University of Southampton

Faculty of Environmental and Life Sciences

School of Psychology

Controlled and Automatic Influences of Multiple Choice Testing

by

Aeshah Alamri

Thesis for the degree of Doctor of Philosophy

July 2022

Table of Contents

List of Accompanying Materials

DOI: https://doi.org/10.5258/SOTON/D2251

Research Thesis: Declaration of Authorship

Print name: Aeshah Alamri

Title of thesis: Controlled and Automatic Influences of Multiple Choice Testing

I declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. Parts of this work have been published as:
   Alamri, A. 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*, *48*(5), 752–768. https://doi.org/10.1037/xlm0001138
   Signature: Date: 25/7/2022

Acknowledgements

﴿And my success is not but through Allah. Upon him I have relied, and to him I return﴾ Quran, Surat Hud, 11:88.

First of all, I would like to praise and thank Allah for giving me the strength to complete this research, considering the difficult time that I and the whole world went through in the last two years due to the pandemic.

Second, I could not have accomplished this thesis without the support I received from many individuals. I would like to take this opportunity to express my appreciation to those who have played an important role in my PhD journey.

Many thanks go to my supervisor Dr Philip Higham for his valuable guidance and continuous support, for which I am very grateful. I would also like to extend my thanks to King Saud University and the Saudi Cultural Bureau in London for their financial support, which enabled me to pursue my research.

I would also like to thank my external examiner Prof Tim Hollins, and my internal examiner Dr Tina Seabrooke for taking the time to read my thesis, I enjoyed discussing it with both of you.

To my family, especially my parents Salehah and Ali, whom I have not met for three long years, I offer my deepest gratitude for their love, tears, and prayers. I would also like to offer my special thanks to my sister, Ibtissam, and my friend Asiyah, for our cheerful talks that encouraged me even in the darkest moments.

Finally, my greatest thanks go to my beloved husband Mohammed, who gave up his job in Saudi Arabia to be with me and support me during my study. I am very indebted to you. Also, to my beautiful son Sultan, please accept my apologies for not always being with you during the last four years. To you both, I dedicate this work.

Introduction

Testing is one of the most common methods used to assess students' learning outcomes. Educators generally use tests as a reliable and feasible way to evaluate students' performance, especially at the end of a course, while students encounter testing either as part of an educational module or as a requirement for admission to an educational institution. Many types of tests have been utilized to assess learning performance, including open-ended, short answer, true/false and multiple-choice (MC), to name a few. Although researchers have reported that each type of test might assess different kinds of learning and cognitive processes (e.g., Ozuru et al., 2013; Polat, 2020; Smith & Karpicke, 2014), MC questions are considered a popular format that can be used for multiple purposes.

The development of MC questions can be traced to Fredrick J. Kelly in 1916, when he designed a test with questions that required students to circle the correct answer rather than generate an answer. Kelly provided the first guidelines for producing MC questions; for example, each question should test a single problem whose answer is either fully wrong or right, and all students should be able to understand the questions in the same way, so questions should not be phrased ambiguously (Gierl et al., 2017; see Haladyna, 2004 for information about designing MC questions). Since then, MC questions have been used widely to accomplish different goals due to their advantages, including the fact that they are easy to mark, students can answer many questions in a short time, and more than one topic can be covered in an exam. In addition to these advantages, however, many disadvantages are associated with MC testing; these include the fact that creating the test can consume more time than creating tests in other formats (e.g., open-ended questions); generating good questions, especially for high cognitive levels, can be

difficult; and since students can guess the correct answer, the test might not accurately reflect a learner's level of knowledge (Dubins et al., 2016; Fellenz, 2004).

Positive Testing Effect

Despite the possible negatives associated with using MC questions as an assessment method, this type of testing can be considered the most common format used in different grades and courses and for different types of exams, including educational and international tests (Gierl et al., 2017). Less familiar, however, is the use of these questions as a learning tool to promote retention, which is widely known as the testing effect (for a review of the testing effect, see Roediger & Karpicke, 2006). Previous literature has reported that taking tests on to-be-learned material has a positive effect on learning compared with other learning techniques, such as study or summarisation, whether or not feedback is provided (see Dunlosky et al., 2013 for a review of learning techniques). Typically, studies that examine the testing effect involve participants reading for a certain amount of time before taking an initial test. After a short delay, participants take a final test, which consists of repeated questions as well as new questions that act as control questions. Originally, researchers focused on comparing the performance of an initially tested group with another group that did not take an initial test or sometimes engaged in a filler task. Later, to equate the exposure time of the groups, scholars started adding a different control group, one that was asked to study or restudy the same material, and then compared the performance of this control group to the tested group's performance (Butler & Roediger, 2007). In either case, prior studies showed the benefits of using MC testing as a learning tool that facilitated performance on the final test, which is usually presented in a cued recall (CR) format (Fazio et al., 2010; for a review of the consequences of MC testing, see Marsh et al., 2007).

One of the earliest studies involving initial MC testing was a study by Spitzer (1939) examining the effect of MC testing on later performance after long and short delays. Two groups of participants read two texts about peanuts and bamboo. One of the two groups took an initial MC test with no feedback provided, while the other group engaged only in the reading session without being initially tested (the control group). Then, both groups took the final CR test at different intervals: immediately or after one day, seven days, 14 days, 21 days, one month, or two months. The findings showed that the initially tested group outperformed the control group when they took the final test 1, 7, 14 and 21 days after the initial test. However, similar results were not observed in the one- and two-month interval conditions. Spitzer concluded that as long as the final test was taken no longer than three weeks after the initial test, an initial test with MC questions could serve as a powerful learning tool.

In a later study, Nungester and Duchastel (1982) compared the consequences of taking an initial MC test to those of studying the material or working on a filler task. After the reading session, participants either answered CR/MC initial questions with no feedback, studied the material for an equal amount of time, or answered filler questions. After two weeks, participants returned to take the final test. The researchers modified the test format for the tested group by changing the MC and CR questions in the initial test to CR and MC formats, respectively. The findings demonstrated a great enhancement in the tested group's performance compared to the other groups, regardless of the initial test format. This suggests that taking an initial MC test can produce better retention than merely studying the material, resulting in a better later performance.

Most studies investigating the testing effect have employed passage-reading sessions before participants take the initial test (e.g., Nungester & Duchastel, 1982;

Spitzer, 1939). Other studies, however, have examined the MC testing effect using general knowledge questions without a prior reading session and have reported similar benefits of initial MC testing. For example, Butler et al. (2008) had participants answer general knowledge MC questions with corrective feedback being provided for half of the questions, while no feedback was provided for the other half. After completing a filler task, participants took the final CR test, which contained some previously tested items as well as new (control) items. The findings demonstrated enhanced performance on the tested items compared to the new items, regardless of whether feedback was provided during the first test. The best performance, however, was observed in the feedback condition, suggesting that following MC questions with corrective feedback produces larger benefits for later performance compared to when no feedback is provided.

The benefit of utilizing initial MC testing has been reported not only in lab-based experiments but also in an educational context using authentic educational materials. For example, McDaniel et al. (2012) had college students take either a weekly MC/CR quiz with corrective feedback or read the information presented in the quiz as statements. Another group of students who were not involved in any activity later acted as a control group. After three weeks, students took the final exam, which was presented in MC format. The results indicated that, regardless of the initial test format, taking an initial test enhanced students' performance on the final exam compared to the reading statements and the control conditions; that is, taking an MC test was not less effective than taking a CR test in terms of promoting retention in the final test. Similar benefits were reported with different grades, such as high school (e.g., McDermott et al., 2014), middle school (e.g., McDaniel et al., 2013) and elementary school (e.g., Marsh et al., 2012).

Prior studies have investigated the difference in the size of the testing effect when the initial test is an MC test compared to a CR test; some studies reported that CR tests had more benefits than MC tests. For example, Butler and Roediger (2007) had participants first watch lectures via a video recording and then either study a summary of the lectures or take an initial MC or CR test with corrective feedback on half of the questions and no corrective feedback on the other half. Another group only watched the lecture with no subsequent initial test or study (the control group). In the final CR test, which took place after a month, both the tested and study groups outperformed the control group regardless of whether or not feedback was provided. Surprisingly, no difference was observed between the MC group and the study group. The CR group's performance on the final test was significantly better than that of participants in the MC and study conditions. The researchers suggested that answering CR questions required participants to produce responses, which might involve more retrieval processes than MC questions require. On the one hand, these results are consistent with other studies that reported enhancement in the CR group over the MC and study groups (e.g., Kang et al., 2007). On the other hand, the results did not replicate the majority of the research, which has shown that testing is a more powerful learning tool than any other learning technique, such as studying (e.g., Dunlosky et al., 2013). Butler and Roediger suggested that the summary that participants in the study condition were provided included all the critical facts that they had to know from the lecture, which was enough to enhance the performance of the study group to reach that of the MC group.

Similarly, a study by Little and Bjork (2012) investigated the different effects produced by initial MC and CR tests, however, their findings contrasted with those of Butler and Roediger (2007). After a reading session, participants took either an initial MC

or CR test without feedback. All the MC questions on the initial test were designed with competitive lures[1] that were plausible answers for each question. An example of an MC question with competitive lures is the following: “How many inches long is an average ferret tail? a. 7–10, b. 20, c. 5” (boldface indicates the correct answer). After a five-minute delay, participants took the final CR test, which consisted of some previously tested questions and other questions that were new (control questions). Regardless of the initial test format, the results showed an enhanced performance on the tested items compared to the new items. However, a greater enhancement occurred after the MC initial test than after the CR initial test. These findings are quite surprising in light of prior research suggesting that CR initial testing is more beneficial than MC initial testing (e.g., Butler & Roediger, 2007; Kang et al., 2007). Little and Bjork concluded that MC testing is not necessarily less effective than other test formats, as it can motivate learners to engage in more retrieval processes to reject incorrect alternatives and select the correct ones. Thus, designing these questions with competitive lures can produce a retrieval process similar to or even better than that produced by CR tests.

Indeed, the nature of MC questions was found to be useful in terms of promoting the retrieval of information, as MC questions present the targeted answer as part of the question. For instance, Cantor et al. (2014) investigated whether taking an MC test could activate marginal knowledge and thus enhance final performance. Marginal knowledge is information that a learner might have stored in their memory but to which they have no easy access when it is needed. The researchers selected 84 marginal knowledge questions that learners failed to answer most frequently when they encountered them in a CR format yet answered them correctly when they were presented in an MC format. After

[1] “Lures” is another term for incorrect alternatives.

the researchers determined the materials to be used, participants took an initial MC test, which included half of the 84 marginal knowledge questions, as the other half would serve as control questions in the final test. After different intervals (immediately, five minutes or ten minutes), participants completed the final CR test, consisting of tested questions as well as control questions. The results showed that participants performed better on the tested items than on the control items, despite that no feedback was provided during the initial test. Although performance dropped slightly after the five- and ten-minutes delays, the testing effect of the MC questions remained. These findings suggest that MC questions are particularly effective in activating knowledge that learners may struggle to access by stimulating them to retrieve this information.

In a review of the testing effect of MC questions, Glass and Sinha (2013) concluded that MC questions are powerful learning tools as they enhanced learning for different student populations and with the use of different test formats in the final test. MC testing also enhanced the short- and long-term retention of information, as benefits remained after different delay conditions between the initial and the final tests. In a meta-analysis of 181 separate experiments, Adesope et al. (2017) concluded that, in terms of the testing effect, MC questions were more likely to enhance retention ($g = 0.70$) than CR questions ($g = 0.58$). These findings are consistent with a more recent meta-analysis of the testing effect by Yang et al. (2021), which found that taking an initial MC test was more beneficial for retention ($g = 0.57$) than taking an initial CR test ($g = 0.32$). Together, these results should encourage educators to use MC testing to promote retention, given the ease of marking this type of test.

Negative Testing Effect

Although the majority of previous studies have shown the advantages of using MC testing to promote learning outcomes, other studies have found that, under certain circumstances, MC testing can impair later performance; this is termed the negative testing effect. The literature indicates that when it comes to recognition testing (e.g., MC), exposing learners to incorrect information can lead them to memorise false knowledge as part of the retrieval process. Specifically, when learners encounter MC questions, they are usually exposed to one correct option and three incorrect options (lures). Learners' exposure to lures can increase the likelihood that they will produce one of them on the later test as a correct answer (Meyer, 2011; Roediger & Marsh, 2005).

Roediger and Marsh (2005) examined the positive and negative effects of taking an initial MC test on a final CR test. After reading a passage, participants took the initial MC test with no feedback provided. Half of the questions were based on the passage (studied questions), while the other half were not (unstudied questions). Each MC question had two, four or six alternatives (i.e., the correct answer plus one, three or five lures). Then, participants took the final CR test, which included some tested questions as well as new questions (control questions). The results from the final test demonstrated a positive testing effect, as participants performed better on the tested items compared to the new items. However, the researchers also found a negative testing effect, as participants tended to intrude with lures from questions on the initial test on the final CR test. The negative testing effect was associated with the unstudied questions as well as with a greater number of lures. Thus, although MC testing typically enhanced later retention, it was also a source of errors that may impair later performance.

A later study by Butler and Roediger (2008) replicated Roediger and Marsh's (2005) study and investigated whether providing corrective feedback during the initial test could

reduce the negative testing effect. The procedure was similar to that used in the study by Roediger and Marsh, except that the researchers in this study compared participants' performance based on the presence and absence of corrective feedback during the practice test. In addition, instead of two study conditions (studied and unstudied), the researchers added an additional condition where participants were required to study the passage twice (restudied). The results of the final CR test showed both positive and negative testing effects and indicated that providing corrective feedback resulted in more accurate answers and reduced the probability of participants endorsing lures in the final test (i.e., lure intrusion) compared to the no-feedback condition. The negative testing effect was found to be associated with the no-feedback condition and with spending less time studying. However, when they received corrective feedback, participants managed to differentiate between the correct option and the lures, which prevented them from later selecting the lures as the correct answers. The researchers concluded that providing corrective feedback helped participants to correct their previous errors and keep their previous correct responses, consequently increasing the positive effect of initial MC testing.

Subsequent studies examined the negative testing effect of MC questions with different manipulations and showed similar results. For example, focusing on how long the negative testing effect can persist, Fazio et al. (2010) investigated the negative testing effect when the final CR test was taken immediately and when it was taken after a one-week delay. In both cases, the negative testing effect remained, although it was slightly less in the delay condition. Marsh et al. (2009) conducted a study with undergraduate and

high school students who answered MC questions acquired from SAT subject tests[2] without a prior reading session. Consistent with previous studies, the findings showed a negative testing effect in both groups; however, the performance of the high school participants was worse than that of the undergraduate participants. Also, participants' performance on the final test was correlated with their performance on the initial test. That is, participants who scored low on the initial test showed more lure intrusion than participants who performed better on the initial test. The researchers attributed these findings to the different levels of prior knowledge held by the two groups. Specifically, in the absence of previous knowledge, participants could not distinguish the lures from the correct answers, which resulted in more lure intrusion.

Indeed, Roediger et al. (2010) found that the negative testing effect of MC testing was associated with the number of incorrect answers selected during the initial test. That is, 78% of the incorrect answers on the final test were selected previously on the initial test, which could explain why providing corrective feedback helped to reduce this effect. These findings are consistent with those of Meyer (2011), who examined positive and negative testing effects with younger (18–25) and older (55–60) participants. While the results indicated a positive testing effect in both the younger and older participants, no negative testing effect was found in either group. However, a closer analysis showed a negative testing effect only when participants performed poorly on the initial test. Consistent with Marsh et al. (2009), the results suggest that the negative testing effect is more prominent with participants who lack knowledge, leading them to guess the answers to MC questions. Therefore, when Butler and Roediger (2008) instructed

[2] The SAT is a standardized test that is used widely in the USA to test high school students prior to admission to undergraduate university programs.

participants to answer “I do not know” instead of guessing, participants showed less lure intrusion, which resulted in a reduction in the negative testing effect.

Previous research has demonstrated that MC questions produce a positive testing effect (e.g., McDaniel et al., 2012) and a negative testing effect (e.g., Roediger & Marsh, 2005). However, it should be noted that most of the studies that investigated the negative testing effect of MC questions reported a greater positive than negative effect. Also, studies have shown that the negative testing effect occurred in the absence of corrective feedback or when participants had less study time; nevertheless, a robust positive testing effect was found in both conditions. These results should promote more use of MC questions in testing as taking MC tests can be more useful than not being tested at all, despite the negative testing effect. Providing corrective feedback and having students spend more time studying to-be-learned material can help to overcome the negative consequences of MC testing and increase the benefits.

Positive Effects of MC Testing on Related Items

The positive testing effect of MC questions was observed not only with repeated tested items but also with related untested items. Related items, according to Little, Bjork and colleagues (e.g., Little & Bjork, 2015), are pairs of questions that are related conceptually (e.g., both questions in a given pair are about capitals of European countries) and use a lure from a practice MC question as the correct answer to a final CR question. 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 an initial MC test (boldface indicates the correct answer), followed by, “What is the capital of Finland?” presented on the final CR test. Note that in the related question on the final test, “Oslo” is no longer the correct answer; the correct answer is “Helsinki”, which was one of the lures

on the initial MC question. In several studies, Little, Bjork and colleagues have examined whether initial MC testing could enhance learning performance on related untested items in a way that is similar to what has been reported with tested items.

In one study, Little and Bjork (2015) had participants read an expository text and then take an initial MC test in which some of the questions included competitive lures, while others did not. For example, the competitive version of the lures for the previously mentioned question used in Little and Bjork's 2012 study (see above) is “a. 7–10, b. 20, c. 5”, while in the uncompetitive version, the first two lures were replaced with “1,500” and “3,500”, respectively, which makes the alternatives less plausible. No feedback was provided during the initial test. After a short delay, participants took a final CR test which contained repeated (tested) questions, related (untested) questions, and new (control) questions. The findings indicated that participants performed better on the tested items compared to the new items, no matter what type of lures were used in the initial MC test. However, participants' performance on the related items was better than their performance on the new items only when the questions on the initial test were accompanied by competitive lures. As mentioned earlier, these results showed the importance of using competitive lures with related questions to provoke extensive retrieval processes, while questions with uncompetitive lures were very easy, resulting in fewer benefits from the initial test.

Little and Bjork’s (2015) study provided a different view of the robust testing effect of MC questions, showing that this effect is present not only with repeated items but also with related items. However, note that these findings were the result of the intensive retrieval of information that occurred during the initial test. Indeed, Chan et al. (2006) concluded that the level of retrieval that occurred during the initial test was an important

determinant in participants' later performance on questions with related information. Retrieving information for one question can improve the probability of recalling the correct answer to a related question. Learners might search for all the related information they have about the topic, leading them to correctly answer not only the main question but also the related question. Even with other test formats, research by Hinze and Wiley (2011) found that initial testing that is more likely to provoke extensive retrieval, such as a free-recall format, enhanced participants' performance on novel questions better than a fill-in-the-blank format. Although both tests required participants to generate answers, the free-recall questions required a more detailed answer than the fill-in-the-blank questions, thereby more effortful retrieval. Likewise, answering related MC questions with competitive lures involves more intensive retrieval than is required to answer questions with uncompetitive lures.

Similar results to that obtained by Little and Bjork (2015) were observed in several studies. For example, Little and Bjork (2012; see also Little & Bjork, 2016; Little, 2018) examined whether the enhanced performance on related items would persist when the final test was considerably delayed. Participants took an initial MC test after a reading session without receiving any corrective feedback. All the MC questions presented during the initial test were constructed with competitive lures. Participants then took the final CR test after either a five minutes delay or 48 hours delay. Participants' performance on the final CR test was slightly reduced after 48 hours delay compared to the five minutes delay; however, the results indicated an enhancement in performance on both the repeated and related questions compared to the control questions regardless of the length of the interval between the initial and the final tests. These findings showed that MC testing on related items was useful for both short- and long-term retention.

Research has identified that the information retrieval that occurs during the initial test is an important determinant of performance on related items in the final test. Thus, some studies have compared a more retrieval-provoking MC format with the standard MC format (i.e., select a single option). For example, Sparck et al. (2016) had participants answer either standard MC questions or confidence-weighted MC questions in the initial test. The confidence-weighted format requires participants to rate their confidence in each of the options relative to the others, thereby involving more retrieval processes. The results of the final CR test revealed better performance on the related items compared to the control items, regardless of the initial testing format. However, the best performance was observed when the confidence-weighted format was used during the initial test. Utilizing a confidence-weighted format forced participants to engage in deeper processing to exclude the incorrect options and choose the correct one.

To increase the amount of processing required, Little et al. (2019) designed an elimination MC format that required participants to identify the correct answer and to provide reasons for rejecting the unchosen options. In this study, the initial test did not follow a reading session; instead, participants started by answering general knowledge questions which came in an elimination format. The results of the final CR test showed better performance on the repeated and related items compared to the control items. Also, answering related questions more accurately in the final test was associated with participants providing more information in their rejection of lures during the first test. Together, these results provide strong evidence of the importance of boosting lure elaboration during initial MC testing either via competitive lures (e.g., Little & Bjork, 2016) or the format of practice tests (e.g., Little et al., 2019), as it can induce more cognitive processing than other types of testing, such as the CR format (e.g., Little et al., 2012).

Negative Effects of MC Testing on Related Items

Previous studies that investigated the testing effect of MC questions on related items mostly used a CR test format in the final test, in which participants performed better on related items than on new items (e.g., Little & Bjork, 2015). Would the same findings be observed with related items when the final test utilizes a test format that is associated with familiarity-based responses, such as MC questions? Higham et al. (2016) addressed this question in several experiments that examined the effects of initial MC testing on related items similar to those used by Little, Bjork and colleagues. However, as the final test was an MC test rather than a CR test, the researchers repeated the same set of alternatives for each pair of questions in the initial and final tests. For example, a question on the initial test is, "What is one of the most blatant examples of how the media can induce public opinion? a. celebrity endorsements, b. advertisements, c. live telecasts of events, d. images of attractiveness norms". The related question on the final test is, "What has been blamed for the prevalence of eating disorders among women in the US? a. celebrity endorsements, b. advertisements, c. live telecast of events, d. images of attractiveness norms".

Higham et al. (2016) had participants read expository texts, take an initial standard MC test with feedback provided after each question, and then take a final MC test either immediately or after a seven-day retention interval. The final MC test contained some repeated questions from the initial test, some related but untested questions and some new questions. The findings demonstrated that participants performed better on the repeated items than on the new items. However, the results showed that taking a practice MC test resulted in worse performance on related items relative to new items, regardless of the retention interval. An analysis of participants' answer selections on the

final test indicated that the impairment was due to participants largely endorsing the corrective feedback from the initial test as the correct answer on the final test, although it was no longer correct.

Moreover, the findings showed that participants continued to erroneously select the corrective feedback even when the repeated items were dropped from the final test as they might encourage participants to falsely recognise the related items as repeated. The impairment remained even when only one option (the corrective feedback) was repeated in the two questions and when participants were asked to read the questions aloud to ensure that the question stems were not ignored. Together, these results showed a different pattern than that observed in the CR studies (e.g., Little & Bjork, 2015). Although Little and Bjork (2015; Little et al., 2019) found that lure intrusions occurred occasionally in CR related items, they did not overshadow the positive testing effect of initial MC questions. In contrast, endorsing the corrective feedback in the final MC test in Higham et al.'s (2016) study negatively affected participants' performance on the related items in the final test compared to the new items.

One might consider the impairment on MC related items observed in Higham et al.'s (2016) study as another example of the negative testing effect (e.g., Roediger & Marsh, 2005). However, it is noteworthy that participants' worse performance in the former study was due to having received corrective feedback during the initial test; that is, participants selected the corrective feedback on the later MC test when the answer was no longer correct, which harmed their overall performance. On the other hand, the original negative testing effect occurred when participants erroneously selected one of the lures during the initial test, received no feedback and then repeated the same error when facing the question again in the final CR test. However, providing corrective

feedback during practice tests was found to reduce the negative testing effect and increase the positive testing effect (Butler & Roediger, 2008). Thus, with the original negative testing effect, the problem is repeating uncorrected errors. In contrast, the effect in Higham et al.'s (2016) study was attributed to selecting corrective feedback when it was no longer a correct answer.

Theoretical Mechanisms of the Testing Effect

Few studies have attempted to explore the theoretical mechanisms underlying the testing effect. One of these studies, conducted by Chan and McDermott (2007), aimed to identify the potential applications of the dual-process model in the testing effect. The dual-process model proposes that two types of processes – recollection and familiarity – underlie the performance of memory (for a review of the dual-process model, see Yonelinas, 2002). In three experiments, Chan and McDermott had participants study word lists, then either take initial free recall tests (the testing condition) or answer maths questions (no testing condition), and then take final recognition tests. By using different measures, the researchers found that the positive testing effect relayed on enhancing recollection processes, but not familiarity. They also suggested that the testing effect can be observed when the final test comes in a recognition format but only if the nature of the recognition test promoted recollection processes (e.g., by using very similar lures that could not be rejected by relying on familiarity). However, if participants can easily recognise the target answer with no need to engage in effortful recollection, then familiarity may mask the testing effect. These findings are consistent with the findings of other studies that reported the contribution of recollection processes to the positive testing effect (e.g., Pu & Tse, 2014; Verkoeijen et al., 2011).

Ozuru et al. (2013) investigated text comprehension which was measured by two different testing formats – MC versus open-ended questions; they argued that answering open-ended questions relies roughly on controlled processes (i.e., recollection), while MC questions encourage reliance on automatic processes (i.e., familiarity). This is due to the nature of each testing format; that is, recollection tests provide very limited information and might promote the activation of several cues related to the desired information in a controlled retrieval process. Conversely, recognition tests, such as MC tests, present the correct answer as part of the question, which might encourage more reliance on familiarity. Consisting with Chan and McDermott (2007), the researchers suggested that increasing the similarity of the lures in MC questions might discourage learners from relying on automatic processes and strengthen controlled ones.

The controlled process is considered slow, conscious, and requires more attentional resources, whereas the automatic one can be fast, less conscious, and requires fewer attentional resources (Jacoby, 1991; Neely, 1977). Thus, memory performance that relies on automatic influences might produce opposite results to that that depends on controlled influences.

Regarding the testing effect of MC tests, both positive and negative effects (tested items) can be roughly attributed to the controlled influences of memory. The positive testing effect occurs when participants retrieve information in a previous episode (i.e., practice test) and recall it again in the later CR test (e.g., McDaniel et al., 2012). Similarly, the negative testing effect occurs when participants selected a lure on the initial MC test, and because they did not receive corrective feedback (i.e., they did not know it was a lure), they recall it again on later CR tests believing that it was a correct answer (e.g., Roediger & Marsh, 2005). Thus, when providing corrective feedback (e.g., Butler &

Roediger, 2008) the result is an increase in the positive testing effect and a reduction in the negative testing effect which might be considered evidence of controlled processes. Indeed, this is in agreement with Fazio et al. (2010) who suggested that controlled processes underly both positive and negative testing effects.

In terms of related items, the facilitation observed in the CR studies can be attributed to the controlled memory processes. Specifically, participants intensively retrieve information about the lures and recall it later due to using competitive lures (e.g., Little & Bjork, 2015) or an initial test format that promotes lure elaboration (e.g., Sparck et al., 2016). Answering difficult MC practice questions followed by a final test format that can provoke deep processing (i.e., CR) results in enhanced performance. In contrast, when the final test format is one that might encourage familiarity-based responses (i.e., MC) and in which the options for related questions are the same, automatic processes dominate performance (e.g., Higham et al., 2016). This is due to participants falsely recognising the related questions as repeated and therefore selecting the corrective feedback, which is no longer the correct answer, causing performance to be impaired.

The literature to date suggests that whether performance is facilitated or impaired on related items depends on the format of the final test, that is, CR and MC respectively. Indeed, Chan and McDermott (2007) found that the positive testing effect was always observed when the final test was recollection (e.g., CR tests) but not when the final test was a recognition test (e.g., MC tests). In light of the dual-process model, we aimed to compare participants' performances on related items in final MC versus final CR tests. We also aimed to investigate the circumstances under which performance on related items can be impaired/enhanced in final MC tests.

Summary

In this research, we investigated the interplay between controlled and automatic processes of retrieval practice whilst utilizing initial MC testing to examine its consequences on repeated and, more importantly, related untested items (versus new items). We used related pairs that were analogous to those used in prior research (e.g., Little & Bjork, 2015); that is, they queried the same topic (e.g., both questions in a given pair were about capitals of African countries), and for the MC version, each pair had the same set of alternatives.

In Paper 1 (comprising three experiments), we compared participants' performance on final MC and CR tests whilst increasing the level of lure elaboration that occurred during the initial MC test. The format of the initial MC test varied in terms of the retrieval level that each test provoked, from the lowest level of retrieval (standard MC; i.e., select a single option) to the medium level (ranking; i.e., rank order the options) to the highest level (elimination; i.e., provide reasons for rejecting unchosen options).

In Paper 2, Experiment 1 with two groups of participants, we examined the extent to which providing or withholding corrective feedback during the first test impacted participants' performance on the final MC test. In Experiment 2, we compared two MC groups, one of which received opposition instructions prior to the final test, while the other received regular instructions. Also, to identify the role of false recognition in answering MC related items, we asked participants to identify each question in the final test as "old" (i.e., it was seen on the first test) or "new" (i.e., it was not seen on the first test).

In Paper 3, we investigated whether controlled processes can be observed in final MC testing whilst using authentic educational materials. In Experiment 1, we manipulated two variables: first, the presence/absence of repeated items in the final MC test, and

second, whether the questions appeared to participants as a single test or as two separate tests. In Experiments 2 and 3, we shifted our investigation to a real educational context. In Experiment 2, we manipulated the interval separated between the initial test and the final test. In Experiment 3, we manipulated the presence/absence of corrective feedback during the initial test.

By conducting these eight experiments and using different measurements, we hope to provide some evidence of the role of controlled and automatic processes in MC testing and highlight the different consequences of using MC testing as a learning tool.

Paper 1 [The Dark Side of Corrective Feedback: Controlled and Automatic Influences of Retrieval Practice]

Manuscript published: Journal of Experimental Psychology: Learning, Memory, and Cognition

Abstract

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

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

The Dark Side of Corrective Feedback: Controlled and Automatic Influences of Retrieval Practice

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

Positive and Negative Testing Effects

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

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

Positive Effects of MC Testing on Related Items

Thus, although some research showed the negative effect with MC testing, the positive testing effect typically outweighed the negative effect. This result indicates that even with the negative testing effect, taking initial MC tests remains advantageous. This

[1] "Lures" is another term for incorrect alternatives.

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

Little and Bjork (2015) demonstrated that the benefits of practice testing using materials like these depended on the plausibility of the lures. They had participants first read two expository texts (e.g., information about ferrets), take a practice MC test without feedback, and then take a final CR test. For the practice test, there were two types of MC questions: questions with competitive lures versus questions with uncompetitive lures. For example, one question with competitive lures was "*How many inches long is an average ferret tail? a. 7-10, b. 20,* ***c. 5***" (boldface indicates the correct answer). Uncompetitive lures were less plausible given the question; for example, to create the uncompetitive version of question on ferrets' tail length in inches, the first two lures were replaced with "*1500*" and "*3500*", respectively. For the final CR test, there were three types of questions: previously tested questions, related untested questions, and new questions that were generated from a passage that participants had read but not been tested on during practice. The findings from the final CR test showed best performance on previously tested questions – the standard testing effect. More critically, there was also better performance on the related untested items compared to the new

questions. However, this advantage only occurred if the questions were accompanied by competitive lures, not if the lures were uncompetitive. Little and Bjork (2015) argued that competitive lures on the first MC test prompted elaboration and retrieval of information about the question, which facilitated participants' later ability to answer related but untested questions on the CR test. Because the questions with uncompetitive lures were so easy, the same information retrieval was not necessary during practice, and so no facilitation was observed on the final test.

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

Theoretical Mechanisms

The prior literature suggests that there are both positive and negative effects of MC practice testing. In our view, these positive and negative effects (tested items) correspond roughly to controlled influences of memory (see Chan & McDermott, 2007 for a similar dual-process perspective on testing effects). Indeed, Fazio et al. (2010) suggested that both positive and negative effects of MC testing are attributed to controlled influences of memory. In the positive effect, participants consciously utilize information retrieved

during practice to answer the later test. Similarly, the negative testing effect occurs when participants selected a lure during practice, and as a result of not receiving feedback, it was recalled again on later CR tests. Therefore, providing corrective feedback during practice tests limited the negative effect and boosted the positive effect.

In terms of the related items, if hard practice MC questions are used which can only be answered with intensive retrieval (Little & Bjork, 2015) or if the practice MC format requires lure elaboration (Sparck et al., 2016), the result is a controlled influence of memory which facilitates later CR performance. In all cases that showed controlled effects of retrieval practice, the final test was CR. However, in Chan and McDermott (2007) research, a testing effect was always observed in final CR tests but not when recognition tests, such as MC, were used as the final test. If both the practice and final tests were MC – tests that are often associated with familiarity-based responding (e.g., Ozuru et al., 2013) – would controlled influences from lure elaboration during practice as observed by Little and colleagues be overshadowed by automatic influences?

Higham et al. (2016) tested this possibility in several experiments using untested related items like those used in Little and colleagues' paradigm. After a reading session, participants took an MC initial test. However, unlike Little and Bjork (2015), participants were provided with corrective feedback after each question in the practice test. Moreover, instead of using a CR test format in the final stage of these experiments, a format likely to encourage effortful retrieval, they used a standard MC test. Untested related questions on the final MC test were analogous to those for the final CR test used in prior research; that is, they queried the same topic, but a lure from the practice test was now the correct option. The main difference was that the same set of options used on the practice MC test were explicitly presented as options on the final MC test. Thus, in

most experiments, the options on the final MC test consisted of the previous correct answer (which was now a lure), a new correct answer (which was previously a lure), and two additional lures. The question that Higham et al. (2016) addressed was whether participants would continue to show facilitated performance due to controlled memory influences, as in Little and colleagues' research. They reasoned that false familiarity derived from explicit presentation of the previously correct option, whose processing would have been reinforced by feedback during practice (regardless of whether it was chosen), might override any controlled influences of memory. If so, then performance on untested related questions would show worse, not better, performance compared to new items.

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

that if the final test format allows for quick familiarity-based responding, any controlled influences attributable to lure elaboration during practice might be overshadowed by the false familiarity of the corrective feedback.

Little and Bjork (2015; see also Little et al., 2019) examined the issue of false recollections/intrusions (e.g., Jacoby et al., 1993) in their research with final CR tests and concluded that, although they occasionally occurred, they did not overshadow the controlled memory influence. However, given the robustness of Higham et al.'s (2016) false feedback effect with MC final tests, and the fact that lure intrusions occur with final CR tests, we believe further research on automatic and controlled memory influences in both MC and CR is warranted.

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

Current Study

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

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

Experiment 1

The first experiment explored the consequences of using different final test formats on the learners' outcomes. Participants first answered a set of MC general knowledge questions on an initial test. On a second test taken immediately afterwards, some questions were repeated, some were related but untested, and some were new. Like Higham et al. (2016), we hypothesized that performance on the MC related items will be impaired compared to new questions due to the false familiarity of the corrective feedback option that is repeated from the practice phase. In contrast, because no option

is repeated between phases if the final test is CR, any misleading effects of the corrective feedback will be limited. Instead, we predicted that controlled memory influences would prevail if the second test is in CR format, and performance on related items will be facilitated compared to new items.

Method

Participants

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

Design

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

Materials and Procedure

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

The experiment was conducted online using Qualtrics survey software, and two groups of participants were fully instructed on the procedure prior to taking the survey. After reading the instructions, participants answered two attention-check questions about the instructions to ensure that they read and understood them. The experiment started with taking the initial test, consisting of 22 standard MC questions for about seven minutes. For each question in the initial test, a question stem was presented with four alternatives below it. Participants selected a single alternative by clicking the radio button that appeared next to it and then clicked "Next" to receive corrective feedback. The feedback was provided by Qualtrics software which highlighted the correct answer with a

green colour and a red colour for the three incorrect alternatives. Clicking “Next” again advanced to the next question. After completing the initial test, participants engaged in a filler task of answering basic mathematics questions for about four minutes. Until this point, the procedure was identical for both groups.

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

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

Results

Initial Test Performance

The initial test was in the same MC format for both the MC and CR groups. (The group names refer to the format of the final test.) Although the two groups were treated the same on the initial test, we compared performance to check for sampling error. For the MC group, the mean proportion of items which participants answered correctly on the first MC test was .62 (*SD* = .17), while it was .64 (SD = .19) for the CR group. An independent samples *t*-test showed that this difference was not significant $t(168) = -0.59$, $p = .56$, $d = .09$, suggesting that there was no evidence that the MC and CR groups were different.

Final Test Performance

Final-Test Accuracy. We conducted a 2 (final test type: MC, CR) x 3 (question type: new, repeated, related) mixed-factor Analysis of Variance (ANOVA) with final test accuracy as the dependent variable (see Figure 2). We found a significant main effect of final test type, $F(1, 170) = 26.39$, $p < .001$, $\eta_p^2 = .13$, such that accuracy was significantly higher on the MC test (*M* = .69; *SD* = .16) compared to the CR test (*M* = .55; *SD* = .18). We also found a significant main effect of question type, $F(2, 340) = 204.68$, $p < .001$, $\eta_p^2 = .55$. Paired-sample *t*-tests revealed that there was no significant difference in accuracy between the new (*M* = .53, *SD* = .24) and related (*M* = .51, *SD* = .24) items, $t(171) = -0.81$, $p = .42$, $d = .06$, but accuracy was significantly higher on the repeated items (*M* = .82, *SD* = .18), compared to the new, $t(171) = -16.82$, $p < .001$, $d = 1.36$, and related items, $t(171) = -17.73$, $p < .001$, $d = 1.45$. Finally, we found a significant interaction between final test type and question type, $F(2, 340) = 4.22$, $p = .02$, $\eta_p^2 = .02$. For the CR group, paired-sample *t*-tests showed higher performance on the repeated items (*M* = .78, *SD* = .19) compared to related items (*M* = .44, *SD* = .22), $t(86) = -13.48$, $p < .001$, $d = 1.60$, and new items (*M* = .44, *SD* = .24), $t(86) = -13.63$, $p < .001$, $d = 1.55$, while there was no significant difference

between new and related items, $t(86) = 0.16$, $p = .87$, $d = .02$. The same general pattern was observed in the MC group, but the differences between performance on the repeated items and the other two item types were smaller, resulting in the interaction. Specifically, participants performed better on the repeated items ($M = .86$, $SD = .16$) versus related items ($M = .58$, $SD = .23$), $t(84) = -11.07$, $p < .001$, $d = 1.39$, and new items ($M = .62$, $SD = .21$), $t(84) = -10.53$, $p < .001$, $d = 1.30$, while there was no significant difference between new and related items $t(84) = -1.37$, $p = .17$, $d = .15$ [2].

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

This analysis produced five mutually exclusive second test possibilities (see Table 1): (a) a correct answer on the second test that matched the previous answer on the initial test (correct/previous answer); (b) a correct answer that was neither the previous answer nor the corrective feedback on the first test (correct/other); (c) an incorrect answer that

[2] Note that there is no adjustment of alpha for the multiple follow-up t-tests in this analysis or the other analyses in this paper.

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

We conducted a 2 (final test type: MC, CR) x 3 (final-test answer type: previous answer, corrective feedback, other) mixed-factor ANOVA with the probability of answering with each type of answers as the dependent variable. We found no significant main effect of final test type, $F < 1$. However, we found a significant main effect of final-test answer type $F(2, 340) = 30.66$, $p < .001$, $\eta_p^2 = .15$. Paired-sample *t*-tests showed no significant difference between the probabilities of answering with the corrective feedback ($M = .23$, $SD = .30$) and other ($M = .29$, $SD = .36$), $t(171) = 1.64$, $p = .10$, $d = .19$, but both probabilities were higher than that for the previous answer ($M = .08$, $SD = .16$), $t(171) = 6.43$, $p < .001$, $d = .65$, and $t(171) = 7.24$, $p < .001$, $d = .78$, respectively. Finally, we found a significant interaction between test and question types $F(2, 340) = 32.71$, $p < .001$, $\eta_p^2 = .16$. For the MC group, paired-sample *t*-tests revealed that participants overwhelmingly selected the corrective feedback from the initial test ($M = .30$, $SD = .32$) compared to other answer ($M = .15$, $SD = .19$), $t(84) = -3.67$, $p < .001$, $d = .59$, and previous answer ($M =$

.10, *SD* = .17), *t*(84) = 5.62, *p* < .001, *d* = .80. However, in contrast to the MC group, participants in the CR group were more likely to produce other answers (*M* = .46, *SD* = .45) compared to corrective feedback (*M* = .15, *SD* = .23), *t*(86) = -5.63, *p* < .001, *d* = .85, and previous answer (*M* = .05, *SD* = .15), *t*(86) = -7.92, *p* < .001, *d* = 1.20.

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

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

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

Discussion

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

Despite the failure to replicate those aspects of the results, when we examined the probability of endorsing the corrective feedback between the related and new items,

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

The analysis of final-test response types (given an incorrect initial answer) shed some light on the nature of these automatic processes. First, it is noteworthy that not all familiar options were endorsed at the same rate on the second MC test. For example, previous wrong answers were largely avoided, a result that replicates Higham et al.'s (2016) finding. Participants in the MC group appeared to be biased to selecting the option that was previously correct rather than incorrect. In contrast, participants tended to avoid both the previously incorrect answer and the corrective feedback on the CR test. Instead, they responded with a new "other" answer.[3]

There are several possible reasons for these different patterns of responding between the two test types. One possibility is that the presence of the options on the MC test, which always matched those presented during practice, acted as strong retrieval cues for the earlier presentation of the related questions. The strong match between the related pairs may have falsely led participants to believe that the question was repeated, which would warrant choosing the corrective feedback from the first test. The absence of the matching options on the CR test may have led participants to notice that the question

[3] Obviously, given the results of the previous analysis and the CR data in Figure 4, participants did intrude with the corrective feedback more often with related items than new items. However, compared to the MC group, their tendency to endorse the corrective feedback was less, at least when the practice test question was answered incorrectly (Table 1).

stem had changed, and that a new answer was required. Importantly, though, if participants detected at a high rate that the question stem had changed on the CR test, it was not enough to control the automatic influence of memory. As noted above, CR participants still chose the corrective feedback for related questions at a higher rate than for the new questions.

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

Experiment 2

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

controlled memory processes that might be working in conjunction with automatic influences in this paradigm.

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

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

Method

Participants

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

Design and Materials

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

Procedure

The procedure was identical to that in Experiment 1 with some exceptions. First, we modified the standard MC format in the initial test to use ranking. Rather than selecting the correct answer, participants were asked to rank the four alternatives provided for

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

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

Results

Initial Test Performance

All participants completed the initial MC ranking test and were treated the same at this stage of the experiment. Nonetheless, we compared the response accuracy of the two experimental groups to test for sampling error. The mean proportion of questions answered correctly on the first MC test in the MC and CR groups was .59 (*SD* = .20) and .61 (*SD* = .19), respectively. An independent sample *t*-test indicated that the difference was not significant, $t(169) = -0.45$, $p = .65$, $d = .07$, suggesting that there was no evidence that the MC and CR groups were different.

Final Test Performance

Final-Test Accuracy. We conducted a 2 (final test type: MC, CR) x 3 (question type: new, repeated, related) mixed-factor ANOVA with final test accuracy as the dependent variable (see Figure 5). The analysis revealed a non-significant effect of final-test type $F(1, 170) = 3.08$, $p = .08$, $\eta_p^2 = .02$, such that accuracy was not significantly different between the MC group ($M = .58$, $SD = .15$) and the CR group ($M = .54$, $SD = .17$). There was a significant main effect of question type, $F(2, 340) = 285.30$, $p < .001$, $\eta_p^2 = .63$. Paired-sample *t*-tests revealed that there was no significant difference in accuracy between the new ($M = .44$, $SD = .24$) and related items ($M = .41$, $SD = .24$), $t(171) = -1.74$, $p = .08$, $d = .15$, but accuracy was significantly higher on the repeated items ($M = .82$, $SD = .19$) compared to the new $t(171) = -20.38$, $p < .001$, $d = 1.80$, and related items $t(171) = -19.32$, $p < .001$, $d = 1.95$.

The main effect of question type was qualified by a significant interaction with final test type, $F(2, 340) = 19.59$, $p < .001$, $\eta_p^2 = .10$. Paired-sample *t*-tests revealed that the interaction was due to the CR group performing better on the related items ($M = .45$, $SD = .23$) than the new items ($M = .39$, $SD = .22$), $t(86) = 2.57$, $p = .01$, $d = .29$, and significantly better on the repeated items ($M = .77$, $SD = .19$) compared to new items, $t(86) = -14.71$, $p < .001$, $d = 1.82$, and related items, $t(86) = -11.89$, $p < .001$, $d = 1.49$. In contrast, participants in the MC group performed better on the new items ($M = .50$, $SD = .23$) than the related items ($M = .36$, $SD = .24$), $t(84) = -4.77$, $p < .001$, $d = .59$, and significantly better on the repeated items ($M = .88$, $SD = .16$) compared to new items, $t(84) = -14.06$, $p < .001$, $d = 1.92$, and related items, $t(84) = -17.28$, $p < .001$, $d = 2.56$.

Final-Test Answer Types. We conducted the same analysis as in Experiment 1, conditioning the final test answers on incorrect initial test answers, producing the same

five mutually exclusive possibilities (see Table 2). As before, we converted part marks in the CR group to full marks so that final test responses could be classified as either correct or incorrect. Then we conducted a 2 (final test type: MC, CR) x 3 (final-test answer type: previous answer, corrective feedback, other) mixed-factor ANOVA with the probability of answering with each type of answers as the dependent variable. The main effect of final test type was not significant, $F < 1$. However, we found a significant main effect of final-test answer type $F(2, 340) = 40.20$, $p < .001$, $\eta_p^2 = .19$. Paired-sample *t*-tests showed that there was no significant difference between the probabilities of answering with the corrective feedback ($M = .32$, $SD = .39$) and other option ($M = .28$, $SD = .36$), $t(171) = -0.86$, $p = .38$, $d = .10$, but both probabilities were higher than that for the previous answer ($M = .07$, $SD = .15$), $t(171) = 8.13$, $p < .001$, $d = .85$, and $t(171) = 7.08$, $p < .001$, $d = .78$, respectively. Also, we found a significant interaction between test type and answer type $F(2, 340) = 67.62$, $p < .001$, $\eta_p^2 = .28$. For the MC group, paired-sample *t*-tests revealed that participants overwhelmingly selected the corrective feedback from the initial test ($M = .50$, $SD = .45$) compared to other answer ($M = .11$, $SD = .17$), $t(84) = -7.39$, $p < .001$, $d = 1.14$, and previous answer ($M = .11$, $SD = .17$), $t(84) = -7.37$, $p < .001$, $d = 1.14$. For the CR group, participants were more likely to produce other answers ($M = .46$, $SD = .41$) compared to corrective feedback ($M = .15$, $SD = .19$), $t(86) = 6.60$, $p < .001$, $d = .97$, and previous answers ($M = .03$, $SD = .11$), $t(86) = -8.90$, $p < .001$, $d = 1.40$.

False Endorsements of Corrective Feedback. As in Experiment 1, we analysed the probability of endorsing the corrective feedback between the related and new items. We conducted a 2 (final test type: MC, CR) x 2 (question type: new, related) mixed-factor ANOVA with the probability of answering with the corrective feedback as the dependent variable (see Figure 6). The ANOVA revealed a significant main effect of final test type,

$F(1, 170) = 147.7$, $p < .001$, $\eta_p^2 = .46$. The probability of answering with the corrective feedback was higher in the MC group ($M = .34$, $SD = .15$) compared to the CR group ($M = .11$, $SD = .09$). We also found a significant main effect of question type, $F(1, 170) = 89.34$, $p < .001$, $\eta_p^2 = .34$. The probability of answering with the corrective feedback was higher for the related items ($M = .32$, $SD = .28$) compared to new ($M = .12$, $SD = .13$). Finally, we found a significant interaction between final test type and question type, $F(1, 170) = 12.04$, $p < .001$, $\eta_p^2 = .07$.[4] Paired-sample *t*-tests revealed that participants in both groups were more likely to endorse the corrective feedback for related items compared to new items, but the difference was larger in the MC group (related: $M = .47$, $SD = .29$; new: $M = .21$, $SD = .14$), $t(84) = 7.29$, $p < .001$, $d = 1.16$, compared to the CR group (related: $M = .17$, $SD = .16$; new: $M = .04$, $SD = .07$), $t(86) = 6.39$, $p < .001$, $d = .99$.

Discussion

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

[4] This interaction needs to be treated with caution given that feedback intrusions for new questions in the CR group were near floor.

lures. This retrieval enhanced participants' overall performance with related items on the second test.

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

Our attempt at increasing elaborative retrieval during the initial test by using a ranking format rather than a standard one produced clear benefits to controlled processing as the results showed facilitation for related versus new items on the CR test. However, the results suggest that the CR performance was affected as well by automatic influences, as evidenced by corrective feedback intrusions in CR. On the other hand, the difference in accuracy between related and new items on the MC test was larger (and statistically significant) in Experiment 2 compared to Experiment 1 (where the difference was not significant). Also, the tendency to respond with corrective feedback on related questions on the MC test was greater in Experiment 2 (M = .47, SD = .29) than in Experiment 1 (M = .24, SD = .20), $t(147) = -6.00$, $p < .001$, d = .93. Thus, rather than the ranking format increasing controlled influences on the final MC test – influences that we

expected would counter automatic influences – it appears to have enhanced automatic influences instead.

Experiment 3

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

Method

Participants

Based on the power analysis reported in Experiment 1, we aimed to recruit at least 128 participants. We actually tested 246 participants, but due to the demanding

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

Design and Materials

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

Procedure

The procedure was mostly identical to Experiment 2 with a few exceptions. First, instead of using the ranking format for the initial test, we replicated Little et al.'s (2019, Experiment 1) design and used an elimination format. On the initial test, each question

[5] Little et al. (2019) did not report eliminating any participants in their study, although we employed the same design on the first test and used the same platform (MTurk).

stem came with four alternatives and below that, we added four empty boxes where participants were asked to type the chosen option in the first box with no need to provide any reason for this option. However, participants were asked to write the three unchosen options and their reasons for rejecting each one in the three remaining boxes. Prior to the initial test, participants were given an example question of how they should respond to the questions in the initial test. Like Experiment 2, each question in the initial test was displayed for 15 seconds before the "Next" button was displayed, which allowed participants to advance to the next question. Participants were not forced to provide reasons for rejecting each lure, so they could advance after the 15 seconds had elapsed without the need to provide a reason for each rejected lure. The rest of the procedure was identical to that of Experiments 1 and 2.

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

Results

Initial Test Performance

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

elimination process by merely identifying the correct and incorrect answers and only occasionally providing some reasons.

To test for sampling error, we again compared the two groups on initial test accuracy even though they were treated the same at that point in the experiment. The mean proportion of initial MC test questions answered correctly was .69 (*SD* = .19) and .66 (SD = .20) for the MC and CR groups, respectively. An independent sample *t*-test showed that the difference was not significant $t(160) = 1.18$, $p = .24$, $d = .19$. Thus, as in Experiments 1 and 2, there was no evidence that the MC and CR groups were different.

Final Test Performance

Final-Test Accuracy. We conducted a 2 (final test type: MC, CR) x 3 (question type: new, repeated, related) mixed-factor ANOVA with final-test accuracy as the dependent variable (see Figure 7). The main effect of final test type was not significant, $F(1, 160) = 1.09$, $p = .30$, $\eta_p^2 = .01$; such that there was no significant difference in accuracy between the MC (*M* = .57, *SD* = .19) and CR (*M* = .54, *SD* = .18) tests. However, there was a significant main effect of question type, $F(2, 320) = 227.8$, $p < .001$, $\eta_p^2 = .59$. Paired-sample *t*-tests revealed that there was no significant difference in accuracy between the new (*M* = .43, *SD* = .25) and related items (*M* = .43, *SD* = .26), $t(161) = -0.30$, $p = .76$, $d = .03$, but accuracy was significantly higher on the repeated items (*M* = .81, *SD* = .22) compared to the related items, $t(161) = 16.76$, $p < .001$, $d = 1.60$, and new items, $t(161) = 17.66$, $p < .001$, $d = 1.61$. The main effect of question type was qualified by a significant interaction between final test type and question type, $F(2, 320) = 21.7$, $p < .001$, $\eta_p^2 = .12$. Paired-sample *t*-tests revealed that the interaction was due to the CR group performing better on the related items (*M* = .49, *SD* = .22) than the new items (*M* = .37, *SD* = .23), $t(80) = 5.16$, $p < .001$, $d = .51$, and significantly better on the repeated items (*M* = .76, *SD* = .21)

compared to the new items, $t(80) = -12.13$, $p < .001$, $d = 1.75$, and the related items, $t(80) = -9.97$, $p < .001$, $d = 1.25$. In contrast, participants in the MC group performed better on the new items ($M = .49$, $SD = .25$) than the related items ($M = .36$, $SD = .28$), $t(80) = -4.22$, $p < .001$, $d = .48$, and significantly better on the repeated items ($M = .86$, $SD = .20$) compared to the new items, $t(80) = 12.89$, $p < .001$, $d = 1.60$, and the related items $t(80) = 15.23$, $p < .001$, $d = 2.01$.

Final-Test Answer Types. We conducted the same analysis on final-test answer types as in Experiments 1 and 2, producing the same five mutually exclusive second test response rates for items incorrectly answered on the initial test. As before, part marks in the CR group were converted to full marks so that responses could be classified as either correct or incorrect. We then analysed the three possibilities where the answers were incorrect on both the first and final tests (see Table 3). A 2 (final test type: MC, CR) x 3 (final-test answer type: previous answer, corrective feedback, other) mixed-factor ANOVA, with the probability of answering with each type of answer as the dependent variable, revealed that the main effect of final test type was not significant, $F < 1$. However, there was a significant main effect of final-test answer type $F(2, 320) = 26.39$, $p < .001$, $\eta_p^2 = .14$. Paired-sample *t*-tests showed that there was no significant difference between the probabilities of answering with the corrective feedback ($M = .28$, $SD = .40$) and other ($M = .28$, $SD = .37$), $t(161) = 0.04$, $p = .97$, $d = .00$, but both probabilities were higher than that for the previous answer ($M = .07$, $SD = .16$), $t(161) = 6.31$, $p < .001$, $d = .68$, and $t(161) = 6.61$, $p < .001$, $d = .73$, respectively. Moreover, we found a significant interaction between test type and answer type $F(2, 320) = 39.11$, $p < .001$, $\eta_p^2 = .20$. For the MC group, paired-sample *t*-tests revealed that participants overwhelmingly selected the corrective feedback from the initial test ($M = .43$, $SD = .48$) compared to other answer ($M = .14$, $SD = .21$),

t(80) = -5.43, *p* < .001, *d* = .81, and previous answers (*M* = .11, *SD* = .21), *t*(80) = -5.62, *p* < .001, *d* = .89. For the CR group, participants were more likely to produce "other" answer (*M* = .42, *SD* = .42) compared to corrective feedback (*M* = .14, *SD* = .24), *t*(80) = 5.24, *p* < .001, *d* = .82, and previous answer (*M* = .04, *SD* = .10), *t*(80) = -7.73, *p* < .001, *d* = 1.24.

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

Discussion

[6] This interaction needs to be treated with caution given that feedback intrusions for new questions in the CR group were near floor.

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

In contrast, the enhanced lure elaboration associated with elimination testing did nothing to temper automatic influences on the final MC test. In fact, the accuracy difference between the related and new questions was as great (or greater) in this experiment (M_{diff} = .13) as it was in the previous two experiments (Experiment 1: M_{diff} = .04; Experiment 2: M_{diff} = .14). The feedback analysis also showed that the enhanced likelihood of endorsing the corrective feedback for related questions compared to new ones was again comparable for MC participants in this experiment (M_{diff} = .21) as in the previous ones (Experiment 1: M_{diff} = .07; Experiment 2: M_{diff} = .26). Moreover, the final-test answer type analysis showed that MC participants still showed a tendency to respond with corrective feedback when answers on the earlier test were wrong. Together, these results suggest that even if it is ensured that participants are processing practice test

questions at a deep level, it does not guarantee that there will be positive transfer of controlled processes to a later test. If the later test is MC, lure elaboration can even worsen the automatic effects (cf. related vs. new difference in Experiment 1 vs. Experiment 3).

General Discussion

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

Related Versus New Questions

Of greater interest to the current research was how participants performed on the final test with questions related to ones answered (with feedback) from the practice phase versus new questions. Here, the effects of retrieval practice were more equivocal, and depended on the format of both the initial and final tests. In Experiment 1, which used the standard "choose the one best option" MC format during practice, there were no differences in accuracy between the related and new items on the CR final test. Following Little and colleagues (e.g., Little & Bjork, 2015; Little et al., 2012; Sparck et al., 2016), we attributed the failure to observe a difference in CR to low levels of lure elaboration during the practice test, which would have limited any controlled memory influences on the later test. Unless the practice test encourages retrieval of information about the lures, either by the lures being competitive (Little & Bjork, 2015; Little et al.,

2012) or by the practice test format encouraging deep retrieval (Sparck et al., 2016), then accuracy on related test items has not typically been enhanced relative to new items. Therefore, a standard initial MC may not be enough to promote retrieval and produce controlled influences, especially if it is self-paced as in Experiment 1, which might have encouraged participants to finish the task very quickly without really taking the time to consider each of the lures. Indeed, in an analysis of the time participants spent completing each experiment, we found that the average time increased monotonically across experiments, from 18 minutes in Experiment 1, to 27 minutes in Experiment 2, to 34 minutes in Experiment 3. Independent sample *t*-tests showed that each successive time increase across experiments was significant (Experiment 1 vs. 2: $t(329) = -6.52$, $p < .001$, $d = .72$; Experiment 2 vs. 3: $t(313) = -5.35$, $p < .001$, $d = .60$).

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

For the final MC test in Experiment 1, we also found no difference in accuracy between the related items compared to new items. This result was somewhat surprising given Higham et al.'s (2016) results. They found across a variety of experimental contexts that accuracy on related questions suffered because participants tended to erroneously endorse the option on the second test that corresponded to the first test's corrective feedback. However, a closer analysis of the data suggested that this tendency did exist in the MC group. That is, the probability of specifically selecting the corrective feedback was

higher for related questions than new questions. Moreover, an analysis of the types of second test responses participants made after answering the practice question incorrectly showed that corrective feedback was the favourite choice. However, this tendency was not strong enough to manifest itself in accuracy differences between related and new items.

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

For the MC test, however, the switch to a ranking format for the first test in Experiment 2 rendered answers to related items being significantly lower compared to new items, replicating Higham et al. (2016). Furthermore, analysing the probability of selecting the corrective feedback showed a large increase in selecting the corrective feedback as an answer for the related items in Experiment 2 (M = .47, SD = .29) compared to Experiment 1 (M = .24, SD = .20), $t(147) = 6.00$, $p < .001$, d = .93. As in Experiment 1, MC participants overwhelmingly preferred the corrective feedback when responses to both related questions were incorrect, even more in Experiment 2 (M = .50, SD = .45) than Experiment 1 (M = .30, SD = .32), $t(145) = 3.61$, $p < .001$, d = *.56*. These findings are quite

surprising in that the ranking format likely fostered at least somewhat more elaborative retrieval during the practice test than standard testing, which should have allowed participants to use controlled processes to counter automatic influences on the second test. One possible reason for this result might be that ranking caused participants to focus on encoding the corrective feedback, perhaps because of enhanced curiosity (e.g., see Potts et al., 2019), whilst mostly ignoring the question stem. Consequently, when the related questions were shown during the second test, the familiar corrective feedback was highly available in memory, but the fact that questions had changed was not noticed, resulting in multiple selections of the corrective feedback option.

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

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

Theoretical Account of the Results

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

On the other hand, it may be that no MC testing format would yield good related-question performance on the second test. Some researchers have argued (e.g., Ozuru et al., 2013) that recollection tests such as CR present very limited information and

therefore promote generation of several internal cues related to the desired information in a controlled multi-step retrieval process. On the other hand, recognition tests such as MC present the candidate answers as part of the question. Thus, in many cases, the retrieval process is shifted from generation of internal cues to assist in retrieving the to-be-remembered information to reliance on the relative familiarity of the options.

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

However, although the automatic influences were invariant, we noted earlier that not all familiar options were treated the same in MC testing. Previous answers on incorrectly answered practice questions were as familiar as corrective feedback, but they were generally avoided if answers on both tests were incorrect. This result suggests that the automatic memory influence in the MC group did not take the form of general, undifferentiated activation of all repeated options. Instead, the pattern is more consistent

with participants sometimes believing that related questions were repeated ones (i.e., false recognition), which if true, would warrant selecting the corrective feedback.

However, it is important to emphasize that false recognition of MC questions is not the only form that automatic influences took in these experiments. CR participants in Experiment 1, for example, also tended to erroneously produce the corrective feedback more for related questions than new questions and this tendency was greater than for the MC test (M_{diff} = .12 vs. .07, respectively; see Figure 4). This tendency to endorse the feedback for related questions on the CR test occurred even though participants tended to avoid the corrective feedback when they had answered the related question incorrectly during practice. Avoiding the corrective feedback suggests that participants did not believe the question was repeated, mostly likely because they were not misled into this assumption by the presence of repeated response options on the CR test. Additionally, in cases where the second test was MC, Higham et al. (2016) removed repeated items from one of their experiments altogether, and more recently, Alamri and Higham (2022: Paper 3) manipulated the presence/absence of repeated items in a multi-group experiment. In both cases, the researchers observed a slight residual tendency for participants to select the corrective feedback more often for related versus new questions even when there were no items repeated between the tests.

Relationship to the Negative Testing Effect

On the surface, readers might conclude that the poor performance we observed with related questions that we have attributed to automatic memory influences is just another case of the negative testing effect, first reported by Roediger and Marsh (2005). However, the original negative testing effect occurred when participants selected lures during practice, received no feedback on their incorrect selections, and then were

required to answer the same question again in CR format. As a result of not receiving feedback (i.e., they don't know it was a lure), participants recalled the same lures on a later CR test. Importantly, Butler and Roediger (2008) showed that providing feedback during practice reduced the negative testing effect (and increased the positive testing effect), leading them to advise educators to provide feedback during MC practice tests. In contrast, the effect we have examined in our research is *caused* by corrective feedback, not reduced by it. If the questions on the final test were repeated, as in the original negative testing effect scenario, then repeated responding with the feedback would have enhanced performance. However, if related questions are used that have different correct answers, then responding with the corrective feedback on the final test harms performance rather than enhancing it. Thus, with the original negative testing effect, the problem is repeating uncorrected errors. In contrast, the effect we have investigated is caused by responding with feedback when it is no longer appropriate to do so, which reveals a dark side to corrective feedback not previously explored.

Conclusions

Practice answering MC questions can produce robust testing effects. Indeed, Yang et al.'s (2021) recent meta-analysis suggests that MC practice tests produce larger testing effects than practice tests involving recall, at least in real educational contexts. Therefore, if used properly, MC tests can be an effective learning tool. However, there are also pitfalls to avoid. Roediger and Marsh (2005) have reported that in retrieval practice scenarios where question stems are repeated between an MC practice test and a CR final test, previously selected lures can intrude on the CR test (negative testing effect). Our research demonstrates that additional problems can arise when both tests are MC and related questions are used. If MC tests with related questions (that have different correct

answers) are to be used for both practice and final testing, we suggest ensuring that no option is repeated across the tests, particularly if feedback is provided during practice. Using a CR final test will also limit the problem, particularly if deep retrieval is required during the MC practice. Indeed, deep retrieval during practice can sometimes facilitate controlled processing and lead to positive testing effects, as both Little and colleagues (e.g., Little et al., 2012) and our Experiments 2 and 3 have shown. However, even under those circumstances, automatic influences can cause corrective feedback to intrude on the CR final test. Thus, designing related questions should be managed very carefully. We hope that our research will help in this regard by highlighting the kinds of situations that give rise to problems so that they can be avoided.

Paper 1 - Tables

Table 1

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

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

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

Table 2

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

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

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

Table 3

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

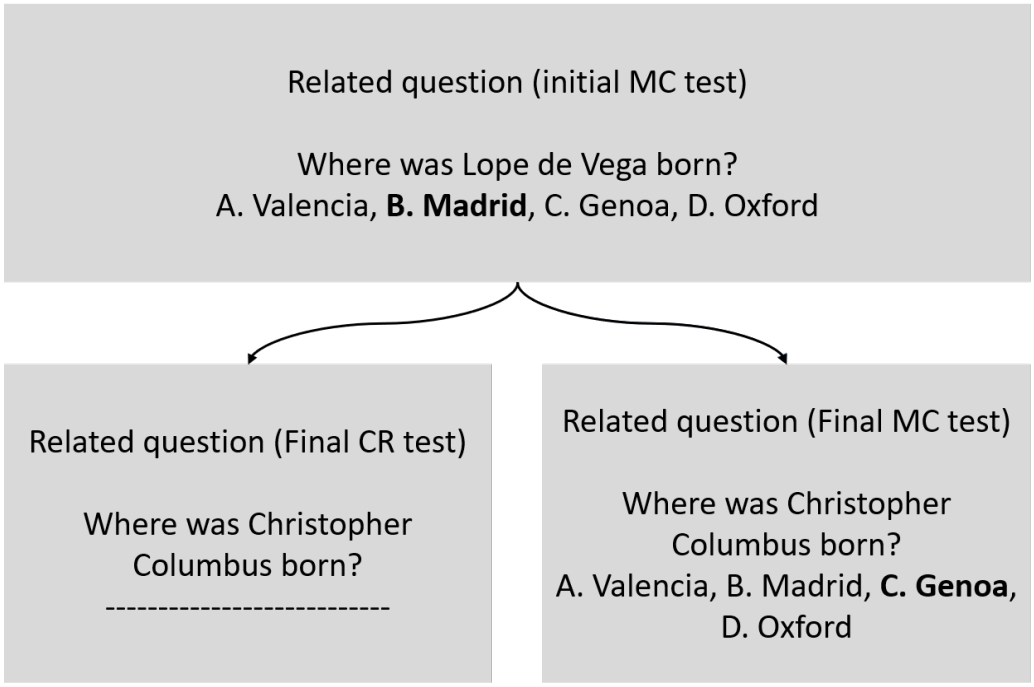
Answer type in the final test	MC	CR
Correct		
Previous answer	0.18 (.20)	0.21 (.31)
Other	0.15 (.24)	0.19 (.24)
Incorrect		
Previous answer	0.11 (.21)	0.04 (.10)
Corrective feedback	0.43 (.48)	0.14 (.24)
Other	0.14 (.21)	0.42 (.42)

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

Paper 1 - Figures

Figure 1

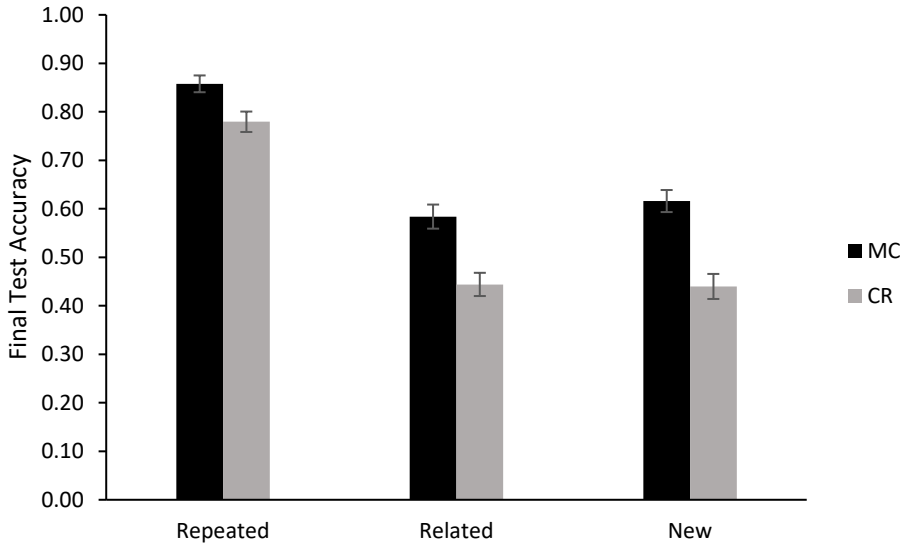
Example of Related Items Presented in Two Final Test Formats.



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

Figure 2

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



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

Figure 3

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

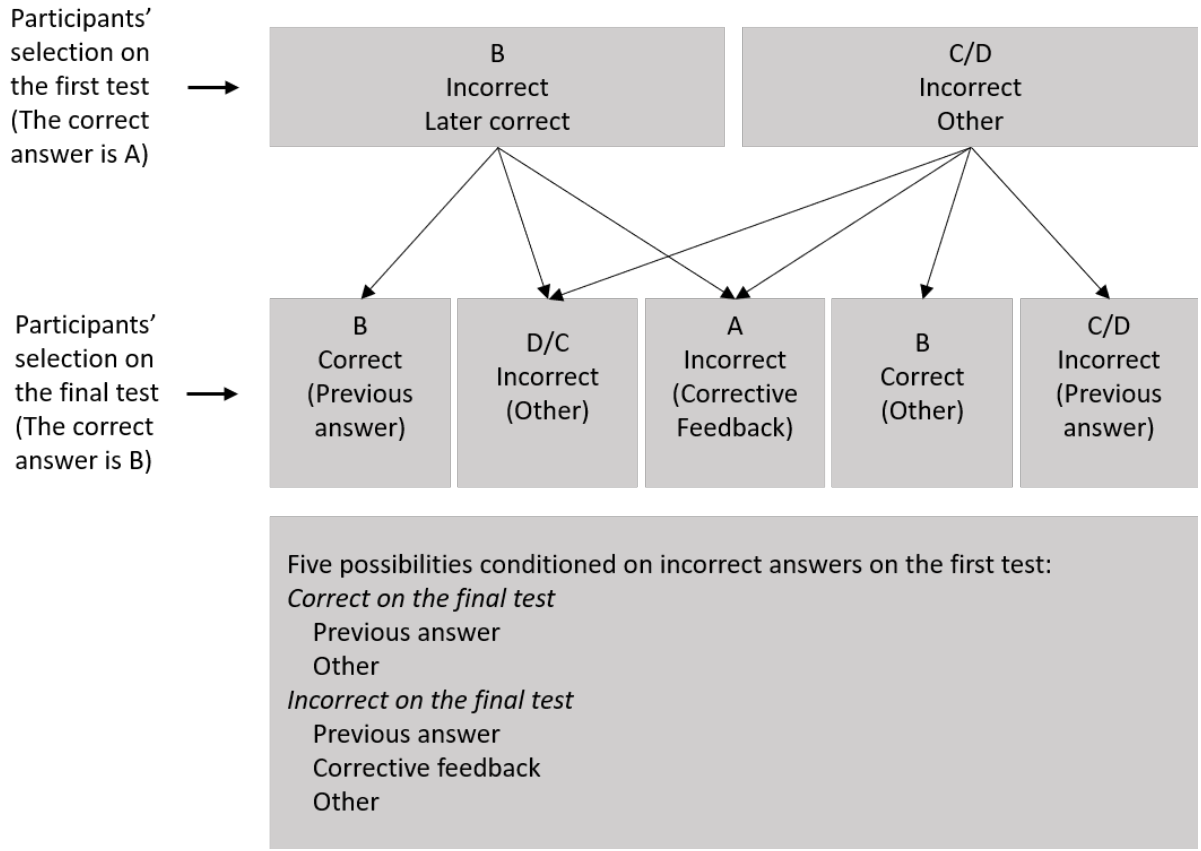
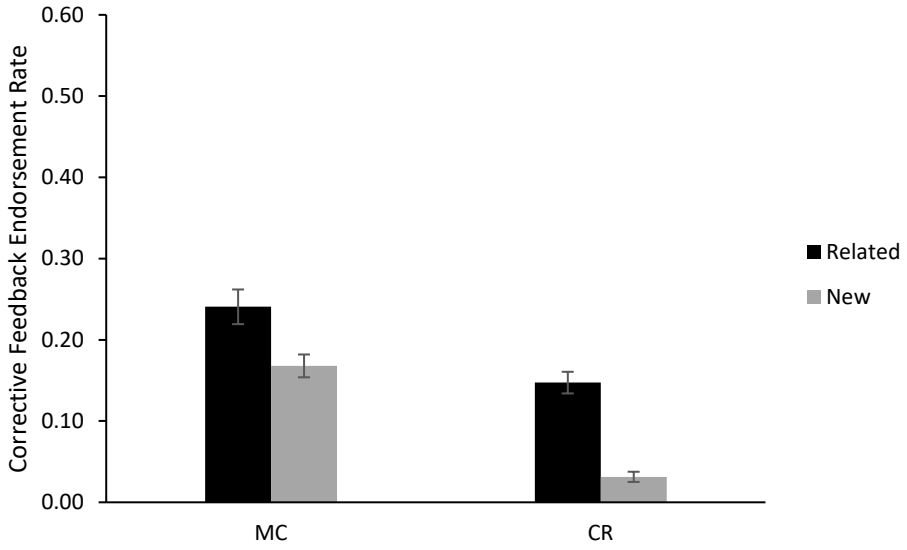


Figure 4

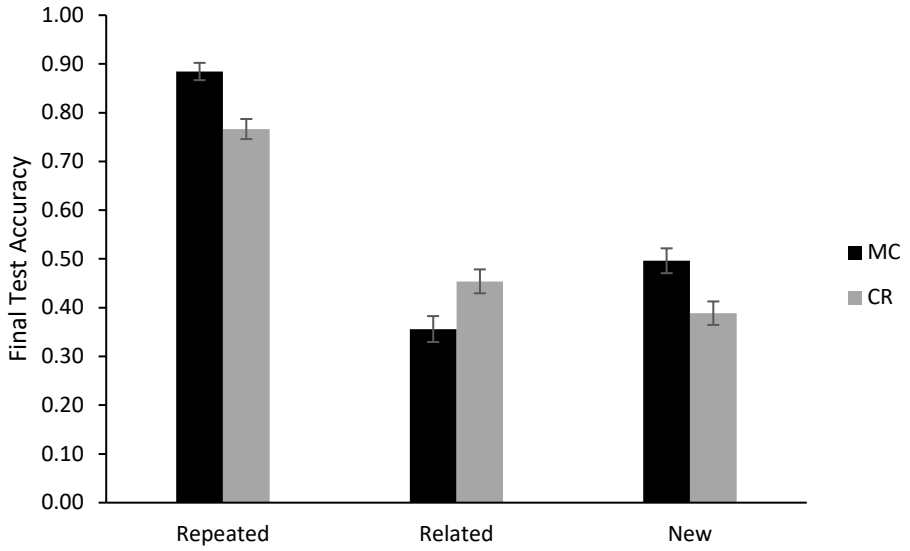
Mean Proportion of Endorsing the Corrective Feedback for Each Question Type Broken Down by Final Test Type in Experiment 1.



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

Figure 5

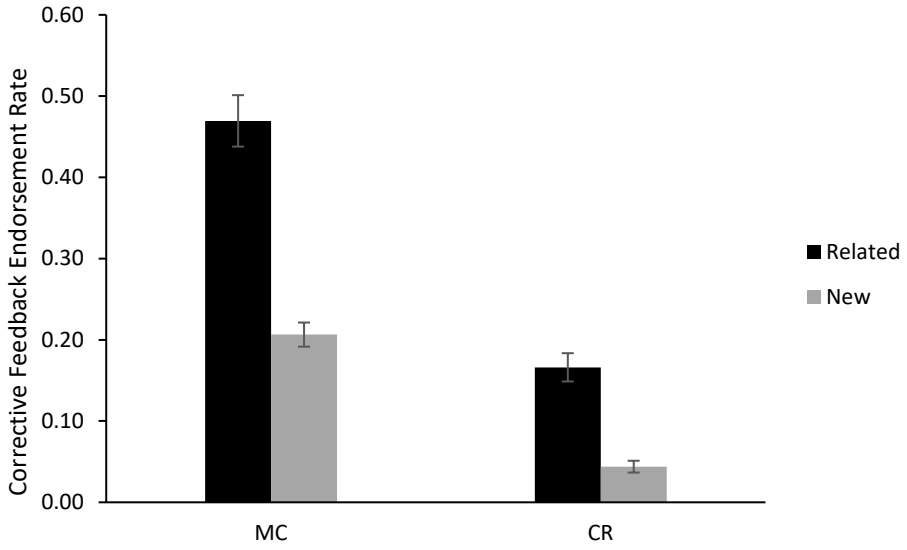
Mean Proportion Correct for Each Question Type Broken Down by Final Test Type in Experiment 2.



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

Figure 6

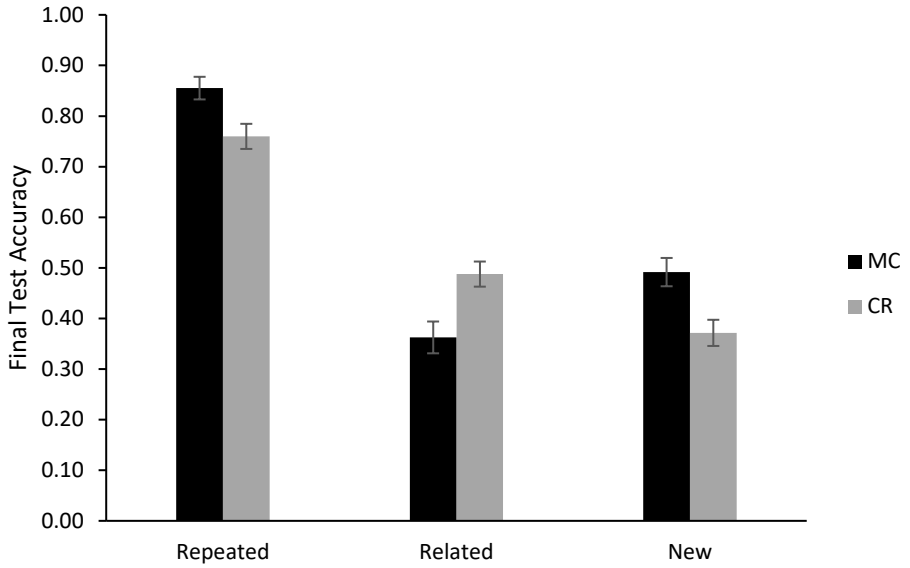
Mean Proportion of Endorsing the Corrective Feedback for Each Question Type Broken Down by Final Test Type in Experiment 2.



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

Figure 7

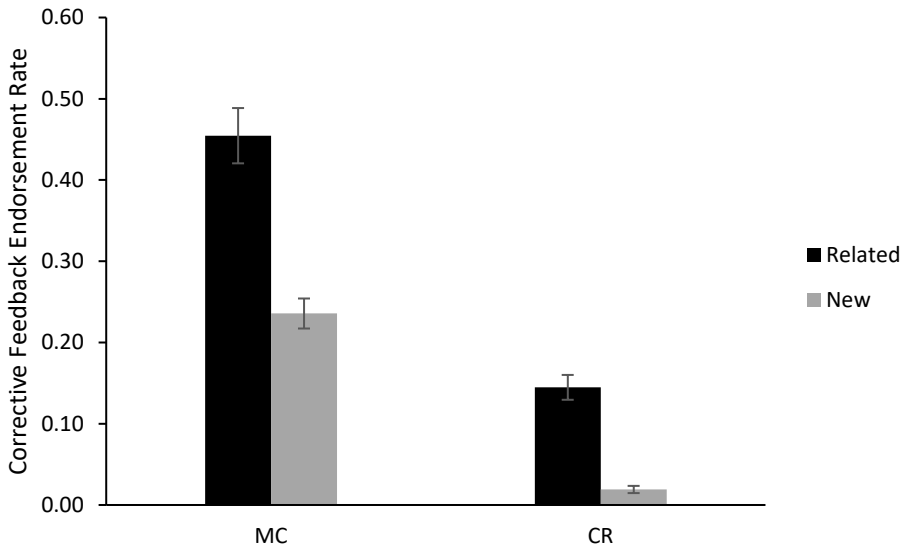
Mean Proportion Correct for Each Question Type Broken Down by Final Test Type in Experiment 3.



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

Figure 8

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



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

Paper 2 [Automatic Influences of Retrieval Practice: The Roles of Feedback, False Recognition, and Opposition Instructions]

Abstract

Multiple-choice (MC) practice tests can enhance later test performance, especially when accompanied by corrective feedback. However, feedback can sometimes be erroneously endorsed if MC questions are related to practice questions but have different answers. This study investigated the role of corrective feedback and false recognition in producing this impairment. Experiment 1 showed that providing feedback on the practice test impaired later related-item performance more than when no feedback was provided. Experiment 2 showed that related-item impairment only occurred when participants believed the related questions were repeated. Experiment 2 also ascertained the nature of the mechanism driving the impairment with opposition instructions. Specifically, participants were told that there were questions related to those on the practice test, but that the answers were never the same. These instructions had minimal effect on performance, suggesting that the impairment is caused by automatic false recognition.

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

Automatic Influences of Retrieval Practice: The Roles of Feedback, False Recognition, and Opposition Instructions

Multiple-choice (MC) testing is a testing format widely used in a variety of contexts. Many educators use MC questions as the main method to assess learners' performance (see Butler, 2018, for a review on using MC testing as a method of assessment). This widespread use is due to advantages such as objectivity when marking learners' answers compared to other testing formats, like open-ended questions (Rauschert et al., 2019). However, some researchers (e.g., Dubins et al., 2016) have criticized this test format because of certain drawbacks. For example, guessing can inflate learners' test scores, which is not a reflection of examinees' level of knowledge.

Positive and Negative Testing Effects

Despite the disadvantages of using MC testing as an assessment method, many studies have shown the benefits of using this test format as a learning tool; that is, taking an MC practice test can enhance later test performance, which is known as the *testing effect* (see Adesope et al., 2017, for a meta-analysis of testing effects). Many studies have reported the positive consequences of taking an initial MC test on later cued recall (CR) performance. For example, Fazio et al. (2010) tested various schedules of retrieval practice using both MC and CR tests. In two schedules of current interest, participants first read expository texts, took an initial MC test immediately, and then either took a final CR test immediately or after one week delay. The CR test contained some previously tested items and new (control) items. The findings showed better performance on the tested items compared to the new items in both the immediate and delayed conditions. These results are consistent with a large literature that has demonstrated a positive

testing effect of taking an MC practice test on retention (e.g., Cantor et al., 2014; Marsh et al., 2012; McDaniel et al., 2012; Yang et al., 2021).

Although there are benefits to MC testing, there can also be negative effects. Specifically, answering MC practice questions exposes participants to misinformation in the form of lures. Roediger and Marsh (2005) examined this issue by having participants take an MC practice test without corrective feedback after reading some text passages. The number of alternatives presented with each question ranged from two to six. Then participants took a final CR test consisting of some repeated questions and new control questions. Although the findings showed a positive effect of MC testing such that tested items had better performance than the new items, a negative effect was observed as well. Specifically, participants tended to intrude with lures from previously tested questions on the CR test, particularly if the number of lures was high and if the lures had been previously selected. A later study by Butler and Roediger (2008) found that providing corrective feedback during the initial test helped to limit the negative testing effect and increase the positive effect.

Positive Effects of MC Testing on Related Items

The positive effect of answering MC questions on later tests extends beyond tested items. For example, Little, Bjork, and colleagues (e.g., Little & Bjork, 2015, 2016: Little et al., 2012, 2019; Sparck et al., 2016) have investigated CR memory performance with untested items that are related to previously answered MC questions. Related items in this context were a pair of questions where a lure from the practice MC question was the correct answer for a CR question on a similar topic. For example, the related question on the initial MC test might be "*Where was Lope de Vega born? a. Valencia,* ***b. Madrid****, c.*

Genoa, d. Oxford" (boldface indicates the correct answer), while the second related question on the final CR test was "*Where was Christopher Columbus born?* ***Genoa***".

To investigate the consequences of initial MC testing on related items, Little et al. (2012) tested two groups of participants who took an initial MC test after reading expository texts. One group was provided with corrective feedback after answering the question, whereas the other group was not. To boost the retrieval process during the initial test, the researchers constructed the MC questions with competitive alternatives (e.g., *How many inches long is an average ferret tail? a. 7-10, b. 20,* ***c. 5***). The final CR test comprised three types of questions: previously tested (repeated) questions, related untested questions, and new control questions. The results revealed a standard testing effect in that performance on the tested items was better than on the new items. More critically, the related items also showed better performance compared to the new items regardless of whether feedback was provided during the first test. Little et al. (2012) suggested that plausible alternatives on the first test provoked retrieval of information about the questions and the alternatives. This retrieval of information about the topic at hand could, in turn, be used to facilitate CR performance with the related questions on the later test.

Later studies have demonstrated similar results as long as the MC format used during practice triggers deep retrieval of information (i.e., lure elaboration). For instance, Little et al. (2019; see also Sparck et al., 2016) designed an elimination MC initial test that tested general knowledge. Elimination testing required participants to choose the correct answer and provide reasons for eliminating the unchosen alternatives. The results of the final CR test showed that participants had more accurate answers on the related questions than the new ones although no corrective feedback was provided during the

initial test. Additionally, questions were answered more accurately on the final test if information was provided to reject the lures during the first test compared to when no information was provided about the lures. The authors concluded that answering MC questions with an elimination format promoted lure elaboration which enhanced later learning. These findings are consistent with several other studies demonstrating the effectiveness of answering MC questions on later performance with related items (e.g., Little & Bjork, 2015, 2016).

Negative Effects of MC Testing on Related Items

The previous studies employed a CR format in the final test and demonstrated facilitation on related items. In contrast, other studies using MC questions in the final test have shown impairment on related items. For example, Higham et al. (2016) investigated the effect of MC initial testing on the final MC test for related items. After reading expository texts, participants took an MC initial test with feedback provided after each question. Subsequently, participants took a final MC test that included some repeated questions, related untested questions, and new control questions. The related questions were like those used in the previous research described earlier that used CR final tests; that is, related questions queried the same topics and a lure from the initial test question was the correct answer for the corresponding question on the final test. In most of their experiments, the MC options were the same for the two related questions. Contrary to the Little, Bjork, and colleagues' results, there was impaired accuracy on the related items compared to the new items.

To investigate the cause of the impairment on related items, Higham et al. (2016) conducted an analysis of participants' responses when they were incorrect on both the first and the final tests. The results of this analysis demonstrated that there was a greater

tendency to select corrective feedback (from the first test) as the answer on the final test compared to other answer possibilities (i.e., previous answers or other options). The effect persisted even when the number of alternatives that matched between the related questions on the initial and final test was reduced from four to one (i.e., the only matching option was the corrective feedback on the first test).

A potentially important difference between studies that have shown related-item facilitation versus impairment is the format of the final test. Throughout their research, Little, Bjork, and colleagues have administered CR final tests whereas Higham et al. (2016) administered MC final tests. To investigate the role that the final-test format might play in producing different outcomes, Alamri and Higham (2022: Paper 1) included both types of final tests in three experiments. The experiments differed primarily in the depth of retrieval that was encouraged during the initial MC test. Standard MC testing (i.e., choose a single favourite answer) was used in Experiment 1. In Experiment 2, participants were required to rank the alternatives from most to least favourite. Finally, in Experiment 3, participants provided reasons for rejecting alternatives (Little et al., 2019). The findings showed that shallow retrieval during the initial test (standard MC) in Experiment 1 resulted in no MC impairment or CR facilitation. However, when the depth of retrieval was increased in Experiments 2 and 3 by using ranking or elimination format, respectively, there was impaired performance with related items (vs. new) on the final MC test. In contrast, there was facilitated performance with related questions (vs. new) on the final CR test. Analysing responses on the final test showed that participants endorsed the corrective feedback on the related items more than the new ones for both

test formats. However, feedback endorsement was larger for the MC test compared to the CR test [1].

Theoretical Mechanisms

Alamri and Higham (2022: Paper 1) attributed the contrasting results between MC and CR tests to opposing automatic versus controlled processes. That is, the enhanced related-item accuracy with the CR test was due to a controlled influence based on consciously recollected information about the question and lures when eliminating them in the initial test. This consciously retrieved information could be used later to facilitate performance with related questions on the final CR test. Conversely, they attributed the impairment observed on the MC test to an automatic influence of retrieval practice which resulted in participants erroneously choosing the corrective feedback at a high rate on related questions. Although the CR results also showed a tendency to endorse corrective feedback to a greater extent for related questions (vs. new), the controlled influence overshadowed the automatic influence on CR tests, resulting in overall accuracy enhancement. However, because the question options including the corrective feedback were explicitly presented again on the final MC test, the automatic influence overshadowed any controlled influence of memory.

Ozuru et al. (2013) differentiated between recollection tests, such as CR, that required learners to generate the answers, and recognition tests, such as MC, that present the answers as part of the questions. Thus, answering related questions on a final CR test may provoke more information retrieval to generate the answer, which is facilitated by controlled influences of memory. In contrast, MC questions might offer a

[1] Because items were counterbalanced between related and new items, this comparison was possible.

shortcut for learners to select the answer they are most familiar with rather than engaging in effortful recollection. Indeed, Chan and McDermott (2007) found that the testing effect was always observed in studies that employed CR questions in the final test but not when using the MC format. As discussed earlier, many studies have investigated controlled influences leading to performance enhancement with related items on final CR tests. In contrast, we believe that further investigation is needed to learn more about the nature of the automatic influence on MC final tests as reported by Higham et al. (2016) and Alamri and Higham (2022: Paper 1).

Current Study

The overarching aim of the current research is to clarify the mechanisms driving impairment on MC final tests following retrieval practice with related items an MC initial test. To achieve this goal, we conducted two experiments where participants took initial and final tests that were both in MC format. In Experiment 1, we compared the performance of MC related items between two groups who took the initial test either with or without corrective feedback. Alamri and Higham (2022: Paper 1) noted that not all options on related questions were chosen at the same rate. For example, if participants answered the initial test question incorrectly, they tended to avoid their previous answer and instead erroneously endorse the corrective feedback. However, it is not currently clear how participants might behave if no feedback is given at all. Although Little et al. (2012) found that feedback had little effect on the facilitation they observed with related questions (vs. new) in their research, their final test was CR not MC. Conceivably, without feedback, the impairment with related final-test MC questions (vs. new) might be reduced or even reversed (i.e., related > new). A reversal might occur if removing

corrective feedback causes automatic influences to be reduced to the point that facilitative controlled influences are able to be expressed.

The main aim of Experiment 2 was to eliminate an alternative account of the impairment with related items. Some critics may take exception to Alamri and Higham's (2022: Paper 1) argument that the impairment they observed with MC final tests was due to an automatic influence. For example, one might consider participants' tendency to select corrective feedback on the final MC test as resulting from the application of a conscious, deliberate strategy rather than the automatic influence. That is, participants may have realised that the questions were related and strategically and deliberately selected the corrective feedback from the earlier question as it was the only option that seemed viable. To distinguish between these possibilities, we used opposition instructions for one condition in Experiment 2 (e.g., Jacoby et al., 1989). These instructions put automatic influences in opposition to controlled influences (e.g., see Jacoby, 1999). Such instructions can provide evidence of whether the impairment with the related items on the MC final test is due to a controlled strategic effect or an automatic memory influence.

A second aim of Experiment 2 was to clarify the nature of the mechanism driving the impairment. Alamri and Higham (2022: Paper 1) argued that repeating the corrective feedback and the other alternatives between the initial and final MC tests may have increased the similarity of the related questions, potentially leading to false recognition (i.e., participants believe the related questions are repeated rather than different questions). If so, that would explain why the corrective feedback was preferred to participants' previous response on the final MC test when the initial test response was wrong. To investigate this possibility, in addition to requiring participants to answer the

MC questions on the final test, we also asked them to indicate whether the question was "old" (repeated from the practice test) or "new" (novel question).

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

Experiment 1

The first experiment examined the consequences of providing versus withholding corrective feedback during the first test on the performance with related items. Participants took an initial MC ranking test where one of two groups was provided with corrective feedback after each question, while the other was not. Following a filler task, all participants took a final MC test in a standard format, which included some repeated questions, related questions, and new questions. Little et al. (2012) found that regardless of whether corrective feedback was provided, taking an initial MC test enhanced performance on the later CR test. However, the role of feedback on impairment with MC final tests has not been tested. We predicted that performance with the related items would be more impaired when providing corrective feedback compared to no-feedback. For the no-feedback group, one possible outcome is that removing feedback would allow controlled processes that had been previously overshadowed by automatic influences to facilitate later performance.

Method

Participants

Across Alamri and Higham's (2022: Paper 1) experiments, the lowest effect size for the interference effect for related MC items (i.e., related items < new items on the final test) was Cohen's f = .24. Based on this result, we conducted 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 were needed. We actually tested 162 participants; however, five participants were excluded after reviewing their performance and responses to attention-check questions. For example, some participants did not rank any question during the first test as instructed. Our final sample involved the remaining 157 participants (female = 63), with ages ranging between 18 and 60 years (*M* = 35.25, *SD* = 9.63) 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 included two groups with 78 participants in the feedback group and 79 in the no-feedback group. Participants in each group were randomly assigned to 12 counterbalancing formats with 6-8 participants each.

Design

The experiment employed a mixed 2 x 3 mixed factorial design with feedback type on the first test (feedback, no-feedback) manipulated between subjects, and question type on the final test (new, repeated, related-but-untested) manipulated within subjects. The dependent variable was the participants' mean performance on the final MC test. We developed 12 surveys using Qualtrics survey software to rotate the questions through the conditions across participants to eliminate item effects. Twenty-two questions were presented on the initial test and 33 questions on the final test. The 33 questions on the final test consisted of 11 repeated (tested) questions, 11 related (untested) questions, and 11 new (control) questions. Questions were presented in a random order for each participant but the MC alternatives for questions were always presented in the same order.

Materials and Procedure

We used 33 pairs of trivia questions as in Alamri and Higham (2022: Experiments 2 & 3). Most questions were originally acquired from Little et al. (2019). However, due to ceiling and floor effects, Alamri and Higham (2022: Paper 1) replaced nine questions (either both members of a pair or only one) with new ones that covered a variety of different general knowledge topics (e.g., mythology, literature, science, history, geography).

The experiment was conducted online utilizing Qualtrics survey software, and two groups were fully instructed on the procedure prior to taking the survey. After reading the instructions, participants answered two attention-check questions about the instructions to ensure that they read and understood them. The experiment started with taking a 22-item initial MC test for about seven minutes. To ensure a reasonable level of encoding, the initial test was presented in a format that required participants to rank the options. For each question in the initial test, a question stem was presented with four alternatives below it as well as four ranking boxes. The boxes were labelled on a scale from *definitely true, probably true, probably false,* to *definitely false*. Participants were required to drag-and-drop the alternatives into the boxes and they were instructed that only one answer was expected in the *definitely true* box which represented the chosen option. Also, to discourage participants from completing the initial test too quickly, participants were not permitted to advance to the next question until 15 s had elapsed. By clicking "Next," participants in the feedback group received corrective feedback. The feedback was provided regardless of whether the chosen option was correct or not (e.g., "*The correct answer is Oslo*"). For the no-feedback group, however, participants were not provided with feedback and instead proceeded to the next question. After completing the initial test, participants engaged in a filler task which involved answering basic mathematics questions for about four minutes.

Following the filler task, both groups took the final MC test for about 10 minutes. The final test contained 33 questions which were divided into three categories: (a) 11 new questions – which acted as the control questions – that were untested and unrelated to the questions from the initial test; (b) 11 repeated questions that had been tested already in the initial test, which was half of the 22 questions that were presented in the initial test; and (c) 11 related untested questions that had not been tested themselves, but were related to previously tested questions (related to the other, non-repeated half of 22 questions that appeared in the initial test). The format of the final test was standard MC where participants were required to select one single answer by clicking the radio button that appeared next to it and then clicked "Next" to move to the next question. No feedback was provided for either group in the final test. The final test was self-paced and the whole experiment took each participant approximately 20-25 minutes to complete.

For the scoring, the initial ranking test was scored as either correct (1) if the correct answer was placed in the "definitely true" box, or wrong (0) if the correct answer was ranked in any of the other three boxes. For the final MC test, the questions were scored as correct (1) if the correct option was chosen and incorrect (0) otherwise.

Results

Initial Test Performance

For the feedback group, the mean proportion of items that participants answered correctly on the first test was .59 (*SD* = .22), while it was .63 (*SD* = .23) for the no-feedback group. An independent samples *t*-test indicated that the difference was not significant, $t(155) = -1.06$, $p = .29$, $d = .17$, suggesting that there was no evidence that the feedback and no-feedback groups were different.

Final Test Performance

Final-Test Accuracy. We conducted a 2 (feedback type: feedback, no-feedback) x 3 (question type: new, repeated, related) mixed-factor Analysis of Variance (ANOVA) with final test accuracy as the dependent variable (see Figure 1). The main effect of feedback type was not significant $F(1, 155) = 1.91$, $p = .17$, $\eta_p^2 = .01$, such that there was significant difference in accuracy between the feedback ($M = .58$, $SD = .17$) and no-feedback ($M = .54$, $SD = .20$) groups. However, there was a significant main effect of question type, $F(2, 310) = 171.41$, $p < .001$, $\eta_p^2 = .53$. A paired-sample *t*-test revealed that accuracy was higher on the new ($M = .55$, $SD = .24$) versus the related items ($M = .41$, $SD = .24$), $t(156) = -7.36$, $p < .001$, $d = .59$, and significantly higher on the repeated items ($M = .74$, $SD = .24$) compared to both the new $t(156) = -10.06$, $p < .001$, $d = .81$, and related items $t(156) = -13.50$, $p < .001$, $d = 1.39$. The main effect of question type was qualified by a significant interaction between feedback type and question type, $F(2, 310) = 58.15$, $p < .001$, $\eta_p^2 = .18$. The interaction was due to both groups performing better on the repeated items compared to new items and on the new items than the related items. However, the differences were larger in the feedback group than the no-feedback group. Specifically, the feedback group performed better on the new ($M = .55$, $SD = .23$), than related items ($M = .34$, $SD = .23$), $t(77) = -7.71$, $p < .001$, $d = .92$, and significantly better on the repeated items ($M = .86$, $SD = .18$) compared to related, $t(77) = -17.67$, $p < .001$, $d = 2.53$, and new items, $t(77) = -13.58$, $p < .001$, $d = 1.49$. Similarly, participants in the no-feedback group performed better on the new items ($M = .54$, $SD = .24$) than the related items ($M = .47$, $SD = .23$), $t(78) = -2.84$, $p < .01$, $d = .28$, and significantly better on the repeated items ($M =$

.62, *SD* = .23), compared to the new $t(78) = -3.10$, $p < .01$, $d = .32$, and related items, $t(78) = -5.76$, $p < .001$, $d = .62$ [2].

Final-Test Answer Types. Prior studies have found that if participants answered the related questions incorrectly on the initial MC test and corrective feedback was provided, they tended to select the corrective feedback for the second pair on the final MC test more than any other answer possibilities (e.g., Alamri & Higham, 2022: Paper 1; Higham et al., 2016). Also, participants tended to avoid their original incorrect selection, despite it likely being highly familiar. We conducted the same analysis to find out whether there was a similar pattern in the current study. Specifically, we calculated the probability of choosing all possible answer types on the final test conditioned on their initial responses being incorrect (see Figure 2 for an illustration). We limited the analysis to incorrect answers on the first test so that we could compare participants' tendency to select their previous answers on the second test versus corrective feedback. Those two alternative possibilities could not be distinguished if the initial test response was correct. Although one of the two groups did not receive corrective feedback, we kept the comparison consistent between the two groups to determine how withholding the corrective feedback affected participants' responses on the final test. In other words, for the no-feedback group, the "corrective feedback" mean corresponds to the likelihood that participants selected the option that would have served as the corrective feedback in the feedback group.

There were five mutually exclusive possibilities of the final test when performance was conditioned on being incorrect in the first test (see Table 1): (a) a correct answer on

[2] Note that there is no adjustment of alpha for the multiple follow-up t-tests in this analysis or the other analyses in this paper.

the second test that matched the previous answer on the initial test (correct/previous answer); (b) a correct answer that was neither the previous answer nor the corrective feedback on the first test (correct/other); (c) an incorrect answer that matched the previous answer on the initial test (incorrect/previous answer) (d) an incorrect answer that matched the corrective feedback that was provided in the first test (incorrect/corrective feedback); and (e) an incorrect answer that was neither the previous answer nor the corrective feedback on the first test (incorrect/other). To compare endorsements of previous answers versus corrective feedback while holding constant the level of accuracy on both tests, we limited our analysis to answers that were incorrect on both tests.

We conducted 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 answers as the dependent variable. We found a significant main effect of feedback type, $F(1, 155) = 8.33$, $p < .01$, $\eta_p^2 = .05$. The probability of answering with one of the three types of answer was higher in the feedback group (M = .72, SD = .49) than the no-feedback group (M = .59, SD = .46). Also, there was a significant main effect of final-test answer type $F(2, 310) = 10.97$, $p < .001$, $\eta_p^2 = .07$. A paired-sample t-test showed that there was no significant difference between the probabilities of answering with the "other" (M = .18, SD = .20) and previous answer (M = .18, SD = .24), $t(156) = -0.14$, $p = .89$, $d = .01$, but both probabilities were lower than that for the corrective feedback (M = .29, SD = .37), $t(156) = 3.41$, $p < .001$, $d = .39$, and $t(156) = 3.04$, $p < .01$, $d = .36$, respectively. Finally, we found a significant interaction between feedback type and question types $F(2, 310) = 39.51$, $p < .001$, $\eta_p^2 = .20$. For the feedback group, a paired-samples t-test revealed that participants overwhelmingly selected the corrective feedback from the initial test (M = .44, SD = .39), compared to "other" answer (M = .14, SD

= .18), $t(77) = 6.18$, $p < .001$, $d = .99$, and previous answer ($M = .13$, $SD = .21$), $t(77) = 6.05$, $p < .001$, $d = .1.01$. However, in contrast to the feedback group, the results indicated that participants in the no-feedback group were more likely to select their previous answers ($M = .24$, $SD = .27$), or "other" answers ($M = .22$, $SD = .22$) compared to corrective feedback ($M = .12$, $SD = .19$), $t(78) = -3.33$, $p < .001$, $d = .51$, and $t(78) = -3.23$, $p < .001$, $d = .47$, respectively. Also, no difference was found between the probability of selecting "other" answer and the previous answer $t(78) = -0.61$, $p = .54$, $d = .08$.

False Endorsements of Corrective Feedback. The previous analysis on the feedback group showed the same pattern as in Alamri and Higham (2022: Paper 1); that is, when participants were provided with corrective feedback in the practice test, they tended to select the corrective feedback in the final test. However, since the analysis was conditioned on providing incorrect answers on both tests, and one of the two groups did not receive corrective feedback in the first test, we conducted another analysis to investigate the rate of endorsing the option associated with the corrective feedback in the feedback group regardless of whether the answers were incorrect on both tests. Also, to exclude the possibility that the corrective feedback was selected merely because it was a plausible option, we compared the probability of endorsing the corrective feedback option between the related and new items for both groups. In other words, the analysis focused on the probability of choosing the particular option that served as corrective feedback for related items in the feedback group. For example, suppose option A was the corrective feedback for the related question X in the feedback group. The likelihood of choosing option A for question X was then computed when participants answered that question with and without feedback, and when it was answered as a related item and as a

new item. The same computation was completed for all questions to produce mean probabilities of selecting the "corrective feedback" for all conditions.

The data were analysed with a 2 (feedback type: feedback, no-feedback) x 2 (question type: new, related) mixed-factor ANOVA with the probability of answering with the corrective feedback option as the dependent variable (see Figure 3). The ANOVA revealed a significant main effect of feedback type, $F(1, 155) = 25.38$, $p < .001$, $\eta_p^2 = .14$. The probability of answering with the corrective feedback was higher in the feedback group ($M = .32$, $SD = .16$) compared to the no-feedback group ($M = .21$, $SD = .12$). We also found a significant main effect of question type, $F(1, 155) = 86.50$, $p < .001$, $\eta_p^2 = .36$. The probability of answering with the corrective feedback was higher for the related items ($M = .35$, $SD = .25$) compared to new items ($M = .18$, $SD = .14$). Finally, we found a significant interaction between feedback type and question type, $F(1, 155) = 45.87$, $p < .001$, $\eta_p^2 = .23$. A paired-sample *t*-test showed that participants in both groups were more likely to endorse the corrective feedback option for related items compared to new items, but the difference was larger in the feedback group (related: $M = .47$, $SD = .26$; new: $M = .17$, $SD = .13$), $t(77) = 9.94$, $p < .001$, $d = 1.47$, compared to the no-feedback group (related: $M = .23$, $SD = .17$; new: $M = .18$, $SD = .14$), $t(78) = 2.18$, $p = .03$, $d = .31$.

Discussion

The findings from Experiment 1 showed that accuracy on the final test was better for repeated items than both related and new untested items, replicating the standard testing effect. In terms of the related versus new item comparison, the final-test accuracy results for both groups replicated those reported by Higham et al. (2016) and Alamri and Higham (2022: Paper 1). That is, related-item accuracy was lower than new-item accuracy. Moreover, analysis of the probability of endorsing corrective feedback

demonstrated that both groups endorsed corrective feedback for related items more than for new items. However, the effect of item type in both of these analyses was qualified by a significant interaction with group, suggesting that the tendency to endorse corrective feedback over the baseline (thereby impairing final-test accuracy) was greater in the feedback group than in the no-feedback group.

A closer examination of the final test answer types for the feedback group showed that when participants answered incorrectly on both tests, they tended to select corrective feedback on the final test. This finding, coupled with participants' tendency to avoid their previous answers on the final test, is consistent with Higham et al. (2016) and Alamri and Higham (2022: Paper 1). In contrast, when no feedback was provided and both test responses were incorrect, the previous answer and "other" options were endorsed about twice as often as the corrective feedback. This result may seem at odds with the analysis on the probability of choosing the corrective feedback, which showed that even for the no-feedback group, there was a tendency to select the corrective feedback option for the related items more than for the new items. However, this seeming anomaly is likely due to the analysis of answer types being conditioned on incorrect initial-test responses whereas the probability of choosing the feedback option was not. Endorsing the corrective feedback option despite not receiving it earlier as feedback during the first test is likely due to cases where the previous answer and corrective feedback were the same (i.e., the answer on the initial test was correct). If the initial test response was correct, then repeating that previous answer would be incorrect on the final test and would be indistinguishable from erroneous endorsements of the corrective feedback. In other words, participants in both the feedback and no-feedback group appear to have shown similar behaviour: to the extent that they falsely believed that a (related) question was repeated, they selected the option they considered to be most likely to be correct

when it was answered earlier. This option was most likely the corrective feedback in the feedback group or their previous answer (which was also the corrective feedback if the initial response was correct) in the no-feedback group.

That said, participants in the no-feedback group were not always simply repeating their previous answers. The rate of choosing an "other" response was also high – much higher than endorsements of the corrective feedback. One possible reason for this outcome is that in the absence of feedback attracting incorrect responses, participants sometimes tried to answer the final-test questions "from scratch" rather than just repeating their earlier answers. By doing so, the likelihood of participants rejecting the corrective feedback, which was always wrong, might have been enhanced, leading to it being the least favourite option. If this occurred, it suggests that removing the feedback had two effects on related items, both of which improved performance. First, it stopped some participants from making errors because they were seduced by the corrective feedback. Second, it encouraged deeper processing of the final-test question, leading to greater understanding that the corrective feedback option was wrong.

Overall, the pattern of data between the feedback and no-feedback groups is consistent with Alamri and Higham's (2022: Paper 1) hypothesis that a major source of the automatic effect in this paradigm is false recognition of related questions. The automatic influence does not take the form of an undifferentiated gain in familiarity for all response options. Rather, different options are favoured depending on which one the participant believed was most likely to be correct on the previous question.

So far, we have described the impairment on related questions as an automatic influence due to false recognition of related questions. Describing the influence as "automatic" is useful because it distinguishes it from the controlled influence that

involves retrieval of information about the lures. However, there is another possible interpretation of the impairment observed with related questions. When encountering a related question on the final test, participants may be reminded of the related question that was answered earlier, consciously recollect their previous answer and the corrective feedback (if provided), and strategically respond with the option that they believe is most likely to be correct. By this account, participants are fully aware that the questions are different, so false recognition is not the source of the error. Instead, participants may reason that the corrective feedback (or their previous answer if no feedback is given) is the best option. A conscious and deliberate strategy of this sort would produce a pattern of data across the feedback and no-feedback groups that is identical to that observed in Experiment 1. However, it is markedly different in nature from Alamri and Higham's (2022: Paper 1) automatic account. In Experiment 2 we examined these two potential interpretations, in addition to the role of false recognition in answering related items.

Experiment 2

In Experiment 2, in addition to asking participants to answer the questions on both the practice and final test, we asked participants to indicate whether each final test question was "old" (seen earlier on the practice test) or "new" (not seen earlier). We then assessed the accuracy of participants' answers separately for each set of items. If our hypothesis that the impairment we observed for related questions was due to falsely recognising the related questions is correct, then the impairment should be limited to items called "old". Related items called "new", on the other hand, may not show any impairment relative to new items, and may even show facilitation. Facilitation might be observed if both automatic and controlled influences are present on the MC final test in this paradigm, but the former influence, which normally overshadows the latter, is

reduced or eliminated. By focusing solely on related items called “new”, (i.e., items for which the automatic influence is small or not present at all), the controlled influence may be revealed.

In addition to the recognition judgment on the final test, the second major change in Experiment 2 was to manipulate instructions between subjects. One group of participants was given regular instructions as in Experiment 1. However, the other group was given *opposition instructions*. Specifically, participants were warned that there were questions on the final test that were related to earlier questions, but to be careful with these questions because the correct answer to these questions was *never the same* as the correct answer on related questions presented earlier. Such instructions would allow us to determine the nature of the related-item impairment by setting controlled and automatic influences in opposition (e.g., Jacoby et al., 1989). Specifically, if participants are consciously and deliberately electing to respond with the corrective feedback because it seemed to be the best strategy for maximizing accuracy, then opposition instructions should remove that tendency. On the other hand, if the influence is more automatic and based on false recognition, as we have hypothesised, then opposition instructions should have a minimal effect.

Method

Participants

Based on the power analysis in Experiment 1, we aimed to recruit at least 128 participants. We actually recruited 182 participants, but after reviewing participants’ performance and responses to the attention checks questions, four participants were excluded. The exclusion criteria included not engaging in the task by avoiding ranking any question during the first test. Thus, the final analysis involved the remaining 178

participants (female = 64), with ages ranging between 20 to 59 years (*M* = 33.41, *SD* = 7.72) from the general population. They were recruited through Amazon Mechanical Turk (MTurk) and were given $2 in return for their participation in the study. The experiment involved two groups with 88 participants in the regular instructions group and 90 in the opposition instructions group.

Design

This experiment employed a mixed 2 x 3 design with two independent variables: instructions type (regular, opposition) was manipulated between subjects, and question type on the final test (new, repeated, related but untested) was manipulated within subjects. The main dependent variable was the participants mean performance on the final test. As in Experiment 1, 12 counterbalancing formats with 6-11 participants each were used to eliminate item effects.

Materials and Procedure

The materials were the same as those used in Experiment 1. The procedure was mostly identical to that in Experiment 1 with some exceptions. First, the format for ranking the options on the first test was slightly modified. 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 most accurate answer and 4 as the least accurate answer. Second, both groups received corrective feedback after each question in the first test as with the feedback group in Experiment 1. Third, the instructions provided prior to the final test were manipulated. Participants in the regular instructions group were simply told that they were going to take a final test. Conversely, participants in the opposition instructions group were told: "*Now you are going to take the final test. Please note there are some questions on the upcoming test that are related*

to, but not the same as, questions you answered earlier. However, please be careful when answering these questions because although they are related to earlier questions, the correct answers are never the same. Therefore, if you choose the same answer just because a question reminds you of a related (but not identical) question from the first test, your answer will be wrong". To discourage participants from moving forward to the final test without reading the instructions, we presented the instructions for a minimum 15 s before permitting participants to advance to the final test. Finally, to understand the role of false recognition in answering the related items, each question on the final test in both groups was followed by a separate recognition question: "*Did you encounter the previous question in the first test?*" Participants answered this question by selecting "*Yes*" or "*No*" before moving to the next question.

For the scoring, the initial ranking test was scored as either correct (1) if the correct answer was placed in the top position, or wrong (0) if the correct answer was ranked in any of the other three positions (2, 3, 4). For scoring the final MC test, the scoring method was identical to that in Experiment 1. Concerning the recognition questions, they were scored with (1) if the answer was "Yes" and (0) if the answer was "No".

Results

Initial Test Performance

The mean proportion of questions answered correctly on the first test in the opposition instructions and regular instructions groups was .63 (*SD* = .20) and .67 (*SD* = .19), respectively. An independent sample *t*-test indicated that the difference was not significant, $t(176) = -1.15$, $p = .25$, $d = .17$, suggesting that there was no evidence that the opposition instructions and regular instructions groups were different.

Final Test Performance

Final-Test Accuracy. We conducted a 2 (instructions type: regular, opposition) x 3 (question type: new, repeated, related) mixed-factor ANOVA with final test accuracy as the dependent variable (see Figure 4). There was no significant effect of instructions type $F(1,176) = 2.62$, $p = .11$, $\eta_p^2 = .01$. However, there was a significant effect of question type, $F(2, 352) = 260.69$, $p < .001$, $\eta_p^2 = .60$. A paired-sample *t*-test revealed that accuracy was significantly higher for the new items ($M = .55$, $SD = .25$) compared to the related items ($M = .43$, $SD = .27$), $t(177) = -6.89$, $p < .001$, $d = .46$. Also, the accuracy was significantly higher for the repeated items compared to the new and related items ($M = .83$, $SD = .19$), $t(177) = 16.08$, $p < .001$, $d = 1.27$, and $t(177) = 20.44$, $p < .001$, $d = 1.72$, respectively. Finally, there was a marginal interaction between instructions type and question type, $F(2, 352) = 2.59$, $p = .08$, $\eta_p^2 = .01$.

Recognition Judgment Accuracy for the Questions. We conducted a 2 (instructions type: regular, opposition) x 3 (question type: new, repeated, related) mixed-factor ANOVA with the probability of an "old" response as the dependent variable (see Figure 5). There was no significant effect of instructions type $F < 1$. However, we found a significant effect of question type, $F(2, 352) = 381.63$ $p < .001$, $\eta_p^2 = .68$. Paired-sample *t*-tests revealed that participants were more likely to call repeated questions "old" ($M = .84$, $SD = .20$) compared to related items ($M = .41$, $SD = .30$), $t(177) = 15.89$, $p < .001$, $d = 1.71$, and new items ($M = .18$, $SD = .26$), $t(177) = 24.63$, $p < .001$, $d = 2.86$. Also, the probability of calling the questions old was higher for the related items compared to the new items $t(177) = 13.09$, $p < .001$, $d = .80$. Note that "old" responses to repeated items were correct (hits), whereas they were not to related or new item (false alarms). Finally, we found no significant interaction between instructions type and question type, $F < 1$.

Final-Test Accuracy Conditioned on Recognition Judgment Accuracy for the Questions. To investigate whether false recognition was the basis of the accuracy impairment for related items, we analysed the probability of answering the questions on the final test correctly when they were called "old" versus "new" (see Table 2)[3]. For the regular instructions group, paired-sample *t*-tests revealed that when calling questions "old", the probability of answering them correctly was higher for the repeated items (M = .90, SD = .14) than for both the related items (M = .17, SD = .27), $t(81) = 21.06$, $p < .001$, d = 3.45, and the new items (M = .53, SD = .38), $t(35) = 5.37$, $p < .001$, $d = 1.26$. Accuracy was also higher for the new items compared to the related items $t(36) = 4.40$, $p < .001$, d = 1.12. However, when calling the questions "new", accuracy for the repeated items (M = .59, SD = .40) was not significantly different from accuracy for the related items (M = .58, SD = .29), $t(50) = 0.95$, $p = .34$, $d = .03$, or the new items (M = .60, SD = .25), $t(52) = 0.30$, p = .76, $d = .03$. Also, there was no difference in accuracy between the related and new items, $t(82) = 1.06$, $p = .29$, $d = .08$.

In the opposition instructions group, when calling questions "old", participants were more likely to answer the repeated questions correctly (M = .84, SD = .20), compared to the related items (M = .22, SD = .26), $t(79) = 18.01$, $p < .001$, $d = 2.71$, and new items (M = .48, SD = .38), $t(43) = 5.73$, $p < .001$, $d = 1.21$. Also, participants had more accurate answers on the new items versus the related items $t(43) = 2.77$, $p < .01$, $d = .81$. When questions were called "new", however, we found no difference in accuracy between the repeated items (M = .54, SD = .41) and the related items (M = .51, SD = .32), $t(53) = 0.89$, $p = .37$, $d = .07$, or the new items (M = .51, SD = .27), $t(56) = 0.64$, $p = .52$, d =

[3] Due to the amount of missing data in this analysis caused by participants calling no items in some conditions "old" or "new", we could not run a mixed-factor ANOVA test similar to what we did with the other analyses or the test would lack power.

.08. Also, we found no difference in accuracy between the related and new items $t(84) = 0.51$, $p = .60$, $d = .00$.

To directly investigate the effect of recognition on the performance of related items, we conducted another analysis comparing participants' performance on related items when they were recognised correctly versus when they were not. For the regular instructions group, paired-sample *t*-tests revealed that when calling related questions "new", the probability of answering them correctly was higher than when calling them "old" $t(77) = -9.25$, $p < .001$, $d = 1.45$. Similarly, participants in the opposition instructions group had more accurate answers when calling related items "new" than when calling them "old" $t(75) = -6.84$, $p < .001$, $d = 1.00$.

Final-Test Answer Types. As in Experiment 1, we analysed participants' responses on the final test conditioned on being answered incorrectly in the first test. Then we conducted a 2 (instructions type: regular, opposition) x 3 (final-test answer type: previous answer, corrective feedback, other) mixed-factor ANOVA with the probability of answering with each type of incorrect answer on the final test as the dependent variable (see Table 3). There was no significant effect of instructions type, $F < 1$. However, there was a significant effect of question type, $F(2, 352) = 52.17$, $p < .001$, $\eta_p^2 = .23$. Paired-sample *t*-tests revealed that the probability of answering with the corrective feedback ($M = .40$, $SD = .42$) was higher than both "other" answer ($M = .16$, $SD = .24$), $t(177) = 6.96$, $p < .001$, $d = .72$, and previous answer ($M = .12$, $SD = .18$), $t(177) = 8.60$, $p < .001$, $d = .89$. Finally, we found no significant interaction between instructions type and question type, $F(2, 352) = 1.72$, $p = .18$, $\eta_p^2 = .01$.

False Endorsements of Corrective Feedback. As in Experiment 1, we analysed the probability of endorsing the corrective feedback between the related and new items. To

do so, we conducted a 2 (instructions type: regular, opposition) x 2 (question type: new, related) mixed-factor ANOVA with the probability of answering with the corrective feedback as the dependent variable (see Figure 6). The ANOVA revealed no significant main effect of instructions type, $F < 1$. However, there was a significant main effect of question type, $F(1, 176) = 99.89$, $p < .001$, $\eta_p^2 = .36$. The probability of answering with the corrective feedback was higher for the related items ($M = .37$, $SD = .27$) compared to new items ($M = .18$, $SD = .16$). Finally, we found a significant interaction between instruction type and question type, $F(1, 176) = 5.56$, $p = .02$, $\eta_p^2 = .03$. A paired-sample *t*-test revealed that participants in both groups were more likely to endorse the corrective feedback for related items compared to new items, but the difference was larger in the regular instructions group (related: $M = .39$, $SD = .26$; new: $M = .15$, $SD = .14$), $t(87) = 9.03$, $p < .001$, $d = 1.13$, compared to the opposition instructions group (related: $M = .36$, $SD = .28$; new: $M = .21$, $SD = .18$), $t(89) = 5.26$, $p < .001$, $d = .63$.

False Endorsements of Corrective Feedback Conditioned on Recognition Judgment Accuracy. To find out whether the endorsement rate of choosing the corrective feedback for related versus new questions differed based on participants' recognition decisions, we analysed the probability of endorsing the corrective feedback conditioned on "old" versus "new" responses (see Table 4)[4]. For the regular instruction group, a paired-sample *t*-test revealed that when calling questions "old", the probability of endorsing the corrective feedback on the related items was higher ($M = .77$, $SD = .30$) compared to the new items ($M = .24$, $SD = .28$), $t(36) = 7.30$, $p < .001$, $d = 1.34$. Also, when calling questions "new", corrective feedback endorsement was higher for related items (*M*

[4] As with our earlier conditional analysis, the data were analysed with *t*-tests rather than ANOVA because of empty cells (i.e., no items in some conditions called "old" or "new").

= .20, *SD* = .28) versus the new items (*M* = .14, *SD* = .14), *t*(82) = 2.17, *p* = .03, *d* = .27. For the opposition instruction group, when calling questions "old", participants endorsed the corrective feedback on the related items (*M* = .69, *SD* = .30), more than the new items (*M* = .27, *SD* = .33), *t*(43) = 5.12, *p* < .001, *d* = 1.83. However, when calling the questions "new", we found no difference between the related items (*M* = .20, *SD* = .28), and the new items (*M* = .20, *SD* = .18), *t*(84) = -0.06, *p* = .94, *d* = .02.

To directly investigate the effect of recognition on the probability of endorsing the corrective feedback, we conducted another analysis comparing the probability of endorsing the corrective feedback on related items when they were recognised correctly versus when they were not. For the regular instructions group, paired-sample *t*-tests revealed that when calling related questions "old", the probability of endorsing the corrective feedback was higher than when calling them "new" *t*(77) = -11.78, *p* < .001, *d* = 1.97. Similarly, participants in the opposition instructions group endorsed the corrective feedback when calling related items "old" more than when calling them "new" *t*(75) = -9.89, *p* < .001, *d* = 1.69.

Discussion

As in Experiment 1, final-test accuracy was better for the repeated items compared to the related and new untested items in both groups. More critically, related-item accuracy was worse than new-item accuracy in both the regular- and opposition-instructions groups. Moreover, participants in both groups tended to select the corrective feedback more for related questions than for new ones. Overall, instruction type (opposition vs. regular) had very little moderating influence on either of these effects. The analyses on final-test accuracy and corrective feedback endorsements suggested that opposition instructions slightly reduced related-new differences in both cases. However,

the former reduction was of only marginal significance, and the effects sizes were very small in both cases (η_p^2 in the range .01 to .03). There were also large residual related-new differences despite the opposition instructions. Finally, analysis of final-test response types when both the practice- and final-test responses were incorrect showed that the corrective feedback was by far the favourite choice, just as it was in the feedback group of Experiment 1, and that this attraction to the corrective feedback did not differ much by instruction type. Together, these results suggest that the impairment is an automatic influence of retrieval practice rather than a consciously controlled strategic one.

In contrast to the negligible effects of opposition instructions, recognition performance strongly moderated the related-new difference. The analysis on final-test accuracy indicated that poor performance with related items in both the regular- and opposition-instructions groups was almost entirely attributable to cases where participants falsely believed items were repeated. An analogous conclusion was suggested by the analysis of the probability of choosing the corrective feedback; if participants believed the related items were repeated, they were two-to-three times more likely to select the corrective feedback compared to baseline. Although there was a small residual related-new difference for items called "new", the effect was only statistically significant in the regular-instructions group, and the effect size was relatively small (Cohen's d = 0.27). Thus, overall, the results of Experiment 2 suggest that the main source of the problem with related items is that participants falsely believe that they are repeated from the first test and that responding with the corrective feedback is therefore warranted.

One final unexpected result is worth noting. Recognition performance did not just moderate the related-new difference in final-test accuracy, it moderated the repeated-

new difference in accuracy as well. In short, there was no testing effect observed if participants falsely believed repeated items were new. This is true even though a liberal definition of the testing effect applies here (repeated vs. new) rather than a conservative one (repeated vs. restudy). This result suggests that a portion of the standard testing effect is at least partly attributable to participants correctly recognising the question and responding with corrective feedback (or possibly a previous response if no feedback is given). Of course, if the items are repeated, this type of responding serves to facilitate final-test performance to produce the testing effect. It is only in cases that the items are related but have different answers, as is the case with the related items used in our experiments, that responding in this way will impair performance.

General Discussion

In this study, we investigated the positive and negative consequences of taking an MC practice test on a second MC test. If questions were repeated between the tests, performance on the second MC test was better than for untested questions, regardless of whether untested questions were related to questions answered previously on the practice test or new questions (i.e., seen only on the second test). This advantage for repeated items was observed even when corrective feedback was absent (Experiment 1) or opposition instructions were given (Experiment 2). These findings are consistent with many studies that have illustrated the positive effect of taking tests as a learning tool (Adesope et al., 2017; Yang et al., 2021).

Related Versus New Questions

The Role of Feedback

Our main aim of the current research was to examine the performance on the related items and how it compared to new-item performance on the final test when both

the initial and final tests were in MC format. In Experiment 1, there was significant impairment to performance with the related items compared to the new items for the feedback group, similar to the results observed by Higham et al. (2016) and Alamri and Higham (2022: Paper 1). Interestingly, when no feedback was provided during the initial test in Experiment 1, the results still showed worse final-test accuracy as well as a tendency to choose the corrective feedback option more for related items than new items. This finding likely resulted from the same mechanism that caused impairment for related questions in the feedback condition: false belief that the related question was repeated, which prompted participants to select the option that was most likely correct earlier. That response was the corrective feedback in the feedback group and the previous response in the no-feedback group.

However, as we noted earlier, the absence of feedback may also have caused participants to consider the final-test questions more closely, thereby reducing the rate of choosing the corrective feedback option (which was always wrong) and increasing the rate of choosing an "other" option (which would sometimes be correct; see Table 1). Nonetheless, a correlational analysis between participants' overall final-test accuracy and the rate of endorsing their previous answer on related questions for the no-feedback group showed a significant inverse correlation $r(77) = -0.53$, $p < .001$, suggesting that the previous answer was an attractive option for participants who were struggling with the final test.

If both the feedback and no-feedback groups were making errors due to false recognition of related questions, then why was there less impairment on related questions in the no-feedback group compared to the feedback group? For example, the no-feedback group showed significantly better performance with the related items when

compared to the feedback group (M = .47, SD = .23; M = .34, SD = .23, respectively), $t(155) = 3.75$, $p < .001$, d = .60. Furthermore, although both groups endorsed the corrective feedback for the related items more than the new items, the endorsement rate difference was higher in the feedback group compared to the no-feedback group (M_{diff} = .30 vs. .05, respectively; Figure 3). In our view, the reason for these differences can be traced to imperfect accuracy on the first test (59-63%). Had their initial-test performance approached 100%, it is conceivable that the feedback and no-feedback groups would have demonstrated an equivalent level of impairment on related items on the final test and a comparable tendency to respond with the feedback option. With an initial-test accuracy rate of 100%, the corrective feedback (feedback group) and the previous answer (no-feedback group) are the same response and guaranteed to yield an error on the final test. However, if some initial-test responses to related questions are incorrect, repeating them on the final test is likely to yield some correct responses (i.e., the incorrect initial-test response is the correct answer to the related final-test question) and a lower rate of feedback endorsement.

Although there are differences in methodology between our paradigm and the paradigm used to investigate the negative testing effect (e.g., MC vs. CR final test; related vs. repeated questions), the finding that participants in the no-feedback group tended to reproduce previous answers that were incorrect is similar to the negative testing effect (Roediger & Marsh, 2005). That is, both findings suggest that selecting lures on an initial MC test without receiving corrective feedback can lead participants to learn those errors and repeat them on later tests. The main difference is that, as long as initial-test accuracy is not at ceiling, reproducing those errors can benefit performance on related questions as used in our research, but it will always impair performance on repeated questions used in negative testing effect research.

In contrast to the large effect of initial-test feedback (vs. no feedback) on final-test performance with related items observed in Experiment 1, Little et al. (2012, Experiment 2) found that feedback produced few differences on the related-new difference (although a large benefit was observed for repeated items). Moreover, the difference was opposite to what we observed in Experiment 1 (i.e., related > new). How might we reconcile these discrepant results? In our view, the answer lies with the nature of the final test. Specifically, our final test was MC, whereas theirs was CR. If the automatic influence that causes the impairment with related questions is based primarily on false recognition, then the presence of retrieval cues at the test will be important. Related MC questions were not only related in terms of the content they queried, they also shared an identical set of options, one of which was the corrective feedback. The matched options and the explicit repetition of the corrective feedback likely acted as a very strong retrieval cue for the earlier encounter with the related question during the practice test, leading to high false recognition rates and a large automatic influence. Indeed, Heist et al. (2014) found that when answering MC questions, most participants, rather than recalling the answer after reading the question, tended to search the alternatives to find the suitable answer. Such a strategy would likely increase the likelihood that the corrective feedback would be considered. In contrast, the absence of matching options and repeated corrective feedback with the CR final test in Little et al.'s research would have limited false recognition and the automatic influence such that the feedback effect was minimal. Instead, controlled influences predominated, producing a benefit, rather than impairment, to performance with related items (vs. new).

Automatic Influence Versus Controlled Strategy

A primary aim of Experiment 2 was to distinguish between two competing accounts of the performance impairment observed with related questions on the final test. The first was originally offered by Alamri and Higham (2022: Paper 1). They argued that answering practice questions and receiving corrective feedback created episodes in memory that were automatically retrieved when participants encountered related questions on the final test. This automatic retrieval prompted participants to choose the same (but now wrong) corrective feedback option. The second account is that the influence is more strategic; that is participants consciously recall the earlier related question and reason that the corrective feedback is likely the best option for the second question as well, particularly if they are unsure of the answer. Indeed, participants might even have realised that the questions were different but inferred that the corrective feedback might be correct for both questions.

The use of opposition instructions (e.g., Jacoby et al., 1989) in Experiment 2 yielded some evidence in favour of the automatic-influences account. In that experiment, we warned participants that the final test contained questions that were similar to ones answered earlier, but that the correct answers to these questions were never the same as the correct answers to the initial-test questions. Such instructions should have substantially weakened any deliberate, strategic responding, thereby reducing the related-new difference, particularly if participants were aware that the questions were different. In contrast, if the influence was more automatic, then related-item impairment should have remained intact despite the opposition instructions. Consistent with the latter account, opposition instructions had very little effect on performance; for the most part, both instruction groups showed equivalent levels of impairment for related (vs. new) questions on the final test regardless of the measure examined. Together, these findings are consistent with many studies that have demonstrated invariance of automatic

influences when automatic and controlled influences are set in opposition (e.g., Higham et al., 2000; Jacoby, 1991, 1999; Jacoby et al., 1993).

That said, we acknowledge that the impairment observed in the performance of the opposition-instructions group might be attributed to other factors which do not reflect automaticity. Specifically, failing to attend to the opposition instructions or forgetting them while answering final-test questions might explain the similar performance observed in the two groups regardless of the type of instructions they received (i.e., opposition vs. regular). Future research might examine the opposition-instructions method under stricter conditions. For example, instead of presenting the instructions a single time before the final test, researchers might consider presenting the instructions multiple times as a reminder prior to answering each question on the final test.

The Role of False Recognition

In Experiment 2, we tested the hypothesis that the source of the automatic influence was false recognition of the related questions. We tested it by asking participants to judge the final-test questions as "old" or "new" after answering them. As expected, the false recognition rate was much higher for related questions than new ones. Moreover, when accuracy was conditioned on the recognition responses, a clear pattern emerged: impairment to related questions accuracy was almost entirely attributable to cases of false recognition. Barring a small residual effect in the analysis of corrective feedback endorsements, there was no difference between the related and new items when they were called "new".

Unexpectedly, a similar pattern emerged when comparing repeated items and new ones. As with the previous analysis, performance differences between these questions

were limited to cases of positive recognition. The suggestion here is that at least some portion of the standard testing effect (i.e., repeated > new) is attributable to the same mechanisms that underpin the impairment with related items (i.e., related < new). That is, participants recognise the repeated questions and produce responses that match the corrective feedback (or previous response if no feedback is provided). We are not aware of any other research bearing directly on the role of question recognition processes as forming the basis of the standard testing effect. In our view, it is a potentially interesting avenue for future research.

Processes and Tasks

Alamri and Higham (2022: Paper 1) argued that although both controlled and automatic influences of retrieval practice occur, it would be a mistake to assume that the former influence is measured solely with CR tests and the latter is measured solely with MC tests. Such a conclusion would constitute a process purity assumption, which equates specific tasks with specific processes, an assumption that has been shown to be problematic in many cases (e.g., Jacoby, 1991). As evidence that such an assumption was unwarranted, Alamri and Higham noted that participants erroneously intruded with the corrective feedback option on related questions at a higher rate than with new questions even when the final test was CR (albeit at a lower rate than with MC final tests). However, Alamri and Higham's studies provided no evidence of the converse, that is, controlled influences on MC final tests. One possible reason for failing to find controlled influences with MC final tests is that automatic influences overshadowed the less pronounced controlled influences. The current Experiment 2 could potentially have revealed controlled effects for related final-test questions that were correctly recognised as "new". In such cases, automatic effects would be at a minimum, potentially allowing less dominant

controlled influences to be expressed. Specifically, correctly recognised related items might have shown facilitation compared to new ones. However, performance on correctly recognised related and new items was remarkably similar, suggesting that controlled influences were negligible in Experiment 2 (Table 2).

Although there was no evidence of controlled influences with MC final tests in the current research, Alamri and Higham (2022: Paper 3) found that when the MC testing stakes are higher, such as with tests administered in real educational environments, controlled influences are observed. Specifically, Alamri and Higham administered an MC test containing some related items to a group of introductory psychology students. The related items had identical options but were separated by different lags. In particular, some related items were separated by several other items, whereas some were presented back-to-back (i.e., in immediate succession). We hypothesised that presenting the related questions in immediate succession in the back-to-back condition would highlight the differences between the questions and the fact that different answers were required. As expected, accuracy on the second related questions was better (vs. new) in the back-to-back condition. However, it was also higher when there was a lag of several intervening items between the related items, although the accuracy benefit was not as great as in the back-to-back condition. Overall, these results indicate that it is possible to observe controlled influences with high stakes MC tests, and to boost them with particular types of sequencing. This will no doubt come as some relief for educators who wish to continue testing their students with MC tests.

Conclusions

The prior literature has consistently shown the positive effect of retrieval practice with MC questions (e.g., McDaniel et al., 2012; see Yang et al., 2021 for a review). More

recently, studies have shown similar benefits of retrieval practice with related untested items, at least when employing the CR format in the final test (e.g., Little et al., 2019). However, research that has examined the effect of retrieval practice on related items when an MC format is used for both the initial and final test has consistently shown a performance deficit, particularly if corrective feedback is provided on the initial test (e.g., Experiments 1 & 2 in the current study; Alamri & Higham, 2022; Higham et al., 2016). Our current results suggest that this impairment was due to the false recognition of the related questions. Furthermore, the impairment was increased by presenting corrective feedback during the first test. Overall, this study sheds light on the nature of the processes underpinning MC testing, as well as the involvement of automatic and controlled influences when taking this type of test.

Paper 2 – Tables

Table 1

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

Answer type on the final test	Feedback	No feedback
Correct		
Previous answer	0.13 (.17)	0.20 (.22)
Other	0.16 (.19)	0.21 (.28)
Incorrect		
Previous answer	0.13 (.21)	0.24 (.27)
Corrective feedback	0.44 (.39)	0.12 (.19)
Other	0.14 (.18)	0.22 (.22)

Table 2

Mean (SD) Proportion of Correct Answers on the Final Test for Different Question Types Conditioned on Being Called "Old" or "New" in Experiment 2

Instruction type	Response	Repeated	Related	New
Regular instruction	"Old"	0.90 (.14)	0.17 (.27)	0.53 (.38)
	"New"	0.59 (.40)	0.58 (.29)	0.60 (.25)
Opposition instruction	"Old"	0.84 (.20)	0.22 (.26)	0.48 (.38)
	"New"	0.54 (.41)	0.51 (.32)	0.51 (.27)

Table 3

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

Answer type on the final test	Regular instructions	Opposition instructions
Correct		
Previous answer	0.15 (.19)	0.15 (.24)
Other	0.17 (.21)	0.17 (.21)
Incorrect		
Previous answer	0.13 (.19)	0.11 (.17)
Corrective feedback	0.43 (.42)	0.38 (.42)
Other	0.12 (.19)	0.19 (.26)

Table 4

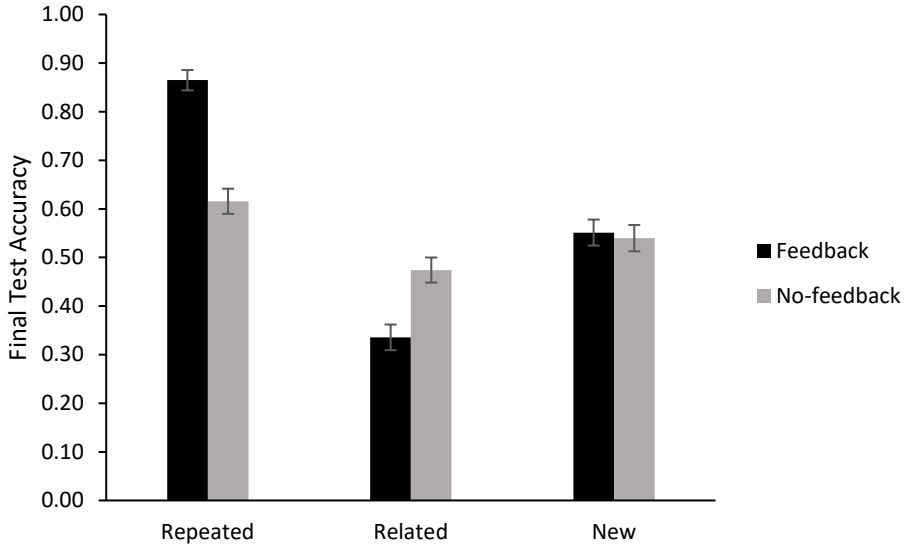
Mean (SD) Likelihood of Endorsing the Corrective Feedback on Test-2 Conditioned on Being Called "Old" or "New" in Experiment 2

Instruction type	Response	Related	New
Regular instruction	"Old"	0.77 (.30)	0.24 (.28)
	"New"	0.20 (.28)	0.14 (.14)
Opposition instruction	"Old"	0.69 (.30)	0.27 (.33)
	"New"	0.20 (.28)	0.20 (.18)

Paper 2 – Figures

Figure 1

Mean Accuracy for Each Question Type Broken Down by Feedback Type in Experiment 1



Note. Error bars indicate standard errors of the mean.

Figure 2

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

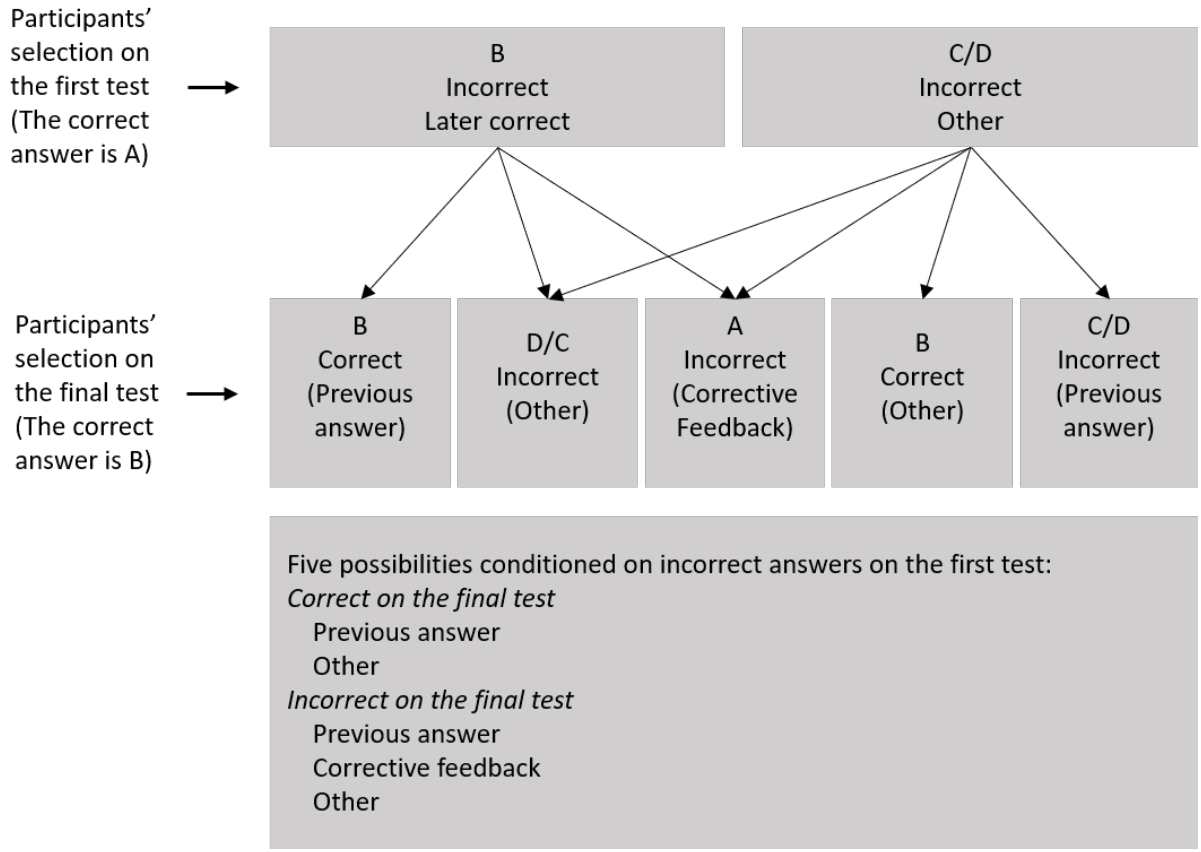
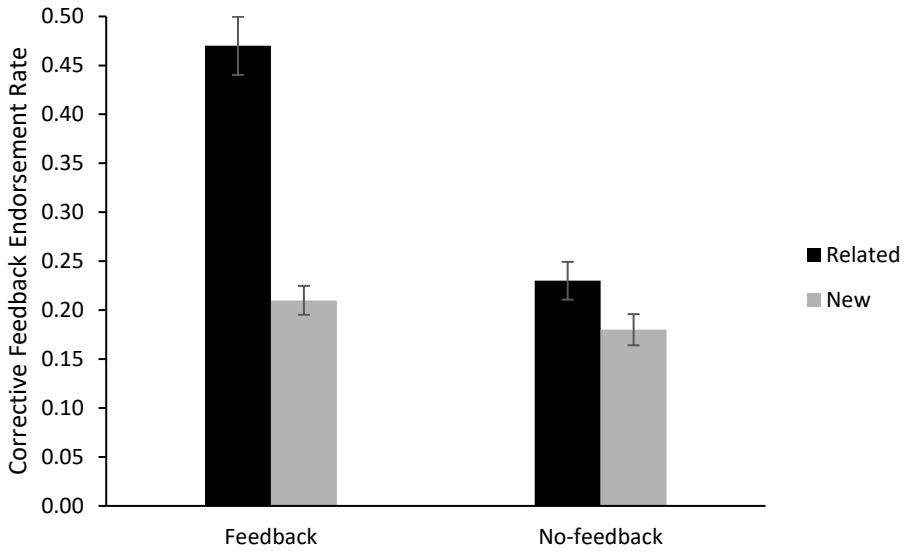


Figure 3

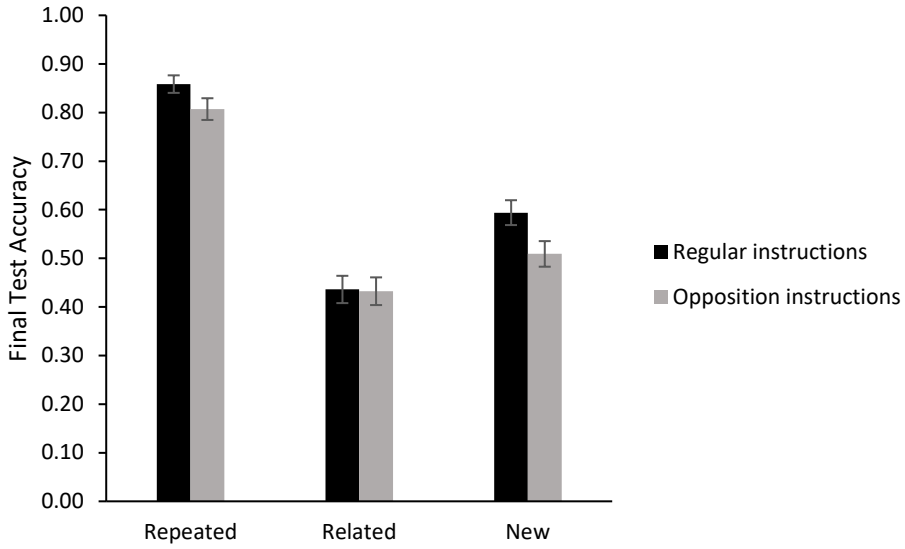
Mean Likelihood of Endorsing the Corrective Feedback for Each Question Type Broken Down by Feedback Group in Experiment 1



Note. Error bars indicate standard errors of the mean.

Figure 4

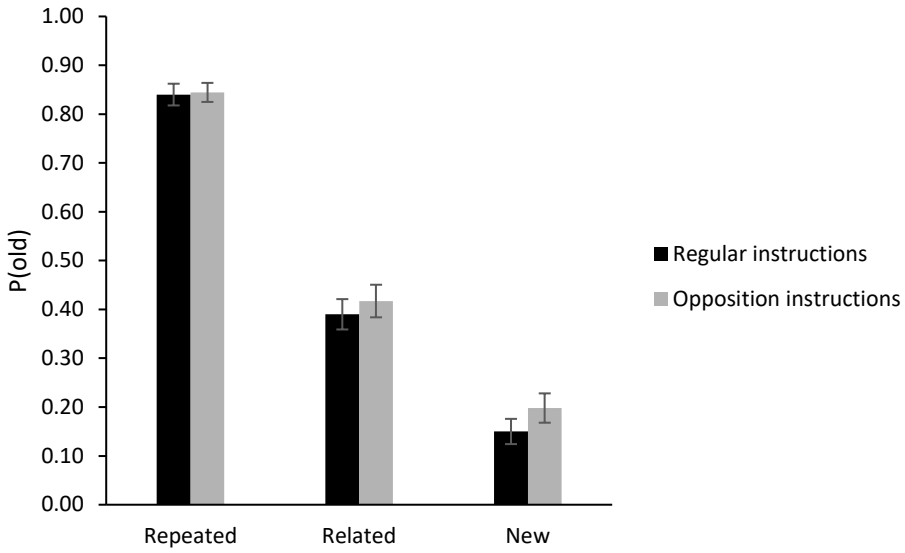
Mean Accuracy for Each Question Type Broken Down by Instructions Type in Experiment 2



Note. Error bars indicate standard errors of the mean.

Figure 5

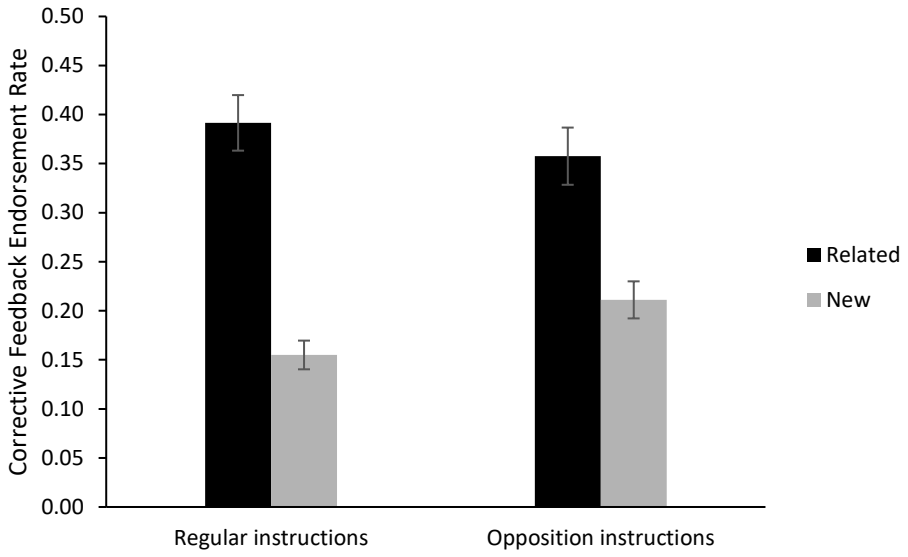
Mean Probability of Calling Each Question Type "Old" Broken Down by Instructions Type in Experiment 2



Note. Bars indicate standard errors of the mean.

Figure 6

Mean Likelihood of Endorsing the Corrective Feedback for Each Question Type Broken Down by Instruction Type in Experiment 2



Note. Bars indicate standard errors of the mean.

Paper 3 [Multiple-Choice Testing: Controlled and Automatic Influences of Retrieval Practice in an Educational Context]

Abstract

Previous studies have shown that taking an initial multiple-choice (MC) test produced both automatic and controlled influences on performance in a subsequent test. Whereas the controlled influence dominated performance when the final test was a cued recall (CR) test, the automatic influence dominated performance when the final test was a MC test. In this study, we examined the involvement of automatic and controlled processes in the performance with MC questions that are related to earlier practice questions, but which have different 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) and increasing the time lags between questions (Experiment 2) 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 CR final tests. 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 testing context, 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 Influences of Retrieval Practice in an Educational Context

Tests have been utilized as the main method to assess students' performance. 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 GRE[1] (Rauschert et al., 2019). This is due to the advantages of using this test format with a large number 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. A large literature has shown that taking an MC practice test can enhance later retention; this is known as the *testing effect*. For instance, McDermott et al. (2014) gave high school students two initial quizzes that contained MC and cued-recall (CR) questions. The first quiz was presented after the material was taught while the other one was given one day before the final test to act as a review quiz. The quizzes covered half of the to-be-learnt material while the other half was not tested. The untested material was used to generate the control questions on the final test. The final test included both MC and CR questions that did not necessarily match

[1] The SAT and GRE (Graduate Record Examination**)** are standardized tests that are used widely in the USA for university admission. The SAT is taken by high school students for undergraduate school admission, while the GRE is taken by undergraduate students for postgraduate school admission.

the format of the questions in the initial tests. The findings demonstrated an enhancement in students' performance on the tested items compared to the control items, regardless of the initial testing format. Interestingly, even when the test format did not match between the initial and final tests, participants still performed better on the tested items compared to the new items. This finding is consistent with other studies that have shown the usefulness of taking a practice MC test on later performance for students at different educational levels, including college students (e.g., McDaniel et al., 2012), middle school students (e.g., McDaniel et al., 2013), and elementary school students (e.g., 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 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 with questions 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 that were acquired from SAT subjects tests without feedback. After a short delay, participants took a final CR test that involved some tested questions as well as untested questions (control). 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 test. As a result of not receiving corrective feedback during the initial test, participants were more likely to produce one of the lures as an answer in the final CR test if the same lure was selected in the initial MC test. As in prior studies (e.g., Butler & Roediger, 2008), the researchers concluded that providing

corrective feedback for questions in the first test is an efficient way to reduce the negative testing effect and increase the positive testing effect.

Positive Effects of MC Testing on Related Items

The positive effect of MC testing extends beyond the tested items; it is found even with untested but related items. Related items, according to Little, Bjork, and colleagues (e.g., Little & Bjork, 2015) are a pair 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 query a similar topic (capital cities of Scandinavian countries). More critically, 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 final 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, 2016; Little et al., 2012; Little et al., 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 "*How many inches long is an average ferret tail? a. 7-10, b. 20,* ***c. 5***". After either a five minutes or a 48-hour 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, this enhancement was greater than that observed with a CR practice test, which showed no enhancement at all at either retention interval. 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 retrieval processes to reject incorrect alternatives, and the products of those retrieval processes can be used later to retrieve correct answers to related questions on the final CR test. This explanation is supported 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: Paper 1; 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, most of 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. In some cases, the option sets for

related questions were matched between the practice and final tests. However, in other experiments, only some of the options matched, but the correct answer on the first test was always an incorrect option on the final 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. 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. Also, participants still selected the corrective feedback for the second related question even when only one MC option (the corrective feedback) matched between the related items pairs. 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, 2016).

To further explore the contrasting final test format results with related items, Alamri and Higham (2022: Paper 1) 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 in 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 utilizing 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 increasing the retrieval depth 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 related items facilitation.

Theoretical Mechanisms

Alamri and Higham (2022: Paper 1) argued that their results were consistent with a dual-process model containing controlled and automatic influences. On the one hand, answering difficult MC practice questions with a test format that encourages lure processing elicits retrieval processes (i.e., CR tests) that can be used in a controlled manner to facilitate performance on later tests (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 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. The result is that participants falsely recognise 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 (2022: Paper 2) follow-up study. One potential criticism of the dual-process account is that the impairment on MC related items was not a result of automatic processes 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 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. However, they were also told that should be 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 to enhance accuracy would be undermined by these instructions. Conversely, if the influence was automatic, then the opposition instructions should make little difference.

In addition to the opposition instructions, Alamri and Higham (2022: Paper 2) 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

automatic influence account; related-item performance on the MC final test was impaired relative to new 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 recognised (i.e., falsely called “old” for the recognition question).

The automatic, false-recognition account was also supported by the results of another experiment from the same study in which the presence of feedback during the practice test was manipulated. Alamri and Higham (2022: Paper 2) 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 “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. Alamri and Higham (2022: Paper 1, 2022: Paper 2) noted that 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: Paper 1) 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[2]. 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: Paper 1) 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 testing, as impairment with MC 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 support Alamri and Higham's (2022: Paper 1, 2022: Paper 2) 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,

[2] Because items were counterbalanced between related and new items, this comparison was possible.

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: Paper 1) 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 three 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 and manipulating variables that we thought would moderate the relative influence of automatic and controlled processes. The tests in Experiments 2 and 3 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 hypothesised 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 if false recognition was the basis of the automatic influence, as Alamri and Higham (2022: Paper 2) contended, then the absence of repeated items might lessen the automatic influence and allow 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 tests might encourage participants to "look back" and search for repeated items, potentially increasing 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 false recognition on performance.

The aim of Experiments 2 and 3 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 than are online participants, thereby enhancing controlled influences. As in Experiment 1, we also manipulated a variable that we hypothesised 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 (2022: Paper 2) 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. 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.

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: Paper 1), 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 distracter 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: Paper 1) experiments, the lowest effect size for impaired performance with related MC items (i.e., related < new on final test) was Cohen's *f* = .24. Based on this result, we conducted 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-but-untested) manipulated within subjects. The main dependent variable was the participants' mean performance on the final MC 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. The three surveys rotated the questions through the repeated, related, and new conditions across participants to eliminate item effects. Twenty-two questions were presented on the initial test and 33 questions on the final test. The 33 final test questions consisted of 11 related (untested) questions, 11 repeated (tested) questions (for the repetition groups only; 11 filler items were used instead in the no-repetition groups), and 11 new (control) questions. Questions were presented in a random order for each participant, but the MC alternatives were always presented in the same order. Figure 1 summarises the design.

Materials and Procedure

Sixty-six questions of SAT subjects test were obtained from (CrackSAT, 2014) 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: Paper 1, 2022: Paper 2). 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 questions were prepared to replace the repeated questions in the no-repetition condition. This 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.

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: Paper 1, 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.

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 tests 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 presented in exactly the same format and covered similar topics as the questions on the initial test. 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: (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) and for the filler (i.e., distractor) questions for all four 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 scoring the initial and final tests in this experiment and the subsequent ones, 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. For the final-test answer type and the false endorsements of corrective feedback, we used (1, 0)

marking method. Specifically, (1) if an answer (e.g., corrective feedback) was ranked first, and (0) if that answer was ranked in the other three positions (2, 3, 4).

Results

Initial Test Performance

The mean rank of participants' correct answers on the first test in the two-tests/repetition, one-test/repetition, two-tests/no-repetition, and one-test/no-repetition groups were 1.91 (*SD* = .33), 1.99 (*SD* = .39), 1.92 (*SD* = .37), and 1.92 (*SD* = .34), 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 there was no evidence that the four groups were different.

Final Test Performance

Final-Test Accuracy. Although there were repeated questions and filler questions depending on group, we focus our analyses on the related and new questions. 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 (based on mean rank) as the dependent variable (see Figure 2). The main effect of test type was not significant, $F(1, 134) = 1.39$, $p = .24$, $\eta_p^2 = .01$, such that there was no significant difference in accuracy between the two-tests (*M* = 1.91, *SD* = .35) and one-test (*M* = 1.85, *SD* = .32) 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* = 1.95, *SD* = .37) compared to the repetition condition (*M* = 1.81, *SD* = .28). 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* = 1.97, *SD* = .38) compared to the related items (*M* = 1.79, *SD* = .44). 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 on the repetition condition performed better on the new items (*M* = 1.94, *SD* = .32) compared to the related items (*M* = 1.68, *SD* = .39), $t(68) = -4.94$, $p < .001$, $d = .74$. However, for the no-repetition condition, the difference between performance on the new items (*M* = 2.00, *SD* = .42), and the related items (*M* = 1.91, *SD* = .46) was not significant, $t(68) = -1.53$, $p = .13$, $d = .21$. No other interaction was significant, largest $F(1, 134) = 1.69$, $p = .19$, $\eta_p^2 = .01$ [3].

Final-Test Answer Types. Previous studies (e.g., Higham et al., 2016; Alamri & Higham, 2022: Paper 1, 2022: Paper 2) 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 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 answer on the initial test (correct/previous answer); (b) a correct answer that was neither the previous answer nor the corrective feedback on the first test (correct/other); (c) an incorrect answer that matched the previous answer on the initial test (incorrect/previous answer) (d) an incorrect answer that matched the corrective feedback that was provided in the first test (incorrect/corrective feedback); and (e) an incorrect answer that was neither the previous

[3] Note that there is no adjustment of alpha for the multiple follow-up t-tests in this analysis or the other analyses in this paper.

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 1. To analyse the data, we focused on responses that were incorrect on both tests (bottom panel of Table 1). 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 ($M = .72$, $SD = .31$) than the no-repetition condition ($M = .60$, $SD = .29$). 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 "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 "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 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. To examine the pattern of answering with corrective feedback in more detail, we conducted another analysis that examined the rates of corrective feedback endorsements regardless of whether the response was correct on the first test To explore the pattern in more detail, we conducted a second analysis that examined the rates of corrective feedback endorsements regardless of whether the response was correct on the first test To explore the pattern in more detail, we conducted a second analysis that examined the rates of corrective feedback endorsements regardless of whether the response was correct on the first test. To control for answer plausibility, we compared the 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 both groups were more likely to endorse the corrective feedback on related questions compared to new, but the difference was larger in the repetition group, $t(68) = 7.32$, $p < .001$, $d = 1.30$, (related: $M = .41$, $SD = .20$; new: $M = .21$, $SD = .12$) compared to the no-repetition group, $t(68) = 2.02$, $p = .046$, $d = .37$, (related: $M = .24$, $SD = .17$; new: $M = .19$, $SD = .10$). No other main effect or interaction was significant, largest $F(1, 134) = .35$, $p = .55$, $\eta_p^2 = .00$.

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. This result suggests that encouraging participants to "solve" the questions rather than "remember" previous answers (Jacoby, 1978) did not moderate the relative influences of automatic versus controlled processes.

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 condition, taking an initial MC test harmed participants' performance on the related items compared to the new items. For the no-repetition condition, 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 on 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 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: Paper 1; 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: Paper 1; Little et al., 2019), it may be a combination of the

explicit presentation of the corrective feedback amongst the MC options and the similarity of the MC questions driving the impairment to be persisted in a testing environment of this sort. Consequently, in Experiment 2, we investigated whether evidence of controlled influences on MC final-test performance might be found 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, as is true of most educational tests, no repeated items were included. 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 reasoned that presenting related items back-to-back would highlight the differences between the questions, invoke deeper processing, and allow controlled influences to override automatic influences. Because of the high-stakes educational context, we hypothesized that we might observe controlled influences with the separated related items as well, but not to the same degree.

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 all or most 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.

Design

The experiment had one independent variable, item 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 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 separated and 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. The three surveys rotated the questions through the three experimental conditions across students.

Figure 5 summarises the design. The first 11 questions of the test were the first questions of the 11 related-separated pairs 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. The first question in each pair counted toward initial test performance whereas the second question counted toward the final test. After that, the 11 control questions were presented. Finally, the second set of questions from the 11 related-separated pairs was presented. These questions counted as the final test for the related-

separated condition. Therefore, the final-test analysis included 11 questions of the back-to-back condition, 11 questions of the separated condition, and 11 new questions. 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 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 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 devices. 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

The mean rank of participants' correct answers on the initial test was 2.30 (*SD* = .27).

Final Test Performance

Final-Test Accuracy. A one-way ANOVA revealed a significant main effect of questions 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 back-to-back questions ($M = 2.42$, $SD = .39$), compared to the separated questions ($M = 2.30$, $SD = .40$), $t(163) = 2.97$, $p < .01$, $d = .30$, and new questions ($M = 2.18$, $SD = .41$), $t(163) = 6.00$, $p < .001$, $d = .61$. Also, participants had more accurate answers on the separated question 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 2). 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 items 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 that there was no significant difference between the probabilities of endorsing the "other" answer ($M = .17$, $SD = .20$) and previous answer ($M = .16$, $SD = .21$), $t(327) = -0.62$, $p = .53$, $d = .05$, 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 back-to-back condition, paired-sample *t*-tests showed that participants were more likely to select "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 "other" answer and the previous answer $t(163) = -0.88$, $p = .38$, $d = .10$. For the separated condition, participants selected "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 back-to-back ($M = .02$, $SD = .05$), $t(163) = 15.71$, $p < .001$, $d = 1.63$, and separated items ($M = .10$, $SD = .11$), $t(163) = -5.76$, $p < .001$, $d = .61$. Also, participants endorsed the corrective feedback on the separated items more than the 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 separated between questions. That is, participants performed better on both back-to-back and separated related 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 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 hypothesised, 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 presenting the related questions back-to-back, automatic influences (in the form of false recognition) would be kept to a minimum because the retention interval between the questions was negligible. 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 back-to-back condition. Noticing changes to related questions may have, in turn, improved students' 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 (2022: Paper 2) 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

As in Experiment 2, 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.

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. To rotate questions through the conditions across students, we created two surveys on Qualtrics. 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. The final test involved 44 questions consisting of 22 related questions (the second questions of the related pairs) and 22 new (control) questions. The initial test was presented in a fixed order while the

related and new questions in the final test were presented in a fixed random order. Also, the MC alternatives were always presented in the same order.

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. Some pairs were identical to those used in Experiment 2, whereas other pairs were completely new to accommodate new lecture material.

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 devices via a special 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 rank of participants' correct answers on the first test was 2.24 (*SD* = .27), whereas it was 2.32 (*SD* = .30) 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 (based on mean rank) as the dependent variable (see Figure 8). There main effect of feedback type was not significant, $F < 1$; such that there was no significant difference in accuracy between the feedback group (*M* = 2.12; *SD* = .29) and the no-feedback group (*M* = 2.15; *SD* = .31). 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* = 2.19, *SD* = .38) compared to the new items (*M* = 2.08, *SD* = .35). 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 3). We then analysed the three rates where the answers were incorrect on both the first and final tests. 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 "other" answer ($M = .21$, $SD = .17$) than corrective feedback ($M = .16$, $SD = .16$), $t(222) = -3.23$ $p < .001$, $d = .29$, or 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 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 9). 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 "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 profoundly deleterious effect on performance (e.g., Alamri & Higham, 2022: Paper 1). 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 also 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 order to gain the benefits of retrieval practice with related questions.

General Discussion

Over three 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: Paper 1; Little et al., 2012, 2019). In contrast, when the final test is MC, automatic influences have been observed (e.g., Alamri & Higham, 2022: Paper 1, 2022: Paper 2; 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., query the same topic and/or contains 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 learning rather than facilitate it. Second, from a theoretical perspective, the dual-process framework of retrieval practice that Alamri and Higham (2022: Paper 1) 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: Paper 1) dual-process model may not be a suitable framework for retrieval practice effects.

Overall, our results showed that controlled influences of retrieval practice do occur 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 performed 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 false recognition of related items was

the source of the automatic influence, just as it was in Alamri and Higham's (2022: Paper 1, 2022: Paper 2) earlier work. In other words, Experiment 1 replicated the finding that only automatic influences were observed when both tests are MC.

However, in Experiments 2 and 3, a very different pattern of results was observed. 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: Paper 1; Little & Bjork, 2015; Little et al., 2012, 2019; Sparck et al., 2016). The fundamental question, then, is what factor(s) caused the difference between Experiment 1 on the one hand, and Experiments 2 and 3 on the other?

In our view, the main difference was that the tests in Experiments 2 and 3 were administered in a formal educational setting whereas the tests in Experiment 1 were not, and this raised the stakes for achieving high marks. Critics may question this assessment because the tests in Experiments 2 and 3 were administered online, just as in Experiment 1, and they were formative, not 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 Experiment 2, 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 the students considered the tests in Experiments 2 and 3 high-stakes, which affected how participants processed the material on the test. For

example, questions may have been read more carefully compared 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 limited false recognition of the related questions 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 back-to-back questions in Experiment 2. These final-test questions were answered the best 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 automatic influences (false recognition) would have been negligible. But more critically, participants were in an ideal position to consciously identify discrepancies between the questions. Doing so 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. Consistent with this observation, the back-to-back methodology in Experiment 2 led participants to confidently reject the corrective feedback rather than simply fail to select it: the probability of ranking the corrective feedback in the last (fourth) position (i.e., the least favourite rank) was highest with the back-to-back related items (M = .41, SD = .21), intermediate with separated related items (M = .29, SD = .16), and lowest with new items (M = .17, SD = .10). These three means were all significantly different from each other: back-to-back > new, $t(163) = 12.52$ $p < .001$, $d = 1.41$; separated > new, $t(163) = 8.06$ $p < .001$, $d = .86$), and back-to-back > separated, $t(163) = 7.41$ $p < .001$, $d = .62$. 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 controlled influences were still observed. The suggestion from these data is that, as long as the context is right and examinees consider the test to be important, controlled influences are fairly robust, even if both tests are MC.

Overall, the current results are consistent with previous research showing 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: Paper 1). MC tests are a riskier option because if two questions are similar to each other but have different correct responses, participants may falsely recognise 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) or if corrective feedback is provided on the initial test (Alamri & Higham, 2022: Paper 2). On the other hand, if automatic influences are limited by, for example, presenting similar questions back-to-back (Experiment 2), or removing corrective feedback (Alamri & Higham, 2022: Paper 2), then these problems are less critical. However, the current experimental series suggests that the most important factor in determining whether controlled or automatic influences will dominate responding with

MC final tests is examinees' approach to the test. If they take it seriously and care about the results, then controlled influences tend to dominate even when seductive corrective feedback is provided on the first test (Experiment 3). These results both support Alamri and Higham's (2022: Paper 1) dual-process framework of retrieval practice effects.

Conclusions

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? (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 assigning probabilities to all the alternatives? (d) 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.

Paper 3 - Tables

Table 1

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 Repetition	One test Repetition	Two tests No-Repetition	One test 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 2

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 3

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)

Paper 3 - Figures

Figure 1

Schematic Illustrating the Design Used in Experiment 1

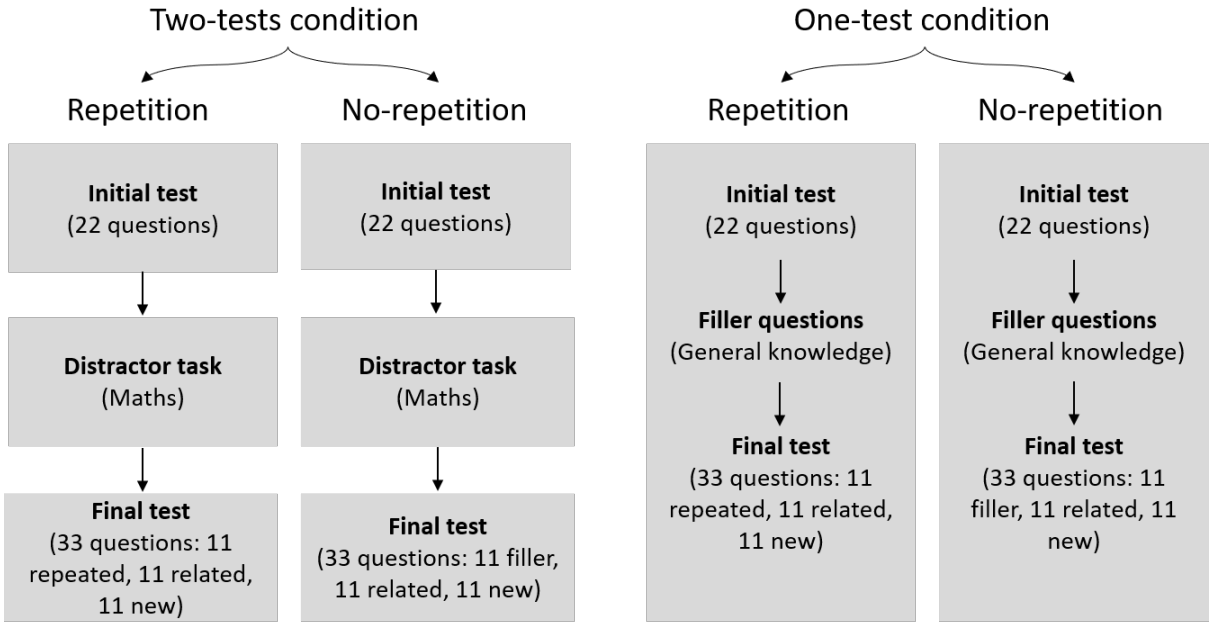
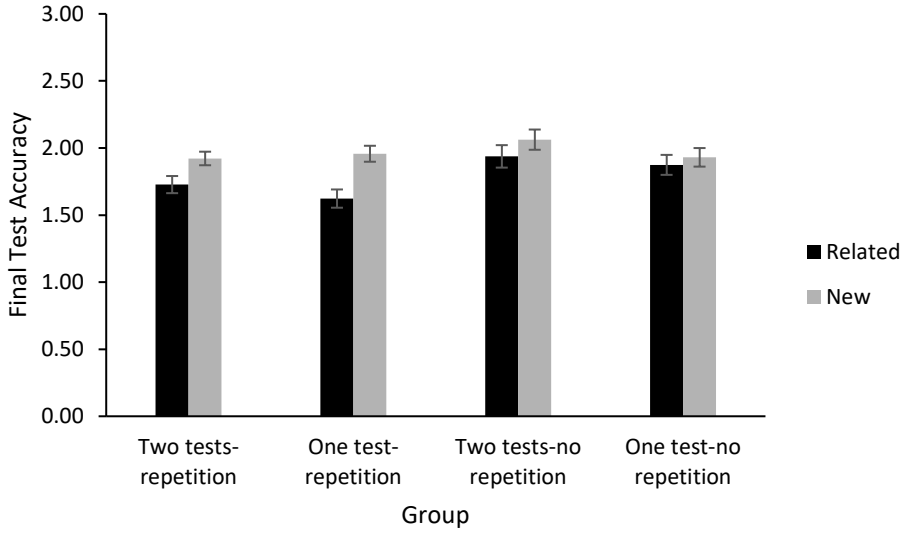


Figure 2

Mean (Rank) 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 rank 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

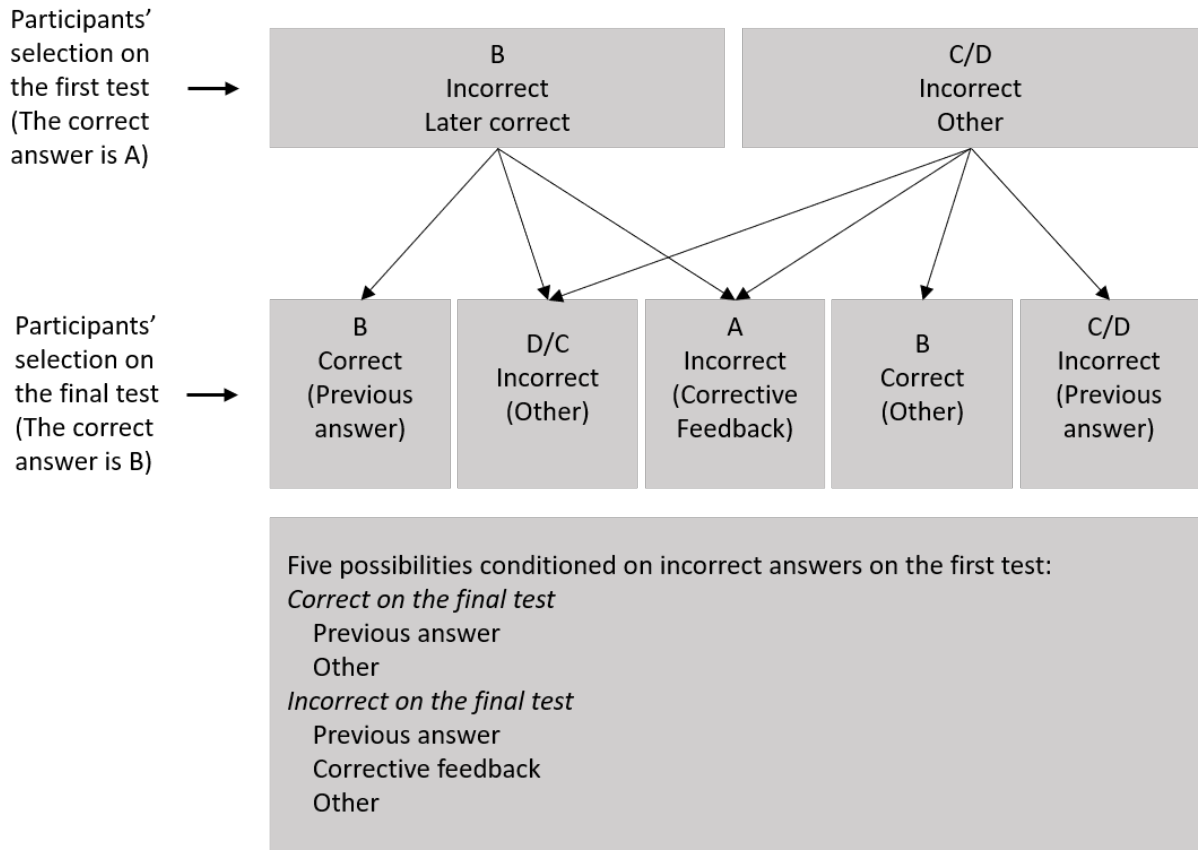
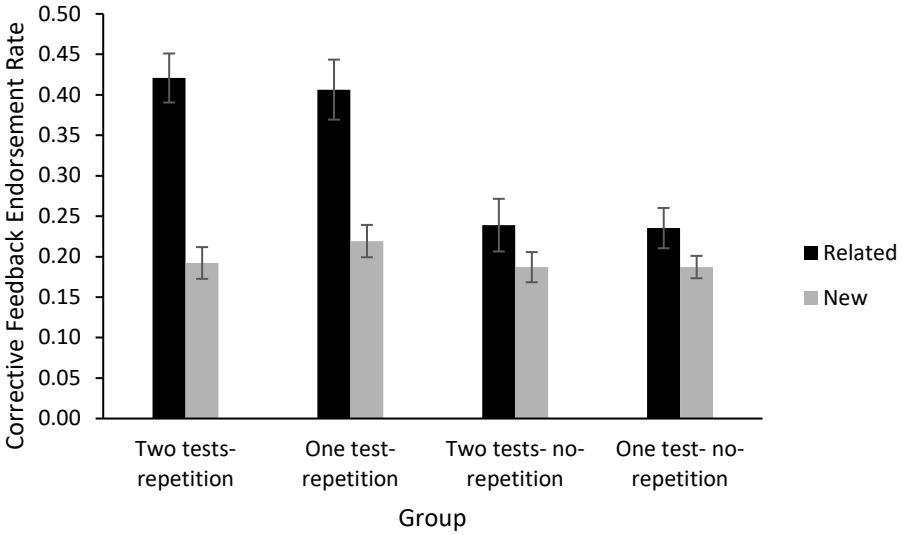


Figure 4

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

Schematic Illustrating the Design Used in Experiment 2

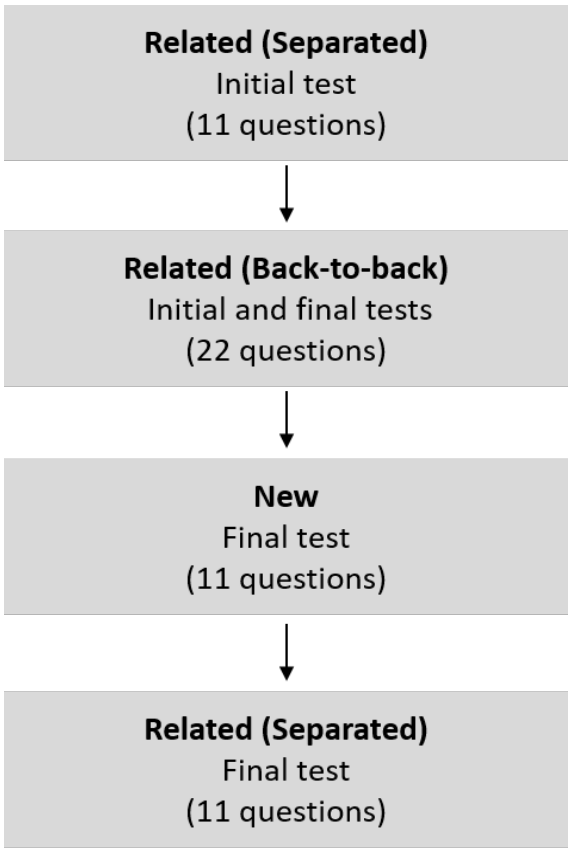
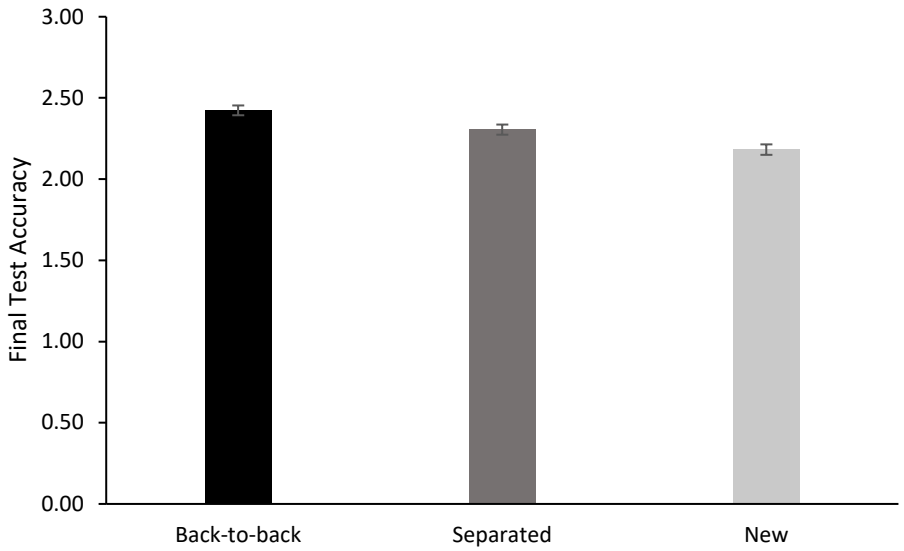


Figure 6

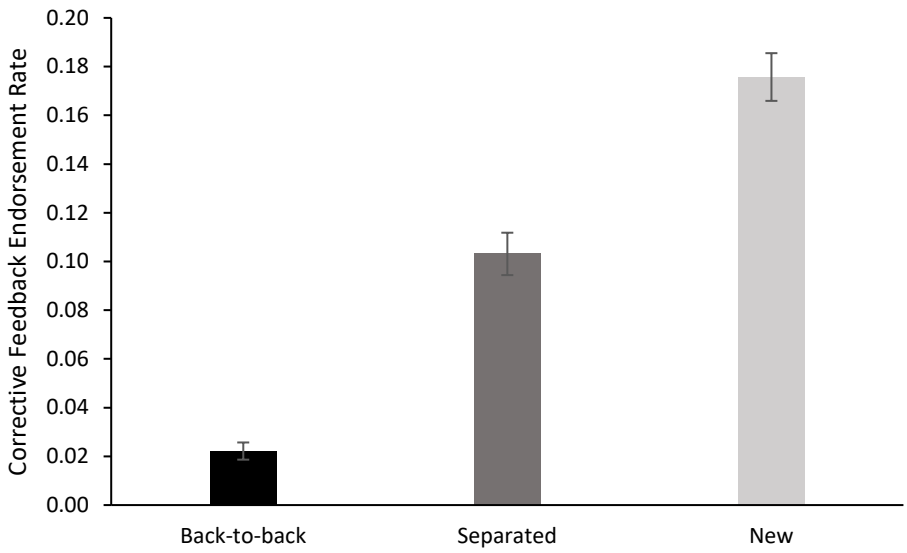
Mean (Rank) Final-Test Accuracy for Each Question Type in Experiment 2



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

Figure 7

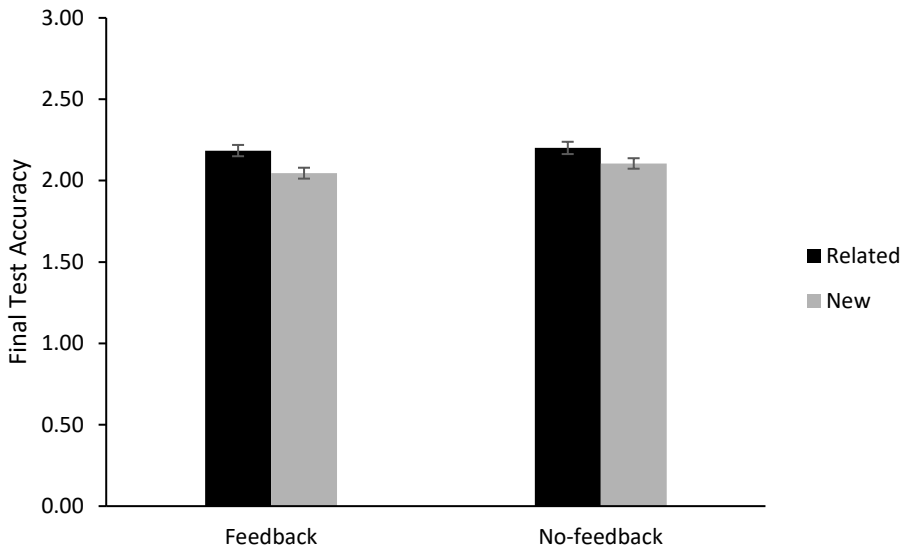
Mean Corrective Feedback Endorsement Rate for Each Question Type on the Final Test in Experiment 2



Note: Error bars indicate standard errors of the mean.

Figure 8

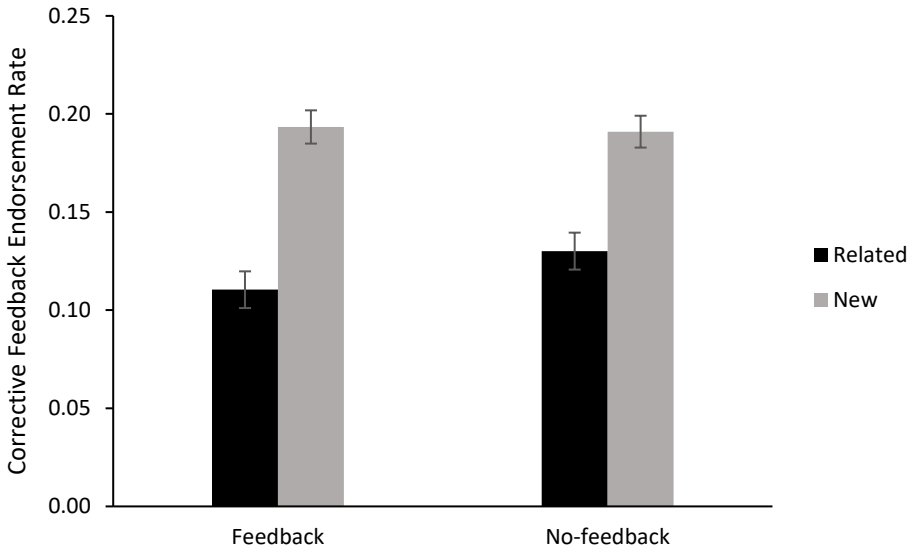
Mean (Rank) Final-Test Accuracy for Each Question Type Broken Down by Feedback Type in Experiment 3



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

Figure 9

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.

General Discussion

In this research, we investigated the consequences of retrieval practices with MC tests on later performance. Across eight experiments, we compared participants' performance on untested related items to their performance on new items to gain a better understanding of the consequences of taking an initial MC test on later retention. Prior research (e.g., Little & Bjork, 2015) showed that taking an initial MC test facilitated performance on related items when the final test used a CR format. On the other hand, a study by Higham et al. (2016) found that answering initial MC questions impaired participants' performance on related items in a final MC test. In this research, we focused mainly, but not exclusively, on examining the impact when the final test used an MC format. By manipulating different variables during the practice tests, conducting experiments in different contexts, and using different measures, we aimed to identify the potential applications of controlled and automatic processes in educational tests.

Paper 1. In three experiments, we compared the consequences of taking an initial MC test followed by a final MC test with those of taking an initial MC test followed by a final CR test. We successively increased the likelihood of retrieval that occurred during initial MC tests by varying the test format. In Experiment 1, we utilized a standard MC format (i.e., select a single option) in the initial test and allowed participants to take the test at their own pace. The results of their performance on both the MC and CR final tests demonstrated no enhancement or impairment compared to the new items. We attributed these findings to the initial test format, which elicited a low level of retrieval.

In Experiments 2 and 3, we increased the retrieval level in the initial test by employing an MC ranking format in Experiment 2 (i.e., participants were asked to rank order the options), and an elimination format in Experiment 3 (i.e., asking participants to provide their reasons for rejecting unchosen options; Little et al., 2019). Also, to prevent participants from finishing the practice test very quickly and to encourage deeper retrieval, we forced participants to wait for a minimum of 15-seconds before they could proceed to the next question. Consequently, we observed impairment in participants' performance on the MC related items on the final test as a result of automatic influences. We also found facilitation in the CR related items, indicating controlled processing. However, analysing the endorsement rate of corrective feedback between related items and new items across the three experiments demonstrated that automatic influences existed in both test formats, taking the form of false recognition in the MC test and false recollection in the CR test.

Paper 2. Based on the findings of Paper 1 and given the large number of studies that have investigated performance on related questions in final CR tests, in Paper 2, we focused on examining the mechanisms underlying impairment on MC final tests. In Experiment 1, we compared two groups of participants who took an initial MC test with or without corrective feedback to see how this would impact their performance on related items in the final MC test. A comparison of participants' performance on related versus new items showed that the findings replicated those reported in Paper 1 (i.e., Experiments 2 and 3), regardless of whether feedback was provided during the initial practice test. That is, participants in both groups (i.e., the feedback group and the no-feedback group) performed worse on the related items, and endorsed feedback at a high

rate on related items compared to the new items, although the difference was larger in the feedback group versus the no-feedback group.

Although the findings of Experiment 1 demonstrated reduced impairment in the no-feedback group, the overall results still showed that automatic processes dominated the performance of both groups (i.e., feedback and no feedback) to a different degree. In Experiment 2, we aimed to provide evidence of the false recognition that we predicted would drive the impairment on MC related items on the final test. Thus, we asked participants to identify the questions in the final MC tests as “old” or “new”. Also, we gave one group opposition instructions and the other group regular instructions and compared the performance of the two groups on the final MC test. The group that received opposition instructions was warned prior to the final test not to repeat the same answer when they encountered similar but not identical questions on the final test. By setting controlled and automatic influences in opposition we aimed to gain a better understanding of the nature of the impairment on related items.

Regardless of the type of instruction they received, both groups demonstrated impairment on related items versus new items and a strong tendency to endorse corrective feedback on the related items compared to the new ones. In terms of recognition, the findings showed that the impairment on related items was due to participants falsely recognising related questions as “old” (i.e., repeated from the initial test). Specifically, the answers selected by participants in both groups were more accurate for the new items than for the related items when they identified them as “old”. However, when participants identified items as “new”, there was no difference in accuracy between related and new items. Thus, although the automatic influence was minimised with items correctly recognised, no facilitation was observed.

Paper 3. Our previous experiments provided evidence of the controlled and automatic processes in the CR final test. For the MC final test, however, there was no evidence of controlled influences as automatic influences dominated MC performance. Thus, the main aim of Paper 3 was to examine whether there are any conditions under which evidence of controlled processes might be found in participants' performance on MC final tests. In Experiment 1, we conducted the test online with MTurk as we did with the previous experiments in Papers 1 and 2. However, instead of the trivia questions used previously, we used questions that were acquired from SAT subject tests. We hypothesised that using questions taken from a high-stakes standardised exam might encourage deep processing and thereby reveal the controlled influence. Also, in an attempt to increase controlled processes, we manipulated the presence/absence of repeated items in the MC final test as well as whether the questions appeared to participants in a single test or two tests.

The participants' accuracy on the final test showed no difference between the one-test and two-test conditions, as their performance on the MC related items remained roughly the same. However, eliminating repeated items from the final test was found to be an important determinant of participants' performance on related items. Specifically, the findings demonstrated impairment on related items when repeated items were included in the final test, while eliminating repeated items resulted in neither impairment nor facilitation on the related items. However, analysing the endorsement rate of corrective feedback indicated that both groups endorsed corrective feedback on related items more than on new items, although the difference was greater in the repetition group. Although these results showed that eliminating repeated items from the final test

helped to temper automatic processes, they still dominated participants' performance on the final MC tests, as evidenced by the results of the false endorsement rate.

Consequently, in Experiments 2 and 3, we shifted our investigation to the educational context to see whether controlled influences could be found in MC testing when the stakes were high. In Experiment 2, we varied the time lag between the related questions; that is, in one condition, the related questions were separated by several other questions; in the other condition, the related questions were presented sequentially (i.e., back-to-back). In Experiment 3, we investigated how providing or withholding corrective feedback during the initial test would affect later MC performance in the educational context. The results of both experiments showed enhanced performance on related items versus new items regardless of the different lags separated between questions as well as the feedback. Also, higher endorsement rate of corrective feedback on new items than on related items.

MC Testing with Repeated Questions

Our overall findings indicate that taking an initial MC test can be beneficial in many scenarios and with different test formats. Specifically, on the final test, participants' performance was better on questions that were repeated than on untested questions, regardless of whether the untested questions were related to questions that were previously answered on the initial MC test or were new questions. This benefit for repeated items was observed with different final test formats (i.e., MC and CR; Paper 1) and regardless of whether corrective feedback was provided in the initial practice test (Paper 2; Experiment 1). These findings are consistent with many studies that have illustrated the positive effect of taking MC tests as a learning tool (e.g., Marsh et al., 2012;

McDaniel et al., 2013; McDermott et al., 2014; Nungester & Duchastel, 1982; Rowland, 2014; Yang et al., 2021).

MC Testing with Related Questions

The main aim of the current research was to examine participants' performance on related items in the final test and compare it with their performance on new items. The results demonstrated both positive and negative effects of taking an initial MC test on performance on related items. When the final test is CR and participants answered difficult MC practice questions on the initial test, the result was enhanced performance. This is due to participants retrieving information about the question and lures when answering the initial MC test (i.e., through ranking or elimination formats), which could be used later to facilitate performance with related questions on the final CR test. These findings replicated those of many previous studies that showed better performance on CR related items (vs. new items) when the level of lure elaboration is boosted during practice testing (Little & Bjork, 2015; Little et al., 2012; Sparck et al., 2016).

On the other hand, when the final test was in MC format, our results demonstrated contrasting performance depending on the contexts in which the test was held. Experiments that were conducted via MTurk replicated the findings of Higham et al.'s (2016) study; that is, participants' performance was impaired on related items compared to their performance on new items, even though Higham et al.'s study was lab-based and recruited undergraduate participants at the University of Southampton. This impairment was due to participants falsely recognising related items as repeated, thereby choosing the corrective feedback at a high rate on related questions (which was no longer correct). Conversely, experiments conducted in a genuine educational context showed that participants performed better on related items than on new items. Seemingly, students in

the educational context retrieved deeply during practice and encoded the details of the question, which aided them in noticing that the related question on the final test was not repeated. The findings of the experiment conducted in the educational context are in line with those of research investigating performance on related items in CR final tests (Little & Bjork, 2015, 2016).

Overall, the results on related questions showed opposing performance, which in our view can be attributed to the stakes that each context may involve. The tests conducted in a formal educational context involved higher stakes (vs. the low stakes of MTurk), which motivated the students to score well. The students in the educational context might have read the questions carefully and invested more in retrieving information related to the target answer, which resulted in limiting false recognition of the related items, allowing them to benefit from the similarity between the questions.

Theoretical Mechanisms

Our findings suggest that taking an initial MC test can facilitate both automatic and controlled memory influences, and which one dominates depends on the testing context, participants' investment in scoring well, and the format of the final test (i.e., CR or MC). That said, our results do not support a process purity assumption (e.g., Jacoby, 1991), that is, the assumption that automatic processes always dominate performance on MC tests, while controlled processes always dominate performance on CR tests. Although some researchers (e.g., Ozuru et al., 2013) differentiated between recollection and recognition tests based on the processes that each type invokes, our results suggest that both controlled and automatic processes take place under certain circumstances in both test formats. However, using CR tests with related items might be better overall at prompting controlled memory processes than MC tests. The absence of the matching options on the

CR test seemed to aid participants to notice that the questions are different, and that a new answer was needed. While, the strong match between the MC related pairs (i.e., query a similar topic and use the same set of alternatives) led participants to falsely recognise the related questions as repeated and select corrective feedback.

Moreover, the results of the current study indicated that some determinants increased the automatic process, resulting in more impairment on MC related questions. These determinants included providing corrective feedback during practice tests (Paper 2, Experiment 1) and including repeated questions in final MC tests (Paper 3, Experiment 1). Both factors seemed to increase the false belief that the related questions were repeated, which prompted participants to overwhelmingly select the corrective feedback as the correct answer in the final MC test. On the other hand, our results showed that presenting related questions back-to-back (Paper 3, Experiment 2) and eliminating corrective feedback from the practice test (Paper 2, Experiment 1) could limit the false recognition of related items and thereby enhance participants' performance. However, our overall findings suggest that participants' attitudes toward the test was the most important factor in determining which of the two types of processing overrode the other. When participants were keen on performing well on the test, controlled processes dominated their performance regardless of whether they received feedback during the initial test (Paper 3, Experiment 3). In contrast, when participants were less motivated to do well on the test, automatic processes dominated their performance.

Another result with a theoretical application is that while increasing the level of retrieval that occurs during an initial MC test (i.e., through lures elaboration) is vital to facilitate controlled processes and enhance later CR performance, it had almost no effect on automatic processes, as evidenced by the false recognition in MC tests and false

recollection in CR tests (Paper 1; e.g., Jacoby et al., 1993). Consistently, our results demonstrated that setting the automatic and controlled processes in opposition (Paper 2, Experiment 2) had little effect on automatic influences, which provides evidence of the invariance of this process. These findings are in line with previous research on the process dissociation paradigm, which showed some factors that affect controlled influences while leaving automatic influences unchanged (e.g., Jacoby, 1991, 1996, 1999).

Practical Applications

Some educators may deliberately construct MC questions that are quite similar to earlier practice questions (between or within tests) to encourage students to think deeply about each question. Although some of our findings might discourage educators from using MC tests, especially with related items, the findings from the experiments conducted in an educational context (Paper 3, Experiments 2 and 3) suggest otherwise. That is, using an initial MC test enhanced students' performance on related items, thereby providing evidence of the usefulness of practice MC tests. Indeed, in a recent meta-analysis of studies conducted in real educational contexts, Yang et al. (2021) concluded that using MC practice tests produces larger testing effects than using practice tests involving recall (e.g., CR tests). This is supported by many studies that have demonstrated the benefits of taking an initial MC test on both tested items (e.g., Marsh et al., 2012; McDaniel et al., 2013) and related but untested items (Little & Bjork, 2012, 2015, 2016; Sparck et al., 2016). However, until further studies are conducted within the educational context, educators might consider taking more precautions to maximize the advantages of MC testing, in particular by not repeating options on the final test that were provided on the initial test. Importantly, our educational-context experiments did not examine the involvement of repeated items within the test; if the MC final test

includes repeated items along with the related items, automatic processes might dominate, potentially undermining learning as a result (Paper 3, Experiment 1).

In terms of corrective feedback, many studies have shown that providing corrective feedback was vital to strong test performance (e.g., Finn & Metcalfe, 2010; Mullet et al., 2014). When the questions are repeated, Butler and Roediger (2008) suggested that providing corrective feedback decreased the negative testing effect and increased the positive testing effect. For the related items, however, some of the findings reported in the current research showed a dark feedback effect on related items. Specifically, providing corrective feedback during practice tests worsened performance on MC related questions compared to when no feedback was provided (Paper 2, Experiment 1). However, considering the findings from the experiment conducted in an educational context (Paper 3, Experiment 3), educators may be confident that even if feedback does not enhance performance, at least it does not harm learning. That is, students in the educational context benefited from answering related questions to the same extent, regardless of whether they received feedback on the initial test. These results are supported by Little et al.'s (2012) findings showing similar facilitation in CR related items, regardless of whether corrective feedback was provided in the practice test.

On the other hand, educators who want to enhance their students' knowledge, especially of related information, might consider using the back-to-back methodology. Our findings showed that presenting related items back-to-back resulted in best performance (vs new and separated conditions), which might be attributed from a different theoretical perspective to participants detecting changes between related questions (Paper 3, Experiment 2). Indeed, studies that have investigated change detection (i.e., noticing changes between original and later events) showed that when

participants detected changes, retention was enhanced (e.g., Garlitch & Wahlheim, 2020; Wahlheim & Jacoby, 2013). Specifically, when two events are presented in a sequence, the later event can act as a reminder of the previous one if the changes are detected and recalled, which results in improving the memory of the earlier event. Accordingly, sequentially presenting related questions can help students identify discrepancies between the questions and potentially promote deeper processing and a greater understanding of the related information. Both educators and researchers might use and investigate the back-to-back method as a potential way to enhance the benefits of retrieval practice.

Limitations and Future Research

Although the overall findings of our experiments are consistent with our exceptions of the contribution of controlled and automatic processes in MC testing, there are other factors that are worth investigating. For example, as the main focus of our study was on the performance of related items which are similar on several dimensions, it is not clear to us how much similarity is needed before the similarity becomes an essential factor that can affect the performance on related items. Higham at al. (2016) investigated the role of the options on the performance of related items; the researchers had only one option (i.e., the corrective feedback option) that was repeated from the initial test on the final test, and impaired performance was still observed on related items compared to new items. However, to our knowledge, no previous studies have investigated similarities in question stems. Specifically, does a related pair in which the questions are worded differently produce less automatic processing than a related pair in which the questions are worded very similarly? Also, to what extent do questions need to be related conceptually (i.e., query similar topics) to produce automatic influences? Our results

indicated that false recognition drove the poor performance on related questions, and similarity between questions might play a major role in increasing/decreasing this false recognition. Thus, a potential avenue for future research would be to examine the similarity dimensions that seduce learners into identifying related questions as repeated questions.

Another potential direction for future research is to conduct studies in an educational context, as our results showed that increasing the time lag between the initial and final tests resulted in more corrective feedback intrusion. However, we do not know how longer intervals between the practice test and the final test, which is typical in the educational context (e.g., an interval of weeks or months), would moderate the controlled and automatic processes. Higham et al.'s (2016) findings showed that regardless of whether the final test was taken immediately after the initial test or after a seven-day interval, the impairment on MC related items in the final test remained roughly the same. However, Little and Bjork's (2012) study with final CR tests showed that facilitation persisted on related items despite a long interval between the two tests (i.e., 48 hours). Since the findings of the current study demonstrated that controlled influences dominated participants' performance on final CR tests and MC tests in the educational context, it is possible to observe similar results as those of Little and Bjork (2012), regardless of the retention interval.

An important factor in the current study was the inclusion of repeated items in the final test, which increased participants' false belief that the related questions were repeated from the first test. However, the findings of our experiments conducted in an educational context were based on final tests in which there were no repeated items. Hence, future research might examine performance on related items when some

repeated questions are also included in the final MC test. Usually, tests in an educational context do not involve repeated questions within the same test. Thus, it is possible that some of our participants were motivated to exclude the possibility that any related questions were repeated (i.e., false recognition), and as a result, looked at each question as new. Based on our results (Paper 3, Experiment 1) we suspected that including repeated questions in the final test might reveal more automatic influence and show another side of the interplay between automatic and controlled influences within the educational context.

In the current study, we did not investigate participants' level of confidence when completing the final test. Our results showed that participants falsely recognised related items as repeated, and therefore they were prompted to endorse corrective feedback as the correct answer on the final test. However, in cases where there were no repeated items on the final test, participants correctly recognised the related questions as new, and no corrective feedback was received in the initial test, the findings suggest that some participants no longer believed that the corrective feedback was the correct answer and thus considered other options. Regardless, no facilitation was observed, which might suggest that participants' selection of one of the other options was a quick guess that was not supported by any deep retrieval indicating controlled processes. Examining participants' confidence in their selections could help us to understand why we did not observe facilitation when the automatic process was minimised. In addition to confidence rate, future studies might examine latencies – especially in cases where automatic influence is minimal, which could provide evidence of whether a quick guess drives participants' performance in cases resulting in no facilitation.

Along the same lines, future research might focus on exploring the controlled influence when using an MC format in the final test that is more likely to evoke intensive retrieval (e.g., an elimination format, as in Little et al., 2019). If participants were forced by the format of the final test to look closely at each option, controlled processes might be revealed. It is possible that the test format we used on the final test encouraged participants to not carefully look into each option. Consequently, presenting the final test in an elimination format might result in more controlled processes, especially in cases where the automatic influence is minimal (e.g., no feedback provided during practice tests).

To provide further support for the automatic influence argument with regard to MC-related items, future research might consider using a deadline methodology. Previous studies (e.g., Higham et al., 2000; Jacoby, 1999) showed that employing a response deadline in a final performance resulted in reduced controlled influence, while it had no effect on automatic influence. Therefore, if the impaired performance on MC-related items was driven by automatic processes, as our results suggest, then future studies should observe results similar to those reported here, regardless of the response deadline.

Conclusions

Through three papers, our practical aim was to contribute to the growing body of research investigating the pros and cons of utilizing MC testing, especially with related questions. Another primary contribution of our research was theoretical, as we aimed to gain a better understanding of the controlled and automatic influences in retrieval practices, a topic that has received little attention in the literature. Our overall findings suggest that both controlled and automatic processes occurred with MC testing, and

which one dominated depends on certain circumstances. With use caution, we recommend that educators continue to test their students with MC questions to enhance retention. In addition, educators might consider using the back-to-back methodology to enhance students' knowledge, especially of related information. We hope that this research helps to shed some light on the nature of the memory processes underpinning MC testing and encourages broader research to explore the involvement of automatic and controlled processes in retrieval practices.

List of References

Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the use of tests: A meta-analysis of practice testing. *Review of Educational Research, 87*(3), 659–701. https://doi.org/10.3102/0034654316689306

Alamri, A., & Higham, P. A. (2022). *Automatic influences of retrieval practice: the role of feedback, false recognition, and opposition instructions*. Manuscript in preparation.

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

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, 48*(5), 752–768. https://doi.org/10.1037/xlm0001138

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

Butler, A. C., & Roediger, H. L. (2007). Testing improves long-term retention in a simulated classroom setting. *European Journal of Cognitive Psychology, 19*(4–5), 514–527. https://doi.org/10.1080/09541440701326097

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

Butler, A. C., Karpicke, J. D., & Roediger, H. L. (2008). Correcting a metacognitive error: Feedback increases retention of low-confidence correct responses. *Journal of Experimental Psychology: Learning Memory and Cognition, 34*(4), 918–928. https://doi.org/10.1037/0278-7393.34.4.918

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

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

Chan, J. C. K., McDermott, K. B., & Roediger, H. L. (2006). Retrieval-induced facilitation: Initially nontested material can benefit from prior testing of related material. *Journal of Experimental Psychology: General, 135*(4), 553–571. https://doi.org/10.1037/0096-3445.135.4.553

CrackSAT. (2014). *SAT Subject Practice Tests*. http://www.cracksat.net

Dubins, D. N., Poon, G. M. K., & Raman-Wilms, L. (2016). When passing fails: Designing multiple choice assessments to control for false positives. *Currents in Pharmacy Teaching and Learning, 8*(5), 598–608. https://doi.org/10.1016/j.cptl.2016.05.005

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest, Supplement, 14*(1), 4–58. https://doi.org/10.1177/1529100612453266

Fazio, L. K., Agarwal, P. K., Marsh, E. J., & Roediger, H. L. (2010). Memorial consequences of multiple-choice testing on immediate and delayed tests. *Memory and Cognition, 38*(4), 407–418. https://doi.org/10.3758/MC.38.4.407

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

Finn, B., & Metcalfe, J. (2010). Scaffolding feedback to maximize long term error correction. *Memory and Cognition, 38*(7), 951-961. https://doi:10.3758/MC.38.7.951

Garlitch, S. M., & Wahlheim, C. N. (2020). The role of attentional fluctuation during study in recollecting episodic changes at test. *Memory & Cognition, 48*(5), 800–814. https://doi.org/10.3758/s13421-020-01018-4

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

Glass, A. L., & Sinha, N. (2013). Multiple-choice questioning is an efficient instructional methodology that may be widely implemented in academic courses to improve exam performance. *Current Directions in Psychological Science, 22*(6), 471–477. https://doi.org/10.1177/0963721413495870

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items*. Lawrence Erlbaum.

Heist, B. S., Gonzalo, J. D., Durning, S., Torre, D., & Elnicki, D. M. (2014). Exploring clinical reasoning strategies and test-taking behaviors during clinical vignette style multiple-choice examinations: A mixed methods study. *Journal of Graduate Medical Education, 6*(4), 709–714. https://doi.org/10.4300/JGME-D-14-00176.1

Hinze, S. R., & Wiley, J. (2011). Testing the limits of testing effects using completion tests. *Memory (Hove, England), 19*(3), 290–304. https://doi.org/10.1080/09658211.2011.560121

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

Higham, P. A., 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

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

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

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

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

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

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

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

Kang, S. H. K., McDermott, K. B., & Roediger, H. L. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of*

Cognitive Psychology, 19(4–5), 528–558. https://doi.org/10.1080/09541440601056620

Little, J. L. (2018). The role of multiple-choice tests in increasing access to difficult-to-retrieve information. *Journal of Cognitive Psychology, 30*(5–6), 520–531. https://doi.org/10.1080/20445911.2018.1492581

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 & Cognition, 43*(1), 14–26. https://doi.org/10.3758/s13421-014-0452-8

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

Little, J. L., Bjork, E. L., Bjork, R. A., & Angello, G. (2012). Multiple-choice tests exonerated, at least of some charges: 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. http://dx.doi.org/10.1037/xlm0000638

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

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

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

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

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

McDermott, K. B., Agarwal, P. K., D'Antonio, L., Roediger, H. L. 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

Meyer, A. N. D. (2011). *The positive and negative effects of testing in lifelong learning*. (Publication No. 3464226). [Doctoral dissertation, Rice University]. ProQuest Dissertations and Theses Global.

Mullet, H. G., Butler, A. C., Verdin, B., von Borries, R., & Marsh, E. J. (2014). Delaying feedback promotes transfer of knowledge despite student preferences to receive

feedback immediately. *Journal of Applied Research in Memory and Cognition, 3*(3), 222-229. https://doi:10.1016/j.jarmac.2014.05.001

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

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

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

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

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

Polat, M. (2020). Analysis of multiple-choice versus open-ended questions in language tests according to different cognitive domain levels. *Novitas-ROYAL (Research on Youth and Language), 14*(2), 76-96.

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

Pu, X., & Tse, C. S. (2014). The influence of intentional versus incidental retrieval practices on the role of recollection in test-enhanced learning. *Cognitive Processing, 15*(1), 55–64. https://doi.org/10.1007/s10339-013-0580-2

Rauschert, E. S. J., Yang, S., & Pigg, R. M. (2019). Which of the Following Is True: We Can Write Better Multiple Choice Questions. *The Bulletin of the Ecological Society of America 100*(1), e01468. https://doi.org/10.1002/bes2.1468

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17*(3), 249–255. https://doi.org/10.1111/j.1467-9280.2006.01693.x

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

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

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin, 140*(6), 1432–1463. https://doi.org/10.1037/a003755

Smith, M. A., & Karpicke, J. D. (2014). Retrieval practice with short-answer, multiple-choice, and hybrid tests. *Memory, 22*(7), 784–802. https://doi.org/10.1080/09658211.2013.831454

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

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

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

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

Verkoeijen, P. P. J. L., Tabbers, H. K., & Verhage, M. L. (2011). Comparing the effects of testing and restudying on recollection in recognition memory. *Experimental Psychology, 58*(6), 490–498. https://doi.org/10.1027/1618-3169/a000117

Wahlheim, C. N., & Jacoby, L. L. (2013). Remembering change: The critical role of recursive remindings in proactive effects of memory. *Memory & Cognition, 41*(1), 1–15. https://doi.org/10.3758/s13421-012-0246-9

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

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language, 46*(3), 441–517. https://doi.org/10.1006/jmla.2002.2864

Accompanying Materials

Paper 1 - Experiment 1

General Knowledge MC Questions

1. diffraction; diffusion; distillation; entropy

1.1 In chemistry, the separating of the constituents of a liquid by boiling it and then condensing the vapor that results is called_____.

distillation

1.2 The spreading of atoms or molecules of one substance through those of another, especially into liquids and gases is known as _____.

diffusion

2. centaur; chimera; minotaur; satyr

2.1 In classical mythology, a creature that is half human and half horse is called a _____

centaur

2.2 In classical mythology, a creature that is half man and half bull is called a _____.

minotaur

3. Aphrodite; Athena; Isis; Venus

3.1 The Greek goddess of love is _____.

Aphrodite

3.2 The Roman Goddess of Love is _____.

Venus.

4. Jupiter; Neptune; Saturn; Uranus

4.1 What is the fifth planet from the Sun?

Jupiter

4.2 What is the sixth planet from the Sun?

Saturn

5. Hamlet; Julius Caesar; Merchant of Venice; Romeo and Juliet

5.1 From what Shakespearian play comes the line, 'This above all: to thine own self be true'?

Hamlet

5.2 From what Shakespearian play comes the phase 'pound of flesh'?

Merchant of Venice

6. All about Eve; Casablanca; Gone with the Wind; Sunset Blvd

6.1 From what classic movie comes the line, 'Frankly my dear, I don't give a damn'?

Gone with the Wind

6.2 From what classic movie comes the line, 'Here's looking at you, kid'?

Casablanca

7. ad hoc; ad nauseam; ergo; et cetera

7.1 The Latin translation of 'therefore' is _____.

ergo

7.2 The Latin translation of 'and so forth' is _____.

et cetera

8. Cirrus; Cumulus; Nimbus; Stratus

8.1 What is the term for lacy or wispy clouds that form at high altitudes, often before a change in the weather?

Cirrus

8.2 What is the term for large, white, puffy clouds that generally appear in fair weather, but that can also form thunderheads on hot days?

Cumulus

9. Anna Karenina; Brothers Karamazov; Crime and Punishment; War and Peace

9.1 _____ is a novel by Feodor Dostoevsky in which a man kills two old women because he believes that he is beyond the bounds of good and evil.

Crime and Punishment.

9.2 _____ is a novel by Leo Tolstoy in which a woman enters a tragic adulterous affair and commits suicide by throwing herself under a train.

Anna Karenina

10. Sir Galahad; Robin Hood; Sir Lancelot; William Tell

10.1 In the tales of King Arthur, who was the young knight whose exceptional purity and virtue enabled him to see the Holy Grail in all its splendour, while many other knights who sought it could not see it at all?

Sir Galahad

10.2 Who was the legendary hero who, famous for his skill as an archer, was forced to shoot an apple off of his own son's head?

William Tell

11. Dirham; Peso; Rupee; Shekel

11.1 Currency in Mexico is called a(n) _____.

peso

11.2 Currency in India is called a(n) _____.

rupee

12. Arthropods; Chordates; Echinoderms; Molluscs

12.1 What is the name of the phylum of invertebrates that contains snails, octopus, and squid?

Molluscs

12.2 What is the name of the phylum within the animal kingdom that includes shrimp, centipedes, insects, and spiders?

Arthropods

13. Aldous Huxley; George Orwell; Jonathan Swift; Kurt Vonnegut

13.1 Brave New World is a novel written by _____.

Aldous Huxley

13.2 Animal Farm is a novel written by _____.

George Orwell

14. Archimedes; Aristotle; Euclid; Ptolemy

14.1 Who was the ancient Greek mathematician, scientist, and inventor best known for his investigations of buoyancy?

Archimedes

14.2 Who was the ancient Greek mathematician that is considered the 'Father of Geometry'?

Euclid

15. Ag; Au; Fe; Hg

15.1 On the periodic table, what is the symbol for Iron?

Fe

15.2 On the periodic table, what is the symbol for Mercury?

Hg

16. Arigato; Obrigado; Sayonara; Xie xie

16.1 How do you say 'good-bye' in Japanese?

Sayonara

16.2 How do you say 'thank you' in Japanese?

Arigato

17. Diego Rivera; Francisco Goya; Pablo Picasso; Salvador Dali

17.1 _____ was a Spanish painter of the twentieth century, well known for cubism and his painting Guernica.

Pablo Picasso

17.2 _____ was a Spanish surrealist painter of the twentieth century, known for his iconic use of melting clocks.

Salvador Dali

18. Pulsar; Quasar; Red Giant; Supernova

18.1 What is the term used for the most distant galaxies seen from the Earth, which are often extremely luminous?

Quasar

18.2 What is the term for a large star in its death throes that suddenly explodes, bringing about a burst of radiation that often briefly outshines the entire galaxy?

Supernova.

19. Eli Whitney; Guglielmo Marconi; Robert Fulton; Thomas Edison

19.1 Who is widely credited with developing the first commercially successful steamboat?

Robert Fulton.

19.2 Who invented the cotton gin, a device for processing raw cotton?

Eli Whitney

20. Helsinki; Leningrad; Oslo; Stockholm

20.1 What is the capital of Norway?

Oslo

20.2 What is the capital of Finland?

Helsinki

21. Kampala; Addis Ababa; Lagos; Nairobi

21. 1 The capital of Kenya is _____.

Nairobi

21.2 The capital of Ethiopia is _____.

Addis Ababa

22. Haemophilia; Hepatitis; Hodgkin's Disease; Polio

22.1 What is the acute and infectious disease (one of the most dreaded childhood diseases of the 20th century) caused by a virus that brings about inflammation of certain nerve cells in the spinal cord?

Polio

22.2 What is the hereditary disease that is caused by a deficiency of a substance in the blood that aids in clotting?

Haemophilia

23. Aztec; Mayan; Incan; Yanomami

23.1 In the 16th century, the _____ empire was overthrown by the Spanish Conquistadores under Hernando Cortes.

Aztec

23.2 Francisco Pizarro overthrew the rulers of the _____ and established the nation of Peru.

Incan

24. Femur; Humerus; Tibia; Ulna

24.1 What is the name of the large bone in the upper leg?

Femur

24.2 What is the name of the large bone in the upper arm?

Humerus

25. Aegean Sea; Black Sea; Caspian Sea; Adriatic Sea

25.1 What body of water is enclosed mostly by Ukraine to the north, Russia to the east, and Turkey to the South?

Black Sea

25.2 What body of water is an arm of the Mediterranean Sea between Italy on the west and Croatia, Montenegro and Albania on the east?

Adriatic Sea

26. Sedimentary; Igneous; Metamorphic; Mantle

26.1 What kind of rock changes from one form to another due to heat or pressure?

Metamorphic

26.2 _____ rock is formed by the cooling and solidifying of molten materials brought from Earth's interior to the surface

Igneous

27. Epoxy; Gasohol; DDT; Napalm

27.1 What chemical weapon (developed in 1943 by Louis Fieser) is used in bombs and flamethrowers, burns intensely, and sticks to its target?

Napalm

27.2 What is the colourless insecticide that is poisonous when swallowed or absorbed through the skin (Rachel Carson wrote The Silent Spring detailing its environmental impact)?

DDT

28. Leif Ericson; Ferdinand Magellan; Christopher Columbus; Marco Polo

28.1 Who was the first man to sail around the Earth, although he was killed on the voyage?

Magellan

28.2 What explorer was one of the first Europeans to travel across Asia, later becoming a government official in China?

Marco Polo

29. S; Hz; Kg; Mol

29.1 The abbreviation for the SI base unit for time is _____.

S

29.2 The abbreviation for the SI base unit for amount of substance is _____.

Mol

30. Hey Jude; Bridge Over Troubled Water; Yer Blues; Imagine

30.1 In which of his hit singles does John Lennon sing of a world at peace and free of religious and national boundaries?

Imagine

30.2 Which 7-min-long Beatles song was written to comfort a child after his parents' divorce?

Hey Jude

31. The Canterbury Tales; The Arabian Nights; Aesop's Fables; Lord of the Flies

31.1 What collection of stories from the 14th century recounts tales about a group of pilgrims who meet at an inn near London?

The Canterbury Tales

31.2 What collection of stories recounts tales that, supposedly, queen Scheherazade told her husband?

The Arabian Nights

32. Mary; Elizabeth; Anne; Victoria

32.1 Queen _____ was the first woman to be crowned queen of England

Mary

32.2 Queen _____ ruled the United Kingdom during a time of economic and imperial expansion

Victoria

33. Caspian Sea; Superior; Tanganyika; Erie

33.1 The largest lake in the world is _____.

Caspian Sea

33.2 The largest lake in North America is _____.

Superior

Paper 1 - Experiments 2 and 3

Paper 2 - Experiments 1 and 2

General Knowledge MC Questions

1. diffraction; diffusion; distillation; entropy

1.1 In chemistry, the separating of the constituents of a liquid by boiling it and then condensing the vapor that results is called_____.

distillation

1.2 The spreading of atoms or molecules of one substance through those of another, especially into liquids and gases is known as _____.

diffusion

2. Centaur; Chimera; Minotaur; Satyr

2.1 In classical mythology, a creature that is half human and half horse is called a _____.

centaur

2.2 In classical mythology a creature that is half man and half bull is called a _____.

minotaur

3. Aphrodite; Athena; Isis; Venus

3.1 The Greek goddess of love is _____.

Aphrodite

3.2 The Roman Goddess of Love is _____.

Venus

4. Uranus; Jupiter; Saturn; Mercury

4.1 What is the largest planet in the Solar System?

Jupiter

4.2 What is the smallest planet in the Solar System?

Mercury

5. Hamlet; Julius Caesar; Merchant of Venice; Romeo and Juliet

5.1 From what Shakespearian play comes the line; 'This above all: to thine own self be true'?

Hamlet

5.2 From what Shakespearian play comes the phase 'pound of flesh'?

Merchant of Venice

6. All about Eve; Casablanca; Gone with the Wind; Sunset Blvd

6.1 From what classic movie comes the line, 'Frankly my dear, I don't give a damn'?

Gone with the Wind

6.2 From what classic movie comes the line, 'Here's looking at you, kid'?

Casablanca

7. ad hoc; ad nauseam; ergo; et cetera

7.1 The Latin translation of 'therefore' is _____.

ergo

7.2 The Latin translation of 'and so forth' is _____.

et cetera

8. Cirrus; Cumulus; Nimbus; Stratus

8.1 What is the term for lacy or wispy clouds that form at high altitudes, often before a change in the weather?

Cirrus

8.2 What is the term for large, white, puffy clouds that generally appear in fair weather, but that can also form thunderheads on hot days?

Cumulus

9. Anna Karenina; Brothers Karamazov; Crime and Punishment; War and Peace

9.1 _____ is a novel by Feodor Dostoevsky in which a man kills two old women because he believes that he is beyond the bounds of good and evil.

Crime and Punishment

9.2 _____ is a novel by Leo Tolstoy in which a woman enters a tragic adulterous affair and commits suicide by throwing herself under a train.

Anna Karenina

10. Sir Galahad; Robin Hood; Sir Lancelot; William Tell

10.1 In the tales of King Arthur, who was the young knight whose exceptional purity and virtue enabled him to see the Holy Grail in all its splendour, while many other knights who sought it could not see it at all?

Sir Galahad

10.2 Who was the legendary hero who, famous for his skill as an archer, was forced to shoot an apple off of his own son's head?

William Tell

11. Valencia; Madrid; Genoa; Oxford

11.1 Where was Lope de Vega born?

Madrid

11.2 Where was Christopher Columbus born?

Genoa

12. Arthropods; Chordates; Echinoderms; Molluscs

12.1 What is the name of the phylum of invertebrates that contains snails, octopus, and squid?

Molluscs

12.2 What is the name of the phylum within the animal kingdom that includes shrimp, centipedes, insects, and spiders?

Arthropods

13. Camilo Cela; Aldous Huxley; Jonathan Swift; Rafael Ferlosio

13.1 Brave New World is a novel written by _____.

Aldous Huxley

13.2 La Colmena is a novel written by _____.

Camilo Cela

14. Archimedes; Aristotle; Euclid; Ptolemy

14.1 Who was the ancient Greek mathematician, scientist; and inventor best known for his investigations of buoyancy?

Archimedes

14.2 Who was the ancient Greek mathematician that is considered the 'Father of Geometry'?

Euclid

15. Carrot; Passion fruit; Coconut; Banana

15.1 Granadilla is another name for_____.

Passion fruit

15.2 Nariyal is the Indian term for _____.

Coconut

16. Arigato; Obrigado; Sayonara; Xie xie

16.1 How do you say 'good bye' in Japanese?

Sayonara

16.2 How do you say 'thank you' in Japanese?

Arigato

17. Diego Rivera; Francisco Goya; Pablo Picasso; Salvador Dali

17.1 _____ was a Spanish painter of the twentieth century, well known for cubism and his painting Guernica.

Pablo Picasso

17.2 _____ was a Spanish surrealist painter of the twentieth century, known for his iconic use of melting clocks.

Salvador Dali

18. Pulsar; Quasar; Red Giant; Supernova

18.1 What is the term used for the most distant galaxies seen from the Earth, which are often extremely luminous?

Quasar

18.2 What is the term for a large star in its death throes that suddenly explodes; bringing about a burst of radiation that often briefly outshines the entire galaxy?

Supernova

19. Eli Whitney; Guglielmo Marconi; Robert Fulton; Thomas Edison

19.1 Who is widely credited with developing the first commercially successful steamboat?

Robert Fulton

19.2 Who invented the cotton gin, a device for processing raw cotton?

Eli Whitney

20. Helsinki; Leningrad; Oslo; Stockholm

20.1 What is the capital of Norway?

Oslo

20.2 What is the capital of Finland?

Helsinki

21. Kampala; Addis Ababa; Lagos; Nairobi

21.1 The capital of Kenya is _____.

Nairobi

21.2 The capital of Ethiopia is _____.

Addis Ababa

22. Haemophilia; Hepatitis; Hodgkin's Disease; Polio

22.1 What is the acute and infectious disease (one of the most dreaded childhood diseases of the 20th century) caused by a virus that brings about inflammation of certain nerve cells in the spinal cord?

Polio

22.2 What is the hereditary disease that is caused by a deficiency of a substance in the blood that aids in clotting?

Haemophilia

23. Aztec; Mayan; Incan; Yanomami

23.1 In the 16th century, the _____ empire was overthrown by the Spanish Conquistadores under Hernando Cortes.

Aztec

23.2 Francisco Pizarro overthrew the rulers of the _____ and established the nation of Peru.

Incan

24. Epidermis; Vertebrae; Cerebrum; Nostrils

24.1 What is the name of the biggest part of the human brain?

Cerebrum

24.2 The outside layer of skin on the human body is called the?

Epidermis

25. Aegean Sea; Black Sea; Caspian Sea; Adriatic Sea

25.1 What body of water is enclosed mostly by Ukraine to the north, Russia to the east; and Turkey to the South?

Black Sea

25.2 What body of water is an arm of the Mediterranean Sea between Italy on the west and Croatia, Montenegro and Albania on the east?

Adriatic Sea

26. Sedimentary; Igneous; Metamorphic; Mantle

26.1 What kind of rock changes from one form to another due to heat or pressure?

Metamorphic

26.2 _____ rock is formed by the cooling and solidifying of molten materials brought from Earth's interior to the surface

Igneous

27. Epoxy; Gasohol; DDT; Napalm

27.1 What chemical weapon (developed in 1943 by Louis Fieser) is used in bombs and flamethrowers; burns intensely, and sticks to its target?

Napalm

27.2 What is the colourless insecticide that is poisonous when swallowed or absorbed through the skin (Rachel Carson wrote The Silent Spring detailing its environmental impact)?

DDT

28. Leif Ericson; Ferdinand Magellan; Christopher Columbus; Marco Polo

28.1 Who was the first man to sail around the Earth, although he was killed on the voyage?

Magellan

28.2 What explorer was one of the first Europeans to travel across Asia later becoming a government official in China?

Marco Polo

29. S; Hz; Kg; Mol

29.1 The abbreviation for the SI base unit for time is _____.

S

29.2 The abbreviation for the SI base unit for amount of substance is _____.

Mol

30. Hey Jude; Bridge Over Troubled Water; Yer Blues; Imagine

30.1 In which of his hit singles does John Lennon sing of a world at peace and free of religious and national boundaries?

Imagine

30.2 Which 7-min-long Beatles song was written to comfort a child after his parents' divorce?

Hey Jude

31.The Canterbury Tales; The Arabian Nights; Aesop's Fables; Lord of the Flies

31.1 What collection of stories from the 14th century recounts tales about a group of pilgrims who meet at an inn near London?

The Canterbury Tales

31.2 What collection of stories recounts tales that, supposedly, queen Scheherazade told her husband?

The Arabian Nights

32. Mary; Elizabeth; Anne; Victoria

32.1 Queen _____ was the first woman to be crowned queen of England

Mary

32.2 Queen _____ ruled the United Kingdom during a time of economic and imperial expansion

Victoria

33. Caspian Sea; Superior; Tanganyika; Erie

33.1 The largest lake in the world is _____.

Caspian Sea

33.2 The largest lake in North America is _____.

Superior

Paper 3 - Experiment 1

SAT Questions

1. tundra; tropical rain forest; taiga; desert

1.1 Northern areas, characterized by permafrost, extremely cold temperatures, and few trees, would be classified as _____.

Tundra

1.2 Coniferous forests, characterized by long, cold winters and short, wet summers, would be classified as _____.

Taiga

2. Thyroid; Pancreas; Parathyroid; Adrenal medulla

2.1 _____ is an organ that secretes the hormone responsible for the flight-or-flight response.

Adrenal medulla

2.2 _____ is an organ that organ secretes a hormone that causes the liver to break down glycogen.

Pancreas

3. Persian Gulf War; Vietnam War; World War I; Spanish-American War

3.1 "Hawks" and "doves" were nicknames given to those who supported and opposed which war?

Vietnam War

3.2 The cry "Remember the Maine!" is associated with which war?

Spanish-American War

4. Concave mirror; Convex mirror; Concave lens; Flat mirror

4.1 An object is placed in front of an optical device, and an image is obtained. Which device would produce an image that is erect, virtual, and reversed from left to right?

Flat mirror

4.2 An object is placed in front of an optical device, and an image is obtained. Which device would produce an image that is inverted, real, and on the same side of the device?

Concave mirror

5. Aztec; Inca; Maya; Olmec

5.1 The creation story contained in the Popol Vuh was written by which culture?

Maya

5.2 In 1521 Hernán Cortés led the Spanish conquest of which Mesoamerican culture?

Aztec

6. Oxygen; Carbon; Nitrogen; Argon

6.1 Which of the following is the third most abundant gas in Earth's atmosphere?

Argon

6.2 Which of the following that its allotrope is the primary absorber of UV solar radiation in Earth's atmosphere?

Oxygen

7. Animalia; Eubacteria; Fungi; Protista

7.1 A unicellular prokaryote belongs to which kingdom?

Eubacteria

7.2 Paramecia are members of which kingdom?

Protista

8. Massachusetts; New York; Virginia; Pennsylvania

8.1 The first shots of the Revolutionary War were fired in which colony?

Massachusetts

8.2 Which colony wrote the 1776 resolution stating that "these united colonies, are, and of right ought to be, free and independent states"?

Virginia

9. pressure decreases; temperature decreases; volume increases; temperature increases

9.1 When a gas undergoes an adiabatic expansion, its _____.

temperature decreases

9.2 When a gas undergoes an adiabatic compression, its _____.

temperature increases

10. Socrates; Aristotle; Epicurus; Plato

10.1 The Greek philosopher who supported the concept of a state ruled by philosophers and built on a structure of class divisions was _____.

Plato

10.2 The Greek philosopher who developed a teaching methodology based on systematic questioning was _____.

Socrates

11. Molality; Mole fraction; Density; Partial pressure

11.1 Which of the following is measured in units of moles/kilogram?

Molality

11.2 Which of the following is a measure of mass per unit volume?

Density

12. A negative; O negative; B positive; AB positive

12.1 People with an A-positive blood type may safely donate blood to those with which blood type?

AB positive

12.2 People with an O-positive blood type may safely receive blood from those with which blood type?

O negative

13. Franklin D. Roosevelt; Lyndon B. Johnson; Harry S. Truman; John F. Kennedy

13.1 Which United States president's domestic programs were collectively known as the Great Society?

Lyndon B. Johnson

13.2 Which United States president's domestic and foreign policy programs were collectively known as the New Frontier?

John F. Kennedy

14. Gamma rays; Ultraviolet; Radio; X-rays

14.1 Which of the following has the shortest wavelength?

Gamma rays

14.2 Which of the following has the longest wavelength?

Radio

15. Rome; Greece; Mesopotamia; Egypt

15.1 Which civilization's code of law was written in the Twelve Tables?

Rome

15.2 From which region did the cuneiform method of writing emerge?

Mesopotamia

16. Nuclear fusion; Alpha decay; Positron emission; Nuclear fission

16.1 What is the principal reaction that is responsible for the energy output of the sun?

Nuclear fusion

16.2 What is responsible for most helium found on Earth?

Alpha decay

17. Imprinting; Reasoning/insight; Classical conditioning; Habituation

17.1 Geese recognize a ticking clock as "mother" if exposed to it during a critical period shortly after hatching, this is known as _____.

Imprinting

17.2 Fish are given food at the same time as a tap on their glass bowl and soon learn to approach when a tap sounds even in the absence of food, this is known as _____.

Classical conditioning

18. Frog; Turtle; Lizard; Sparrow

18.1 Which of the following organisms is able to regulate its own body temperature?

Sparrow

18.2 Which of the following organisms has skin that serves as an accessory organ of respiration?

Frog

19. Opera; Ballet; Vaudeville; Musical comedy

19.1 The huge influx of European immigrants throughout the nineteenth century helped give rise to which of the following forms of entertainment in the United States?

Opera

19.2 Which of the following describes the most popular form of stage entertainment during the late 1800s?

Vaudeville

20. Displacement; Acceleration; Linear momentum; Kinetic energy

20.1 Which quantity can be expressed in the same units as impulse?

Linear momentum

20.2 Which quantity could remain constant if an object's speed is changing?

Acceleration

21. Hinduism; Sikhism; Buddhism; Shintoism

21.1 The Four Noble Truths form the basis of which major Eastern religion?

Buddhism

21.2 The Upanishads contain the core philosophy and central teachings of which world religion?

Hinduism

22. Balance; Funnel; Barometer; Condenser

22.1 _____ is commonly used in the laboratory in a distillation setup.

Condenser

22.2 _____ is commonly used in the laboratory in a filtration setup.

Funnel

23. Small intestine; Large intestine; Stomach; Mouth

23.1 _____ is a structure where most digestion and absorption of nutrients occurs

Small intestine

23.2 _____ is a structure where starch digestion first takes place

Mouth

24. Democratic Party; Republican Party; Progressive Party; Populist Party

24.1 Which political party was formed during the 1890s to address the concerns of U.S. farmers?

Populist Party

24.2 Which political party passed a "gag rule" in 1837 that banned any discussion of abolition in Congress?

Democratic Party

25. Frequency; Amplitude; Wavelength; Velocity

25.1 _____ is described as the number of wave crests passing a given point per unit of time.

Frequency

25.2 _____ is described as the distance between two points or two consecutive waves.

Wavelength

26. salt; spices; iron; grain

26.1 The ancient African city of Meroe became a major trading centre between 250 BCE and 150 CE due to production of _____.

iron

26.2 Aside from gold, the most valuable and highly traded commodity in the West African kingdom of Ghana was _____.

salt

27. A solid; Density; Volume; Weight

27.1 _____ has mass and a definite size and shape.

A solid

27.2 _____ gives the space occupied.

Volume

28. scurvy; pernicious anaemia; pellagra; bleeding

28.1 Severe vitamin K deficiency may cause _____.

bleeding

28.2 Severe vitamin C deficiency may cause _____.

scurvy

29. Willa Cather; Mark Twain; Jack London; Edith Wharton

29.1 Which of the following American regional novelists is known for his or her stories of the Great Plains states?

Willa Cather

29.2 Which of the following novelists wrote mainly about the social class whose lifestyle Thorstein Veblen described in 1899 as the embodiment of "conspicuous consumption"?

Edith Wharton

30. air; metal; liquid; vacuum

30.1 Sound waves cannot travel through _____.

vacuum

30.2 S-waves waves cannot travel through _____.

liquid

31. Dutch; French; Germans; British

31.1 In the mid-1800s, the south-eastern port of Singapore was controlled by the _____.

British

31.2 For the majority of the eighteenth century, South Africa was controlled by the _____.

Dutch

32. Ionic bond; Nonpolar covalent bond; Polar covalent bond; Metallic bond

32.1 What is the type of bond between the atoms in a nitrogen molecule?

Nonpolar covalent bond

32.2 What is the type of bond between the atoms of calcium in a crystal of calcium?

Metallic bond

33. Hair; Epidermis; Guard cell; Cuticle

33.1 The layer that restricts evaporation in humans is _____.

Epidermis

33.2 The layer that restricts evaporation in plants is _____.

Cuticle

The Repeated Items Replacement (The No Repetition condition)

1. Australian Desert; Arabian Desert; The Sahara Desert; Colorado Plateau

Which desert is the largest in the world?

The Sahara Desert

2. Charles Darwin; Isaac Newton; Thomas Edison; Albert Einstein

Which famous scientist introduced the idea of natural selection?

Charles Darwin

3. Venus; Mars; Jupiter; Neptune

What planet is nicknamed the ‘Red Planet’?

Mars

4. Decibels; Kelvin; Watts; Richter

Electric power is typically measured in what units?

Watts

5. Indian Ocean; Atlantic Ocean; Pacific Ocean; Arctic Ocean

What is the location of Bermuda triangle?

Atlantic Ocean

6. Macbeth; Hamlet; Romeo & Juliet; Othello

_____ is Shakespeare's shortest tragedy.

Macbeth

7. Pigeon; Hummingbird; Rabbit; Cat

What animal is associated with ancient Egypt?

Cat

8. China; France; Italy; Germany

Which country gifted the Statue of Liberty to the U.S.?

France

9. H; Ag; Zn; Fe

_____ is the chemical symbol for Iron.

Fe

10. Vladimir Lenin; Joseph Stalin; Karl Marx; Emile Durkheim

The idea of Socialism was articulated and advanced by whom?

Karl Marx

11. Reykjavik; Istanbul; Moscow; Madrid

What is the only major city located on two continents?

Istanbul

General Knowledge Questions (The Filler Task in One Test Condition)

1. Epidermis; Vertebrae; Cerebrum; Nostrils

What is the name of the biggest part of the human brain?

Cerebrum

2. Cirrus; Cumulus; Nimbus; Stratus

What is the term for lacy or wispy clouds that form at high altitudes, often before a change in the weather?

Cirrus

3. Valencia; Madrid; Genoa; Oxford

Where was Christopher Columbus born?

Genoa

4. Uranus; Jupiter; Saturn; Mercury

What is the smallest planet in the Solar System?

Mercury

5. Mary; Elizabeth; Anne; Victoria

Queen _____ ruled the United Kingdom during a time of economic and imperial expansion

Victoria

6. Anna Karenina; Brothers Karamazov; Crime and Punishment; War and Peace

_____ is a novel by Leo Tolstoy in which a woman enters a tragic adulterous affair and commits suicide by throwing herself under a train.

Anna Karenina

Paper 3 - Experiment 2

Introductory Psychology Questions

1. behaviourism; humanism; structuralism; social psychology

1.1 The approach for investigating psychological phenomena Wundt adopted was called _____.

structuralism

1.2 The approach to psychology that stresses positive growth and self-realisation is called _____.

humanism

2. client-centred therapy; the hierarchy of needs; functionalism; classical conditioning

2.1 The idea of _____ can be traced to Rogers.

client-centred therapy

2.2 The idea of _____ can be traced to Maslow.

the hierarchy of needs

3. absolute distance; retinal disparity; convergence; relative distance

3.1 When looking at an object, having a slightly different retinal image for each eye is known as _____.

retinal disparity

3.2 As an object gets closer, we move both eyes inward to keep it in focus. The muscle strain associated with this movement is a cue to depth and is known as _____.

convergence

4. motion parallax; relative size; depth perception; binocular depth

4.1 When driving a car, objects near the car appear to move faster than objects further away. This is an example of _____

motion parallax

4.2 When looking at a picture, objects in the front appear to be larger but we automatically scale size according to distance. This is an example of _____

relative size

5. occlusion; texture gradient; linear perspective; aerial perspective

5.1 When overlooking a natural scene from the top of a mountain, the distant objects seem to be fuzzy. This is an example of _____.

aerial perspective

5.2 When looking at a railway, the two tracks appear to converge on the horizon even though they are parallel. This is an example of _____.

linear perspective

6. Ponzo; Ames room; Muller-Lyer; Moon

6.1 The _____ illusion uses converging parallel lines to suggest that one object is further away than another and appears to be larger when in fact the two objects are the same size.

Ponzo

6.2 The _____ illusion manipulates depth cues to suggest that two people of the same size are the same distance from the observer when in fact they are not, which makes one person look larger than the other.

Ames room

7. motion after-effect; induced motion; Phi-phenomenon; overall intensity

7.1 The feeling of that we are moving on a stationary train when an adjacent train starts to move is an example of _____.

induced motion

7.2 The feeling of moving backward slightly after moving forward for some time and then stopping is an example of _____.

motion after-effect

8. closure; good continuation; proximity; figure-ground

8.1 The Gestalt principle of _____ allows you to fill in the gaps of a radio announcer who is competing with a lot of static that cuts out several of his words.

closure

8.2 The Gestalt principle of _____ states that elements that are arranged on a line or curve are perceived to be more related than elements not on the line or curve.

good continuation

9. frequency theory; place theory; frequency-volley theory; opponent-process theory

9.1 That we can distinguish between tones above 1000Hz, but cells cannot fire any faster at 1000Hz, is an issue for _____.

frequency theory

9.2 That we can distinguish between tones below 1000Hz, but no specific place on the basilar membrane is vibrating more than any other at 1000Hz, is a problem for _____.

place theory

10. response bias; absolute threshold; difference threshold; minimum threshold

10.1 The _____ is the minimum difference between two stimuli that can be detected 50% of the time.

difference threshold

10.2 The _____ is the minimum value of a stimulus that can be detected 50% of the time.

absolute threshold

11. short-sightedness; long-sightedness; dichromatic colour-blindness; monochromatic colour-blindness

11.1 Those with _____ have only two types of iodopsin.

dichromatic colour-blindness

11.2 Those with _____ have only one type of iodopsin.

monochromatic colour-blindness

12. sclera; rhodopsin; iodopsin; pheromones

12.1 _____ is a chemical contained in the rods and functions mainly in dim light.

rhodopsin

12.2 _____ is a chemical contained in the cones and functions mainly in bright light.

iodopsin

13. endorphins; ganglion cells; androstenone; substance P

13.1 When one feels pain, _____ reduce(s) it.

endorphins

13.2 When one feels pain, _____ produce(s) it.

substance P

14. photoreceptor layer; optic disk; cerebral hemisphere; optic chiasm

14.1 The place where optic nerve leaves the eye is known as the _____.

optic disk

14.2 The place where all the information from the eye meets before being rerouted to the thalamus is known as the _____.

optic chiasm

15. a chunking process; deep processing; shallow processing; a rehearsal process

15.1 Processing for meaning, resulting in better long-term memory, is _____.

deep processing

15.2 Processing of superficial features, resulting in poor long-term memory, is _____.

shallow processing

16. elaborative; deep; maintenance; shallow

16.1 Trying to memorise a phone number by keeping it active in short-term memory is an example of _____ rehearsal.

maintenance

16.2 Trying to memorise some information by relating it to information already in memory is called _____ rehearsal.

elaborative

17. episodic memory; semantic memory; sensory memory; iconic memory

17.1 _____ is context-specific memory which encodes time and place.

episodic memory

17.2 _____ is general knowledge about the world.

semantic memory

18. schema-driven; trace-driven; iconic; echoic

18.1 _____ memory retrieval depends on a set of expectations and it is not a literal re-experiencing of the past.

schema-driven

18.2 _____ memory retrieval is a literal re-experiencing of the past.

trace-driven

19. coerced-internalised; coerced compliant; voluntary; transformed

19.1 When innocent people confess to committing a crime with external pressure but they believe they are guilty, it is a _____ false confession

coerced-internalised

19.2 When innocent people confess to committing a crime without external pressure and do not believe they are guilty, it is a _____ false confession.

voluntary

20. confidence; correlation; calibration; resolution

20.1 If we look at whether our confidence ratings and our accuracy match, we are assessing metacognitive _____.

calibration

20.2 If we look at whether our confidence and our accuracy are correlated, we are assessing metacognitive _____.

resolution

21. plurality; grain-size; report; judgment of learning

21.1 Response accuracy can be regulated by the _____ option, which is providing several alternatives as an answer instead of one.

plurality

21.2 Response accuracy can be regulated by the _____ option, which is withholding answers that you are not confident about.

report

22. Kant; Descartes; Locke; Asch

22.1 _____ claimed that knowledge acquisition comes through experience which is structured through innate schemata.

Kant

22.2 _____ claimed that some knowledge is innate.

Descartes

23. negativity bias; positivity bias; Halo effect; primacy effect

23.1 The _____ is a bias in impression formation where earlier information has a stronger influence than later information.

primacy effect

23.2 The _____ is a bias in impression formation where negative information has a stronger influence than positive information.

negativity bias

24. perceptual; cognitive; emotional; behavioural

24.1 Stereotypes are the _____ aspect of group schemas.

cognitive

24.2 Discrimination is the _____ aspect of group schemas.

behavioural

25. Self-perception; Social comparison; Self-fulfilling prophecy; self-discrepancy

25.1 A _____ is when others' expectations about us cause us to behave in a way that confirms those expectations.

Self-fulfilling prophecy

25.2 _____ theory states that when others who are similar to us agree with us, it gives us confidence in the validity of our perceptions, attitudes, and behaviours.

Social comparison

26. consistency; covariation; consensus; distinctiveness

26.1 Based on Kelley's covariation theory, we evaluate _____ when considering whether X behave like most people or only a few of them.

consensus

26.2 Based on Kelley's covariation theory, we evaluate _____ when considering whether X behaves in this way in all situations or only in specific ones.

distinctiveness

27. false consensus; self-serving bias; actor-observer effect; fundamental attribution error

27.1 The _____ is the tendency to consider the cause of others' behaviour to be underlying and unchangeable properties of people.

fundamental attribution error

27.2 The _____ is the tendency to attribute one's success to dispositional characteristics, and failures to situational factors.

self-serving bias

28. anchoring and adjusting; availability; representativeness; language

28.1 The tendency to think that there are more words that begin with R than with R in the third position, is an example of a cognitive shortcut termed the _____ heuristic.

availability

28.2 The tendency to classify something as belonging to a certain category because of its similarity to the typical case is a cognitive shortcut termed the _____ heuristic.

representativeness

29. cognitive; affective; behavioural; perceptual

29.1 One component of attitudes is the _____ component, which describes our feeling toward something.

affective

29.2 One component of attitudes is the _____ component, which describes our beliefs about something.

cognitive

30. Observational learning; Operant conditioning; Classical conditioning; Mere exposure

30.1 _____ can lead to attitude formation, such as when a picture of a lovely home is paired repeatedly with a product so that over time the product becomes likeable.

Classical conditioning

30.2 _____ can lead to attitude formation, such as when repeated listening to a song increase liking of it.

Mere exposure

31. pride; contempt; fear; joy

31.1 _____ is an example of a self-conscious emotion.

pride

31.2 _____ is an example of a moral emotion.

contempt

32. dorsolateral cortex; hippocampus; orbitofrontal cortex; amygdala

32.1 Damage to the _____ is associated with a lack of fear responses in typically fearful situations.

amygdala

32.2 Damage to the _____ is associated with a problem of recognising facial and vocal emotional expression.

orbitofrontal cortex

33. two-factor theory; cannon-bard theory; James-Lange theory; misattribution of arousal paradigm

33.1 The _____ suggested that particular emotions were experienced following a unique pattern of autonomic arousal.

James-Lange theory

33.2 The _____ suggested that emotional and physiological responses are separate.

cannon-bard theory

Paper 3 - Experiment 3

Introductory Psychology Questions

1. absolute distance; retinal disparity; convergence; relative distance

1.1 When looking at an object, we may have a slightly different retinal image for each eye. This is known as _____.

retinal disparity

1.2 When looking at an object that gets closer, we move both eyes inward to keep it in focus. This is known as _____.

convergence

2. depth perception; motion parallax; relative size; binocular depth

2.1 When driving a car, objects near the car appear to move faster than objects further away. This is an example of _____

motion parallax

2.2 When looking at a picture, objects in the front appear to be larger but we automatically scale size according to distance. This is an example of _____.

relative size

3. occlusion; texture gradient; linear perspective; aerial perspective

3.1 Distant objects that look fuzzy due to moisture and particles in the air. This is called _____.

aerial perspective

3.2 Distant objects that look denser less details and closer together. This is called _____.

texture gradient

4. Muller-Lyer; Ponzo; Ames room; Moon

4.1 The _____ illusion uses converging parallel lines to suggest that one object is further away than another and appears to be larger when in fact the two objects are the same size.

Ponzo

4.2 The _____ illusion manipulates depth cues to suggest that two people of the same size are the same distance from the observer when in fact they are not, which makes one person look larger than the other.

Ames room

5. overall intensity; motion after-effect; induced motion; Phi-phenomenon

5.1 The feeling of that we are moving on a stationary train when an adjacent train starts to move is an example of _____.

induced motion

5.2 The feeling of moving backward slightly after moving forward for some time and then stopping is an example of _____.

motion after-effect

6. figure-ground; closure; proximity; good continuation

6.1 The Gestalt principle of _____ allows you to fill in the gaps of a radio announcer who is competing with a lot of static that cuts out several of his words.

closure

6.2 The Gestalt principle of _____ states that elements that are arranged on a line or curve are perceived to be more related than elements not on the line or curve.

good continuation

7. opponent-process theory; frequency-volley theory; frequency theory; place theory

7.1 That we can distinguish between tones above 1000Hz, but cells cannot fire any faster at 1000Hz, is an issue for _____.

frequency theory

7.2 That we can distinguish between tones below 1000Hz, but no specific place on the basilar membrane is vibrating more than any other at 1000Hz, is an issue for _____.

place theory

8. response bias; absolute threshold; difference threshold; minimum threshold

8.1 The _____ is the minimum difference between two stimuli that can be detected 50% of the time.

difference threshold

8.2 The _____ is the minimum value of a stimulus that can be detected 50% of the time.

absolute threshold

9. short-sightedness; long-sightedness; dichromatic colour-blindness; monochromatic colour-blindness

9.1 Those with _____ have only two types of iodopsin.

dichromatic colour-blindness

9.2 Those with _____ have only one type of iodopsin.

monochromatic colour-blindness

10. sclera; rhodopsin; iodopsin; pheromones

10.1 _____ is a chemical contained in the rods and functions mainly in dim light.

rhodopsin

10.2 _____ is a chemical contained in the cones and functions mainly in bright light.

iodopsin

11. androstenone; endorphins; ganglion cells; substance P

11.1 When one feels pain, _____ reduce(s) it.

endorphins

11.2 When one feels pain, _____ produce(s) it.

substance P

12. photoreceptor layer; optic disk; cerebral hemisphere; optic chiasm

12.1 The place where optic nerve leaves the eye is known as the _____.

optic disk

12.2 The place where all the information from the eye meets before being rerouted to the thalamus is known as the _____.

optic chiasm

13. shallow; elaborative; deep; maintenance

13.1 Trying to memorise a phone number by keeping it active in short-term memory is an example of _____ rehearsal.

maintenance

13.2 Trying to memorise some information by relating it to information already in memory is called _____ rehearsal.

elaborative

14. sensory memory; episodic memory; semantic memory; iconic memory

14.1 _____ contains information that is personally meaningful to us.

episodic memory

14.2 _____ contains information that is based on general knowledge.

semantic memory

15. echoic; iconic; schema-driven; trace-driven

15.1 _____ memory retrieval depends on expectations and it is not a literal re-experiencing of the past.

schema-driven

15.2 _____ memory retrieval is a literal re-experiencing of the past.

trace-driven

16. coerced compliant; voluntary; transformed; coerced-internalised

16.1 When innocent people confess to committing a crime with external pressure but they believe they are guilty, it is a _____ false confession

coerced-internalised

16.2 When innocent people confess to committing a crime without external pressure and do not believe they are guilty, it is a _____ false confession.

voluntary

17. confidence; correlation; calibration; resolution

17.1 If we look at whether our confidence ratings and our accuracy match, we are assessing metacognitive _____.

calibration

17.2 If we look at whether our confidence and our accuracy are correlated, we are assessing metacognitive _____.

resolution

18. grain-size; plurality; report; judgment of learning

18.1 Response accuracy can be regulated by the _____ option, which is providing several alternatives as an answer instead of one.

plurality

18.2 Response accuracy can be regulated by the _____ option, which is withholding answers that you are not confident about.

report

19. Asch; Locke; Kant; Descartes

19.1 _____ claimed that knowledge acquisition comes through experience which is structured through innate schemata.

Kant

19.2 _____ claimed that some knowledge is innate.

Descartes

20. Discrimination; Prejudice; Attitude; Stereotypes

20.1 _____ is(are) the emotional aspect of group schemas.

Prejudice

20.2 _____ is(are) the cognitive aspect of group schemas.

Stereotypes

21. Self-perception; Social comparison; Self-fulfilling prophecy; self-discrepancy

21.1 One of the ways to acquire self-knowledge is when others' expectations about us cause us to behave in a way that confirms those expectations. This is known as _____.

Self-fulfilling prophecy

21.2 One of the ways to acquire self-knowledge is when others who are similar to us agree with us which gives us confidence in the validity of our behaviours This is known as _____.

Social comparison

22. consistency; covariation; consensus; distinctiveness

22.1 Based on Kelley's covariation theory, we evaluate _____ when considering whether X behave like most people or only a few of them.

consensus

22.2 Based on Kelley's covariation theory, we evaluate _____ when considering whether X behaves in this way in all situations or only in specific ones.

distinctiveness

23. false consensus; self-serving bias; actor-observer effect; fundamental attribution error

23.1 The _____ is the tendency to attribute the cause of others' behaviour to unchangeable properties of people.

fundamental attribution error

23.2 The _____ is the tendency to attribute one's success to dispositional characteristics, and failures to situational factors.

self-serving bias

24. anchoring and adjusting; availability; representativeness; language

24.1 The tendency to think that things that are easy to imagine occur more frequently than things that are difficult to imagine, is a cognitive shortcut termed the _____

availability

24.2 The tendency to classify something as belonging to a certain category because of its similarity to the typical case is a cognitive shortcut termed the _____ heuristic.

representativeness

25. behavioural; perceptual; cognitive; affective

25.1 One component of attitudes is the _____ component, which describes our feeling toward something.

affective

25.2 One component of attitudes is the _____ component, which describes our beliefs about something.

cognitive

26. Observational learning; Operant conditioning; Classical conditioning; Mere exposure

26.1 _____ can lead to attitude formation, such as when a picture of a lovely home is paired repeatedly with a product so that over time the product becomes likeable.

Classical conditioning

26.2 _____ can lead to attitude formation, such as when repeated listening to a song increases liking of it.

Mere exposure

27. fear; pride; contempt; joy

27.1 _____ is an example of a self-conscious emotion.

pride

27.2 _____ is an example of a moral emotion.

contempt

28. dorsolateral cortex; hippocampus; orbitofrontal cortex; amygdala

28.1 Damage to the _____ is associated with a lack of fear responses in typically fearful situations.

amygdala

28.2 Damage to the _____ is associated with a problem of recognising facial and vocal

orbitofrontal cortex

29. two-factor theory; cannon-bard theory; James-Lange theory; misattribution of arousal paradigm

29.1 The _____ suggested that particular emotions were experienced following a unique pattern of autonomic arousal.

James-Lange theory

29.2 The _____ suggested that emotional and physiological responses are separate.

cannon-bard theory

30. Conformity; Consistency; Obedience; Compliance

30.1 _____ is change of behaviour in response to a directive from people in authority.

Obedience

30.2 _____ is change of behaviour in response to a direct request from someone.

Compliance

31. low-balling; face-in-the-door; door-in-the-face; foot-in-the-door

31.1 The _____ technique begins with making a small request, which is then increased to a larger request.

foot-in-the-door

31.2 The _____ technique begins with making a large request, which is then lowered to a smaller request.

door-in-the-face

32. judgmental; informational; normative; referent

32.1 Being influenced by the behaviour of others because we believe they understand a situation better than we do is _____ conformity.

informational

32.2 Being influenced by the behaviour of others in order to be liked and accepted by them is _____ conformity.

normative

33. influence; facilitation; loafing; inhibition

33.1 Social _____ refers to improved task performance in the presence of others.

facilitation

33.2 Social _____ refers to decreased task performance in the presence of others.

inhibition

34. generalisation; decategorisation; recategorization; specification

34.1 Shifting attention away from overall outgroup characteristics to the individual characteristics of outgroup members can reduce prejudice and is termed _____.

decategorisation

34.2 Focusing on common membership in a superordinate group instead of differences between subgroups can reduce prejudice and is termed _____.

recategorisation

35. Formal operational; Concrete operational; Preoperational; Sensorimotor

35.1 The period in which a child first grasps the concept of object permanence is the _____ period.

Sensorimotor

35.2 The period in which a child increases the ability to represent objects symbolically is the _____ period.

Preoperational

36. conditioning; adjustment; assimilation; accommodation

36.1 New information producing new schemata or changing an existing schema is termed _____.

accommodation

36.2 New information being modified to fit into an existing schema is termed _____.

assimilation

37. Zone of proximal; Theory of mind; Conservation problem; Egocentrism

37.1 _____ is the belief that others see the world exactly like oneself.

Egocentrism

37.2 _____ is the ability to imagine what other people are thinking.

Theory of mind

38. adaptation reaction stage; stage of exhaustion; stage of resistance; alarm reaction stage

38.1 The stage of the general adaptation syndrome at which the organism may experience shock is called the _____.

alarm reaction stage

38.2 The stage of the general adaptation syndrome at which the body is vulnerable to illness is called the _____.

stage of exhaustion

39. reaction-focused coping; emotion-focused coping; problem-focused coping; cognition-focused coping

39.1 Coping that is directed towards one's own personal reaction to a stressor is _____.

emotion-focused coping

39.2 Coping that is directed towards the source of stress is _____.

problem-focused coping

40. emotional; cognitive; behavioural; social

40.1 Obsessive compulsive disorder consists of obsession which is the _____ component.

cognitive

40.2 Obsessive compulsive disorder consists of compulsion which is the _____ component.

behavioural

41. Dissociative disorders; Somatoform disorders; Conversion Disorders; Personality disorders

41.1 _____ are a class of somatoform disorder that involves complaints of wide-ranging physical ailments without apparent biological basis.

Somatoform disorders

41.2 _____ are a class of somatoform disorder that involves physical complaints that resemble neurological disorders but without underlying organic causes.

Conversion Disorders

42. dissociative; dramatic; anxious; eccentric

42.1 Antisocial, borderline, and narcissistic personality disorders belong to Cluster B, which can be described as _____.

dramatic

42.2 Paranoid, schizoid, and schizotypal personality disorders belong to Cluster A, which can be described as _____.

eccentric

43. Selfishness; Alogia; Delusions; Obsessions

43.1 _____ is (are) one of the positive symptoms of the schizophrenic disorders based on the diagnostic criteria of DSM.

Delusions

43.2 _____ is (are) one of the negative symptoms of the schizophrenic disorders based on the diagnostic criteria of DSM.

Alogia

44. Gestalt therapy; psychoanalysis therapy; cognitive behavioural therapy; client-centred therapy

44.1 The goal of _____ is for patients to change maladaptive behaviour, and negative thoughts and feelings.

cognitive behavioural therapy

44.2 The goal of _____ is for patients to gain understanding of their unique potential for personal growth.

client-centred therapy