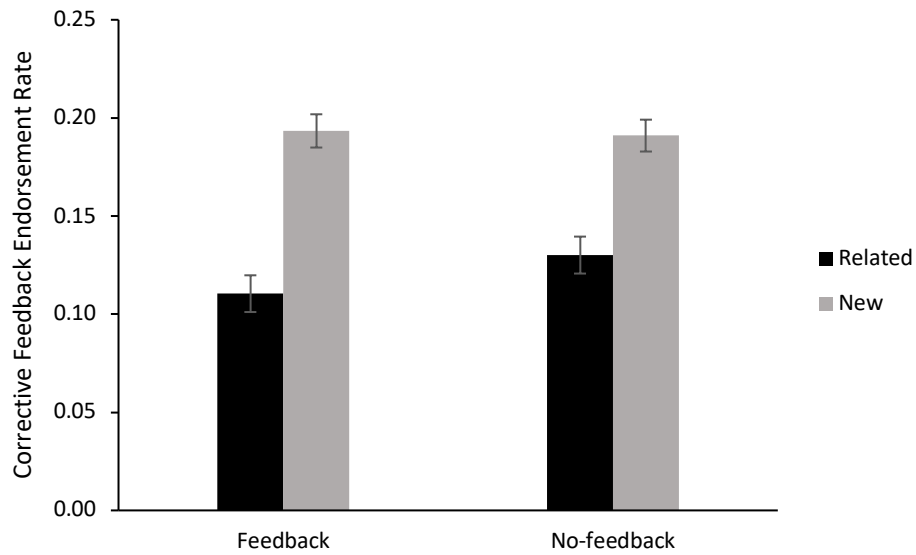


Note: Final-test accuracy was defined as the mean proportion scored out of three possible points per question. Error bars indicate standard errors of the mean.

Figure 10

Mean Corrective Feedback Endorsement Rate for Each Question Type on the Final Test

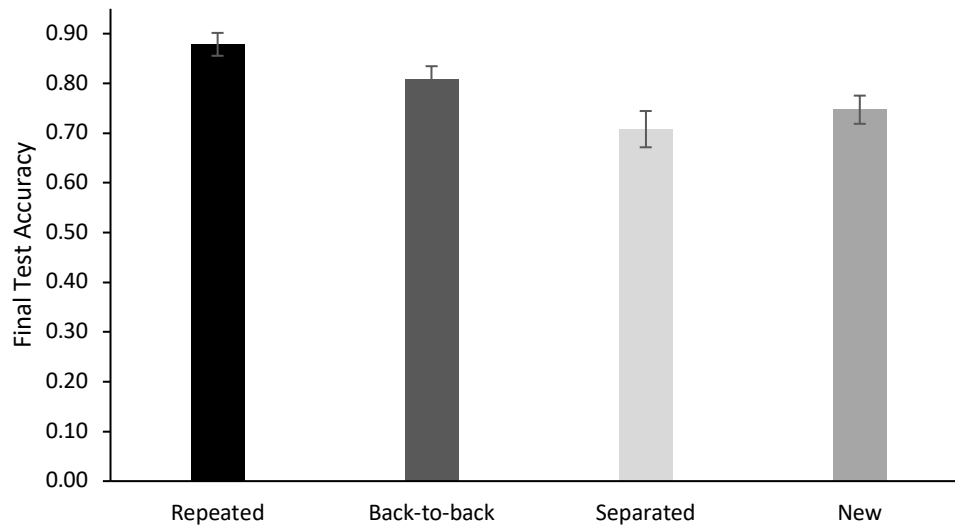
Broken Down by Feedback Type in Experiment 3



Note: Error bars indicate standard errors of the mean.

Figure 11

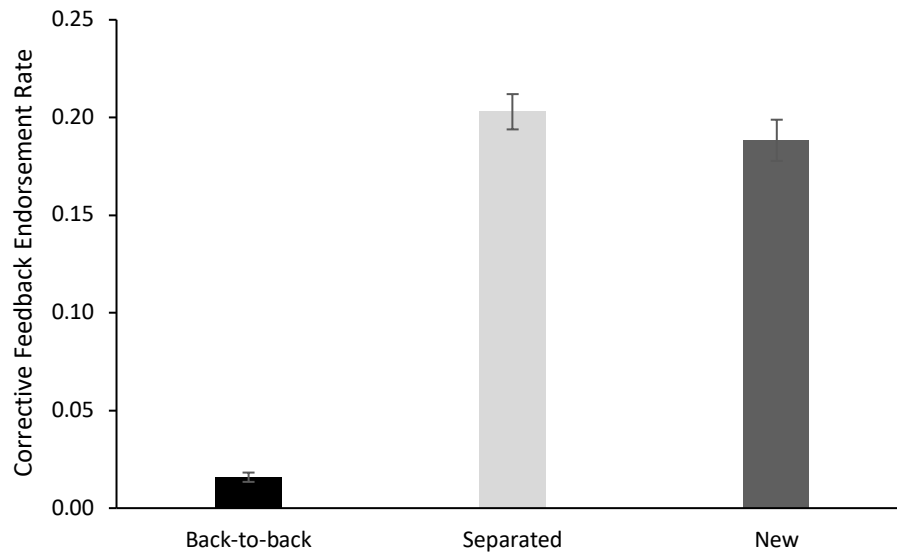
Mean Final-Test Accuracy for the Repeated, Related-Back-to-Back, Related-Separated, and New Question Types in Experiment 4



Note: Final-test accuracy was defined as the mean proportion scored out of three possible points per question. Error bars indicate standard errors of the mean.

Figure 12

Mean Corrective Feedback Endorsement Rate for the Related-Back-to-Back, Related-Separated, and New Question Types on the Final Test in Experiment 4



Note: Error bars indicate standard errors of the mean.