Mem Cogn
DOI 10.3758/s13421-014-0452-8

Optimizing multiple-choice tests as tools for learning

Jeri L. Little · Elizabeth Ligon Bjork

Abstract Answering multiple-choice questions with competitive alternatives can enhance performance on a later test, not only on questions about the information previously tested, but also on questions about related information not previously tested—in particular, on questions about information pertaining to the previously incorrect alternatives. In the present research, we assessed a possible explanation for this pattern: When multiple-choice questions contain competitive incorrect alternatives, test-takers are led to retrieve previously studied information pertaining to all of the alternatives in order to discriminate among them and select an answer, with such processing strengthening later access to information associated with both the correct and incorrect alternatives. Supporting this hypothesis, we found enhanced performance on a later cued-recall test for previously nontested questions when their answers had previously appeared as competitive incorrect alternatives in the initial multiple-choice test, but not when they had previously appeared as noncompetitive alternatives. Importantly, however, competitive alternatives were not more likely than noncompetitive alternatives to be intruded as incorrect responses, indicating that a general increased accessibility for previously presented incorrect alternatives could not be the explanation for these results. The present findings, replicated across two experiments (one in which corrective feedback was provided during the initial multiple-choice testing, and one in which it was not), thus strongly suggest that competitive multiple-choice questions can trigger beneficial retrieval processes for both tested and related information, and the results have implications for the effective use of multiple-choice tests as tools for learning.

Keywords Memory · Educational psychology · Retrieval processes · Learning · Multiple-choice tests · Testing effects

In addition to assessing one's knowledge, taking a test involving recall can also improve one's retention of the tested information. Such improvement occurs, it is argued, because retrieval modifies the memorial representation of the retrieved information in such a way as to make it more recallable in the future than it would have been otherwise (see, e.g., R. A. Bjork, 1975; Carrier & Pashler, 1992). Consequently, cued-recall, short-answer, and free-recall tests that require retrieval are highly regarded as retention-promoting test formats, whereas multiple-choice testing—presumed to require relatively little explicit retrieval—is not.

Indeed, cued-recall testing has been shown to improve the retention of tested—and sometimes even of nontested related—information (e.g., R. C. Anderson & Biddle, 1975; Boker, 1974; Chan, McDermott, & Roediger, 2006; Duchastel, 1981; Frase, 1967, 1968, 1971; McGaw & Groteluesche n, 1972; Rickards, 1976; Rothkopf, 1966; Rothkopf & Bisbicos, 1967; Rothkopf & Bloom, 1970; see Roediger & Karpicke, 2006, for an excellent review of testing effects), whereas the demonstrated benefits of multiple-choice testing have often been smaller (see, e.g., the meta-analyses by R. C. Anderson & Biddle, 1975, and Hamaker, 1986). Furthermore, in studies directly comparing the effectiveness of cued-recall versus multiple-choice testing, cued-recall questions have usually led to better retention of the tested information than have multiple-choice questions, with this result being largely attributed to differences in the types of

J. L. Little · E. L. Bjork
Department of Psychology, University of California, Los Angeles, Los Angeles, CA, USA

J. L. Little (✉)
Department of Psychology, Hillsdale College, 33 E. College St., Hillsdale, MI 49242, USA
e-mail: jerilittle@gmail.com

Published online: 15 August 2014

processing required by these two test formats (Foos & Fisher, 1988). Specifically, many researchers have argued that—whereas cued-recall tests involve retrieval processes known to enhance later recall—multiple-choice tests primarily rely upon recognition processes, which have been shown to lead to lower subsequent retention than do tasks involving substantial retrieval (e.g., Carpenter & DeLosh, 2006; Glover, 1989).

The argument that multiple-choice tests rely primarily upon recognition processes seems, on the surface at least, to be a reasonable critique of multiple-choice testing. Multiple-choice questions do, in fact, expose the correct answer to the learner by presenting it as one of the alternatives, which could obviate the need for retrieval. Not all multiple-choice questions, however, can be answered through recognition processes alone. Consider, for example, the question *What is Saturn's largest moon?* with the choices *Titan*, *Rhea*, *Mimas*, and *Enceladus*, which might be asked on a multiple-choice test following students' study of a passage about Saturn containing information about *Titan*, *Rhea*, *Mimas*, and *Enceladus*, four of Saturn's moons. The test-taker should recognize all of these alternatives as moons from having encountered them during study of the passage, and thus, should view all of them as plausible answer choices. Hence, even though the correct answer is presented as one of the alternatives, additional information might need to be retrieved to discriminate among the alternatives in the attempt to determine or select the most likely correct answer. Indeed, recent work has shown that, relative to cued-recall tests, multiple-choice tests using these types of plausible or competitive alternatives can improve later performance—not only on questions about the information that was previously explicitly tested, but also on questions about information pertaining to the incorrect alternatives (Little, Bjork, Bjork, & Angello, 2012, Exp. 1). To illustrate, Little et al. had participants study two passages and then take either an initial multiple-choice test or an initial cued-recall test without feedback. On the later cued-recall test, participants were better able to answer both previously tested and nontested related questions when they had taken an initial multiple-choice test than when they had taken an initial cued-recall test on the same information.

The reason for these observed benefits—particularly for the enhanced performance on nontested related questions—when given a prior multiple-choice test versus a cued-recall test has not yet been determined. One possibility is that exposure to the incorrect choices in a prior multiple-choice question simply makes those items more accessible. This increased accessibility allows test-takers to recall these incorrect choices more easily in the future, making them more likely to be output as potential answers to later cued-recall questions for which they may sometimes be the correct answer. A more interesting possibility, in our view, is that multiple-choice tests with competitive alternatives actually induce the retrieval of information pertaining to the incorrect alternatives. When, for example, the incorrect alternatives on a multiple-choice question are plausible answers to that question, information presented in the passage specifically pertaining to such competitive alternatives may need to be retrieved in order for the test-taker to select among them. Such processing would likely strengthen the association between such retrieved information and the alternatives. Then, should such information become the basis for a cued-recall question on the final test, the test-taker's ability to answer such a related question should be enhanced. To the extent, however, that incorrect alternatives can be rejected without bringing to mind specific information pertaining to them, as would be the case with noncompetitive alternatives, these processes would be unlikely to occur. Following from this *retrieval hypothesis*, answers to related questions would be more likely to be correctly recalled on a later cued-recall test if they had served as competitive alternatives than if they had served as noncompetitive alternatives on an earlier multiple-choice test.

It is also possible, however, that competitive alternatives could simply be processed more deeply, in a general sense, than noncompetitive alternatives would be. Perhaps test-takers would simply spend more time examining competitive than noncompetitive alternatives, but without retrieving specific information pertaining to them, or at least not information that would be specific enough to improve performance in answering related questions. If so, the test-taker might be more likely to provide such competitive alternatives as answers to related questions, even though the connection between those incorrect alternatives and the information pertaining to them had not been strengthened during the multiple-choice test, and sometimes such responses would be correct. Relevant to this alternative explanation is a result observed by Jacoby, Shimizu, Daniels, and Rhodes (2005), who found (using a *memory-for-foils* paradigm) that memory for information that one needs to select against (e.g., *distractor words* or *lures* in a recognition memory procedure) can be strengthened during an initial recognition task, and that the amount of such strengthening (as evidenced by "yes" and "no" responses on a later recognition test) depends on the depth of processing during study and/or the initial recognition test. Although there are important differences between the materials and procedures employed in the present research and those used by Jacoby et al., their work has helped to motivate the notion that competitive alternatives might simply be recalled better because they are processed more deeply than noncompetitive ones, not because the information pertaining to them is strengthened.

Our primary goal in the present research was to assess whether previously incorrect competitive alternatives would be recalled better than previously incorrect noncompetitive alternatives as correct answers to related questions, and if so, whether this finding could best be accommodated by the retrieval account put forth by Little et al. (2012) or by a general deep-processing account.

A secondary question of interest to us was what would be the effect of the competitiveness of the alternatives on the retention of the previously tested information. One possibility is that the presence of competitive alternatives, as opposed to noncompetitive alternatives, would lead to better retention of the previously tested information. This hypothesis follows from previous work by Whitten and Leonard (1980), who investigated the later recall of target words that had been given—between being studied and tested for recall—one of two types of intervening recognition tests: choosing the target word either from among semantically related distractor words (i.e., a relatively difficult recognition test) or from among semantically nonrelated distractor words (i.e., a relatively easy recognition test). They found that the previously studied target words that had been given the difficult recognition test versus the easy recognition test were then recalled better on a final recall test (although the initial correct recognition performance had been nearly identical in the two conditions). Another possibility, however, is that performance on questions previously tested with competitive alternatives would not be better than performance on questions previously tested with noncompetitive alternatives. A straightforward reason for this second possibility would be that questions containing competitive alternatives would be harder to answer correctly than questions containing noncompetitive alternatives, and if a participant cannot answer a question on an initial multiple-choice test (that is not followed by feedback), it is unlikely that he or she would be able to answer that question later when it was asked as a cued-recall question.

Experiment 1

In Experiment 1, we tested the adequacy of the retrieval-based hypothesis put forth by Little et al. (2012), suggesting that multiple-choice tests with competitive alternatives would induce the recall of information pertaining to those incorrect alternatives to a larger extent than would multiple-choice tests with noncompetitive alternatives. Although such a retrieval-based explanation seems compelling and is consistent with the findings obtained by Little et al. (2012), a general deep-processing account is also consistent with their findings.

Our approach to testing between these two possible explanations was as follows. First, we manipulated the level of competitiveness of the alternatives in the prior multiple-choice questions. Our reasoning for doing so was that the more plausible the incorrect alternatives were as answers, the more competitive they would be, and thus, the more processing would be required in trying to select among them in answering the question. If some of this processing involved the retrieval of information from the passage pertaining to the alternatives—as is assumed to occur in the retrieval-based explanation—then the association between the alternatives and the specific information so retrieved should be strengthened. Consequently, the likelihood that the test-taker would be able to answer later cued-recall questions on the basis of that information would be increased. In contrast, such retrieval processes would not be invoked, or at least would be to a lesser extent, for multiple-choice questions with noncompetitive alternatives. Second, by having two types of incorrect alternatives for each question, our design allowed us to test the adequacy of the general deep-processing account. To illustrate, we used multiple-choice questions that had three alternatives: one correct alternative and two incorrect alternatives. Of the two incorrect alternatives, one was incorrect for the multiple-choice question but correct for the related cued-recall question, and one was incorrect for both questions. If incorrect alternatives are simply more generally strengthened when they are competitive than when they are not (as opposed to having their association to specific information from the passage strengthened via retrieval processes), we should find that the second type of incorrect alternative (i.e., the ones that were incorrect for both questions) would be intruded as incorrect responses to the related questions more often when they were used as competitive incorrect alternatives than when they were used as noncompetitive incorrect alternatives.

To summarize, for our results to be consistent with the retrieval-based explanation, we would expect to find that prior multiple-choice questions using more plausible incorrect alternatives would lead to better performance on related cued-recall questions than would prior multiple-choice questions using less plausible incorrect alternatives. Furthermore, we would not expect to see a difference in the intrusion rates for related questions as a consequence of whether the alternatives were competitive or noncompetitive, whereas if the general deep-processing explanation is correct, we should see such a difference for intrusion rates.

Method

Participants and design

A total of 28 undergraduate students participated for partial credit in psychology courses being taught at the University of California, Los Angeles. To overview the design, participants read two text passages, with one followed immediately by a multiple-choice test (with two types of multiple-choice questions: some containing competitive incorrect alternatives and others containing noncompetitive incorrect alternatives) and one not followed by an immediate test (thus serving as a nontested control passage). Following a filled retention interval, a final cued-recall test was administered on which the participants answered previously tested and nontested related questions from the tested passage and nontested control questions from the nontested control passage. Thus, we employed a 2 (item type: previously tested vs. previously nontested

related) × 2 (question type: competitive vs. noncompetitive) within-subjects design for the testing condition plus a control condition, with all participants serving in both conditions.

Materials

Two passages (of approximately 1,050 words each) were constructed: one about the solar system and one about ferrets, with eight pairs of basic questions associated with each passage.

Examples of four such basic pairs of questions are shown in Table 1. Note that for these questions, although the correct answers to the questions in a pair are different, they are of the same category (e.g., both proper names, terms, or numbers), which was true for all question pairs.

Multiple-choice items for the initial test To convert our pairs of basic questions into a multiple-choice format, four incorrect alternatives were chosen for each pair of questions, on the basis of other information presented in the passage. Of these four incorrect alternatives, two were highly related to one of the questions in the pair (and, thus, could serve as plausible, but incorrect, answers for it), and the other two alternatives were highly related to the other question in the pair (and, thus, could serve as plausible, but incorrect, answers for it). Consider, for example, the first question pair shown in Table 1. Two of the alternatives (*Uranus* and *Saturn*) are outer/gaseous planets and, thus, plausible alternatives for the question about an outer planet, but implausible alternatives for the question about an inner/terrestrial planet. Likewise, the other two alternatives (*Mercury* and *Mars*) are plausible alternatives for the question about a terrestrial planet, but implausible alternatives for the question pertaining to an outer/gaseous planet. (Note, however, that for many of the pairs, the relative competitiveness of the alternatives is less clear without reading the passages.) By manipulating which set of alternatives (competitive or noncompetitive) was presented as answer choices, we were able to create competitive and noncompetitive versions of each basic question. To ensure that each alternative was presented only once for a given participant, each participant answered either the two competitive questions or the two noncompetitive questions for a given pair.

In summary, the six possible alternatives for each of the eight pairs of basic questions (two correct and four incorrect choices) were manipulated so as to allow each basic question to be asked in the form of a three-alternative multiple-choice question that was presented with either competitive incorrect alternatives or noncompetitive incorrect alternatives for a given participant. On the initial test, each participant received eight competitive and eight noncompetitive multiple-choice questions, for a total of 16 questions, with the competitive and noncompetitive multiple-choice versions of each basic question being counterbalanced across participants.

Items for the final cued-recall test Three types of items appeared on the 64-item final cued-recall test for a given participant: (a) previously tested, (b) related, and (c) control items. The previously tested items were the questions that had appeared on the initial multiple-choice test. All 16 of these (eight that had appeared as competitive and eight that had appeared as noncompetitive multiple-choice questions on the initial test) were presented again on the final test, except now in the form of cued-recall questions.

The related items were questions whose answers had previously been incorrect alternatives from the initial multiple-choice test. Two of these were constructed for each of the eight pairs of basic questions for each passage, creating a total of 16 related questions for each passage. Table 2 shows the two related questions that corresponded to each of the four

Table 1 Examples of basic question pairs and the corresponding competitive and noncompetitive alternatives

Basic Question	Correct Answer	Alternatives	
		Competitive	Noncompetitive
Which outer planet was discovered by mathematical prediction rather than by direct observation?	Neptune	Uranus; Saturn	Mercury; Mars
What is the hottest terrestrial planet?	Venus	Mercury; Mars	Uranus; Saturn
Which island(s) has (have) the largest population of feral ferret–polecat hybrids, used to control the rabbit population?	New Zealand	Balearic Islands; Shetland Islands	United Kingdom; Rome
Where are ferrets said to have originated?	Egypt	United Kingdom; Rome	Balearic Islands; Shetland Islands
How many objects in our solar system have natural satellites (moons)?	9	6; 8	30–55; 2,000–50,000
How many AU is the scattered disc away from the Sun?	60–100	30–55; 2,000–50,000	6; 8
When a ferret has black-tipped hairs, it is said to have what coloring?	sable	roans; blaze	gib; sprite
What is the term for an intact male ferret?	hob	gib; sprite	roans; blaze

Table 2 Related questions that correspond to the basic questions presented in Table 1

Related Question	Correct Answer/ Previously Incorrect Alternative
Which planet's axial tilt is 90 degrees to the plane of its orbit (meaning it revolves around the sun on its side)?	Uranus
Which planet was first visited by the Mariner 10?	Mercury
What location has the biggest population of ferrets (not ferret–polecat hybrids)?	Shetland Islands
Where are ferrets still used for hunting today?	United Kingdom
How many AU away from the sun is the Kuiper belt?	30–55
How many planets in our solar system have natural satellites (moons)?	6
A ferret with _____ coloration has alternating white and pigmented hairs.	roans
What is the term for a spayed female ferret?	sprite

The correct answers for these questions were previously incorrect alternatives. For a given participant, the correct answer to a related question had either been a competitive alternative or a noncompetitive alternative in the previous corresponding multiple-choice question

question pairs in Table 1. The related questions for a given passage were never presented as items on the initial multiple-choice tests; instead, they only appeared on the final cued-recall test. In addition, because the same incorrect alternative could appear in a competitive or a noncompetitive version of a multiple-choice question for different participants, there were two types of related questions on the final test: ones for which the correct answer had previously been a competitive incorrect alternative, and ones for which the correct answer had previously been a noncompetitive incorrect alternative.

Finally, the control items presented on the final cued-recall test for a given participant were all the questions about the passage that had served as the control passage for that particular participant. These consisted of the 16 items in the eight pairs of basic questions constructed for that passage, plus the 16 questions that would be the related items for that passage when it was the one given an initial multiple-choice test rather than serving as the control passage. Thus, for a given participant, there were always 32 control items on the final test.

Procedure

The experiment began with all participants being given the first of two passages (Solar System or Ferrets) to read. They were told they had 10 min to read it and to continue studying the passage if they finished reading it before the allotted time was up. Next, participants were given an immediate 16-item multiple-choice test on the passage, which took 3 min to complete, or they engaged in a nonverbal filler task (i.e., playing Tetris) for the same amount of time required to take the multiple-choice test. Questions on the multiple-choice test appeared on a computer screen, one at a time for 10 s each, with the screen going blank for 2 s between successive questions, and participants gave their answers to the questions out loud. The intervening blank screen was employed to ensure that participants would have enough time to give a vocal response before the next question was presented. Corrective feedback was never given. After the 3-min interval during which participants either took the multiple-choice test or played Tetris, the participants were then presented with their second passage for study. If a participant had received a multiple-choice test after the first passage, then that participant engaged in the nonverbal filler task after study of the second passage, and vice versa for the other participants.

Following completion of the above phase of the experiment, a 4-min retention interval was imposed for all participants, during which time they played Tetris. Next, each participant received a final, 64-item cued-recall test containing previously tested, related, and nontested control items. Each item or question was presented by itself for 12 s, with the screen going blank for 2 s between successive questions, and participants gave their answers to the questions out loud. In the final test, all test items were presented as cued-recall questions, and thus were presented without alternatives. The questions appeared in the following order. First, all 32 of the items for the control passage were presented in the first half of the test (with the 16 items corresponding to what would be its related questions when it was the tested passage occurring first, followed by the 16 items corresponding to what would be its basic questions when it was the tested passage). Then, the 16 related items were presented at the beginning of the second half of the test (with those whose correct answer had appeared as a competitive vs. a noncompetitive alternative on the initial test appearing in alternation), and, lastly, the 16 previously tested items were presented (with those corresponding to previous competitive and noncompetitive multiple-choice questions on the initial test appearing in alternation).

We presented the related items on the final test before the previously tested items because we were most interested in how the competitiveness of previously presented incorrect alternatives might affect performance on related questions. Then, in analyzing participants' performance on the final test, we compared performance for related items with that for the control items appearing first on the final test (i.e., in the first fourth of the list), and we compared performance for previously tested questions with that for the control items appearing second on the final test (i.e., in the second fourth of the list) to control for output interference effects.

The order in which the two passages (Solar System and Ferrets) were presented, which one served as the tested versus

the control passage, and whether the first or second passage was the initially tested passage were all counterbalanced across participants.

Results and discussion

Initial multiple-choice test performance

Significantly more noncompetitive questions (M = 86 %, SE = 3 %) than competitive questions (M = 66 %, SE = 3 %) were answered correctly by participants on the initial multiple-choice test, $t(27) = 5.67$, $p < .001$, $d = 1.08$, consistent with the assumption that the presence of competitive alternatives versus noncompetitive alternatives on a multiple-choice question would make such questions more difficult to answer.

Final-test performance

Performance on the final cued-recall test is illustrated in Fig. 1, and as is indicated there, participants seemed able to answer both types of previously tested items better than the corresponding control items. Their performance on related items, however, appears to have been improved (as compared to that for the corresponding control questions) only when the related question pertained to what had been a competitive incorrect alternative on the initial multiple-choice test.

The items corresponding to previously tested multiple-choice questions and the related items appearing for the first time on the final cued-recall test were different sets of questions. Thus, in our analysis of final-test performance for these two types of items, the questions asked about the control passage that appeared in the first quarter of the final test (which would be the related items for that passage when it was the tested passage) served as the baseline control items for related items about the tested passage; those that appeared in the second quarter of the final test (which would be the basic questions for that passage when it was the tested passage) served as the baseline control items for previously tested items. We evaluated performance on previously tested and related items versus their appropriate control items with planned paired-samples t tests.

Related items Participants' performance on questions whose answers had previously appeared as competitive incorrect alternatives on the initial multiple-choice test (M = 47 %, SE = 5 %) was significantly better than their performance on the comparable control items (M = 36 %, SE = 4 %), $t(27) = 2.21$, $p = .04$, $d = 0.42$. In contrast, participants' performance on questions whose answers had previously appeared as noncompetitive alternatives on the initial multiple-choice test was not (M = 36 %, SE = 4 %), $t(27) = 0.1$, $p = .96$. Additionally, planned paired-samples t tests revealed that performance on related questions whose answers had previously been competitive incorrect alternatives on the initial test (shown by the leftmost bar in the right group of bars in Fig. 1) was better than that on questions whose answers had been noncompetitive incorrect alternatives (shown by the middle bar in the right group of bars in Fig. 1), $t(27) = 2.55$, $p = .02$, $d = 0.49$.

Thus, as previously discussed, we observed a pattern of results that is consistent with the notion that multiple-choice tests with competitive alternatives invoke search for and retrieval of specific information from the studied passage

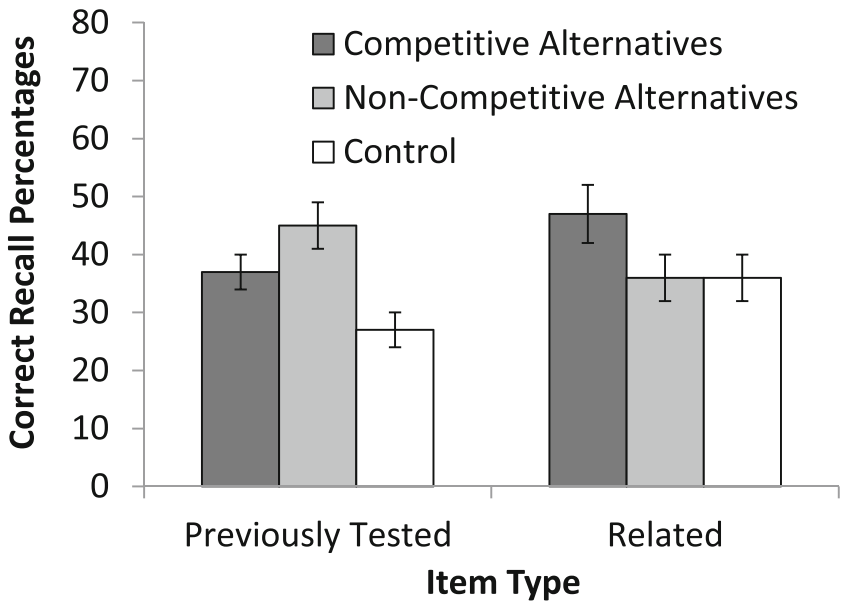

Fig. 1 Correct performance percentages on the final cued-recall test as a function of item type (previously tested or nontested related) and competitiveness of the incorrect alternatives in Experiment 1. The left and right white bars represent control items tested last and control items tested first, respectively. Error bars represent ±1 *SEM*

pertaining to those alternatives. As we also previously discussed, however, this pattern is also consistent with a general deep-processing account. Thus, to evaluate the adequacy of the general deep-processing account, we looked at the extents to which participants intruded incorrect alternatives from the initial multiple-choice test as incorrect answers to the related questions. To review, each multiple-choice question had three alternatives, two of which were incorrect. Of these two incorrect alternatives, one was correct for the related question, and one was never a correct answer. If incorrect alternatives were simply more generally strengthened when they were competitive than when they were not, we should find that never-correct alternatives were intruded as incorrect responses to the related questions more often when they had been used as competitive rather than noncompetitive incorrect alternatives. Instead, however, we found no difference in the intrusion rates of such items when they had served as competitive incorrect alternatives (M = 4 %, SE = 1 %) versus when they had served as noncompetitive alternatives (M = 5 %, SE = 1 %), $t(27) = 0.90$, $p = .38$. Additionally, on the initial multiple-choice test, when questions had competitive alternatives, participants were no more likely to choose the alternative that would later turn out to be the correct answer to the related question (M = 17 %, SE = 2 %) than they were to choose the incorrect alternative that was never the correct answer (M = 16 %, SE = 2 %), $t(27) = 0.26$, $p = .80$, suggesting that these two types of alternatives had similar competitive strength.

Thus, it would seem that when competitive alternatives are used in a multiple-choice question, they are not simply processed more deeply, in a manner that makes those alternatives themselves more accessible and thus more likely to be produced in response to any question for which they would be a plausible—albeit incorrect—response. Rather, they appear to be processed in a way that leads them to be more recallable as correct responses to specific cues (i.e., specific information from the studied passage). We thus see the pattern of results revealed by this intrusion analysis, in combination with the pattern of results for correct performance, as being consistent with the notion that when answering competitive multiple-choice questions, test-takers think about or retrieve information pertaining to them from the previously studied passage.

Previously tested items Participants' performance on questions that had been previously tested with either competitive alternatives (M = 37 %, SE = 3 %) or noncompetitive alternatives (M = 45 %, SE = 4 %) was enhanced relative to their performance on the appropriate control items (M = 27 %, SE = 3 %), $t(27) = 3.10$, $p < .01$, $d = 0.59$, and $t(27) = 4.54$, $p < .001$, $d = 0.87$, respectively. Thus, a testing effect was observed for both types of items. Performance on the cued-recall questions that had previously appeared as multiple-choice questions with competitive alternatives, however, was marginally less than performance on those that had previously appeared as multiple-choice questions with noncompetitive alternatives (the leftmost and middle bars, respectively, in the left group of bars in Fig. 1), $t(27) = 1.76$, $p = .09$, $d = 0.34$.

Because performance on the initial test was marginally worse for competitive multiple-choice questions than for noncompetitive multiple-choice questions, we conducted an analysis for performance on previously tested items, conditional upon having correctly answered that question on the initial multiple-choice test. Looking at only those questions that had been correctly answered on the initial multiple-choice test, performance on the later cued-recall test was numerically greater for items that had originally been presented as multiple-choice items with competitive alternatives (M = 55 %, SE = 5 %) than for those originally presented as multiple-choice items with noncompetitive alternatives (M = 50 %, SE = 4 %), $t(27) = 0.89$, $p = .38$. Although this difference was not reliable and might be the result of item-selection effects, it is worth noting that it is in the opposite direction from the one in the unconditional analysis.

Experiment 2

In Experiment 2, we essentially replicated the design of Experiment 1, but with the inclusion of a feedback manipulation. Examining retention as a consequence of providing feedback on previous multiple-choice tests is important for both practical and theoretical reasons. In most educational contexts, feedback is provided. Thus, it becomes important to know whether the benefits observed in Experiment 1, particularly for competitive related information, would still emerge if feedback were provided. On the basis of the results of Little et al. (2012), it seems unlikely that feedback would disrupt this result; nevertheless, an explicit test of that assumption was definitely warranted.

With respect to the theoretical notions under consideration in the present research, examining the effects of feedback could also provide important relevant findings. In Experiment 1, we found that, although final-test performance for questions previously tested with competitive alternatives was not better than that for questions previously tested with noncompetitive alternatives, performance conditionalized on correct multiple-choice performance was numerically higher. Although this change in the pattern of results may have been the consequence of item-selection effects, it is possible that the conditionalized pattern points toward the idea that competitive alternatives would be better for the retention of previously tested information, but only when a learner actually has access to the correct answer following the initial multiple-choice test. A straightforward test of this idea would be to provide feedback following the initial multiple-choice test, so that learners

would have access to all of the answers following the test, regardless of how difficult particular questions were to answer. With feedback, one might expect competitive alternatives to be recalled better than noncompetitive ones. On the other hand, it is possible that provision of feedback would not lead to increased performance for questions previously tested with competitive alternatives versus those previously tested with noncompetitive alternatives. To the extent that competitive incorrect alternatives induce learners to process *them* and recall information pertaining to *them* (an assumption supported by Exp. 1), less attention might be allocated to forming a relationship between the present question and its correct answer, whether feedback is provided or not.

In sum, the main goal of Experiment 2 was to replicate and extend the findings of Experiment 1. We hoped to obtain additional evidence for the retrieval hypothesis and also to examine whether feedback would alter final test performance for previously tested information.

Method

Participants and design A total of 96 members of the University of California, Los Angeles, and Washington University in St. Louis communities participated for course credit or payment. The design was similar to that employed in Experiment 1, but with the addition of a feedback manipulation. Specifically, half of the participants were randomly assigned to receive feedback during the initial multiple-choice test. Thus, we employed a 2 (item type: previously correct alternative vs. previously incorrect alternative) × 2 (alternative type: competitive vs. noncompetitive) × 2 (feedback: present vs. absent) mixed-subjects design for the testing condition plus a control condition, with all participants serving in the testing and control conditions.

Materials and procedure The materials were the same as those used in Experiment 1. The procedure was also the same as that used in Experiment 1, with the following exceptions. On both the initial and final tests, participants typed their responses. Half of the participants were randomly assigned to receive corrective feedback on the initial multiple-choice test. Those who were assigned corrective feedback had 15 s to answer each question and received corrective feedback (e.g., "The correct answer is Mercury") for 3 s immediately after answering each question. Those who were not assigned corrective feedback had 18 s to answer each question.

The ordering of questions on the final test was slightly different from that in Experiment 1. Questions from the control passage were tested immediately prior to the questions from the tested passage to which they would be compared. Specifically, 16 control items and the 16 items that were related to the previously tested questions (with those whose correct answer had appeared as a competitive vs. a noncompetitive alternative on the initial test appearing in alternation) were presented in the first and second quarters of the test, respectively. Then, the other 16 control items and the 16 previously tested items to which they would be compared (with those corresponding to previous competitive vs. noncompetitive multiple-choice questions on the initial test appearing in alternation) were presented in the third and fourth quarters of the test, respectively.

Results

Initial multiple-choice test performance Significantly more noncompetitive questions (M = 83 %, SE = 2 %) than competitive questions (M = 66 %, SE = 2 %) were answered correctly by participants on the initial multiple-choice test, $t(95) = 8.62$, $p < .001$, $d = 0.88$, consistent with the results of Experiment 1. We found no interaction with feedback condition, $F < 1$.

Final-test performance for related items Correct recall percentages for related items on the final cued-recall test are shown in Fig. 2. As is indicated there, performance appears only to have been improved (as compared to that for the corresponding control questions) when the related question pertained to what had been a competitive incorrect alternative on the initial multiple-choice test. Additionally, this pattern of results appears to be similar in both the feedback and no-feedback conditions.

Participants' performance on questions whose answers had previously appeared as competitive incorrect alternatives on the initial multiple-choice test (M = 46 %, SE = 2 %) was significantly better than their performance on the comparable control items (M = 37 %, SE = 2 %), $t(95) = 4.36$, $p < .001$, $d = 0.43$. In contrast, participants' performance on questions whose answers had previously appeared as noncompetitive alternatives on the initial multiple-choice test was not (M = 37 %, SE = 2 %), $t(94) = 0.15$, $p = .88$. Additionally, planned paired-samples t tests revealed that performance was better on related questions whose answers had previously been competitive incorrect alternatives on the initial test than on questions whose answers had been noncompetitive incorrect alternatives, $t(95) = 4.36$, $p < .001$, $d = 0.45$. Thus, the basic pattern of results observed in Experiment 1 on the final recall test for related items was replicated in Experiment 2. Additionally, the provision of feedback did not interact with item type for related items, $F < 1$, consistent with the earlier findings of Little et al. (2012).

In an analysis of intrusions on the final recall test, we found no difference in the intrusion rates of previously incorrect alternatives that were never correct as responses to related questions, whether they had served as competitive incorrect alternatives (M = 7 %, SE = 1 %) or noncompetitive alternatives (M = 7 %, SE = 1 %), $t(95) = 0.12$, $p = .91$. Additionally,

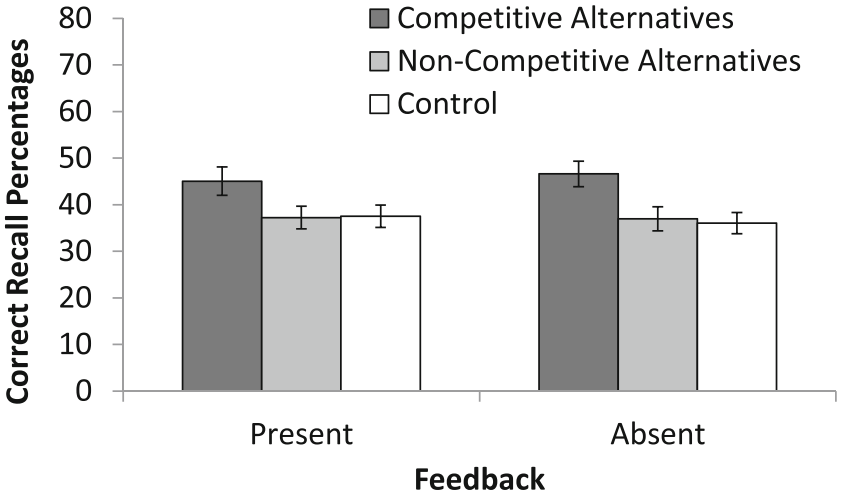

Fig. 2 Correct performance percentages on the final cued-recall test for previously nontested related questions as a function of competitiveness of the incorrect alternatives on the initial multiple-choice test and whether feedback was present or absent on the initial test in Experiment 2. Error bars represent ±1 *SEM*

as in Experiment 1, on the initial multiple-choice test, when questions had competitive alternatives, participants were no more likely to choose the alternative that would later turn out to be the correct answer to the related question (M = 17 %, SE = 1 %) than they were to choose the incorrect alternative that was never the correct answer (M = 15 %, SE = 1 %), $t(95)$ = 1.41, p = .16, suggesting that these two incorrect alternatives had similar competitive strengths. Taken together, these findings provide additional support for the retrieval hypothesis and do not support the general deep-processing hypothesis.

Final-test performance for previously tested items Correct recall percentages for previously tested items on the final cued-recall test are shown in Fig. 3 and, as is indicated there, performance appears to have been improved (as compared to that for the corresponding control questions) both when competitive incorrect alternatives and noncompetitive incorrect alternatives were used on the initial multiple-choice test. Additionally, this pattern of results appears to be similar in both the feedback and no-feedback conditions.

Participants' performance on questions that had been previously tested with either competitive alternatives (M = 52 %, SE = 2 %) or noncompetitive alternatives (M = 52 %, SE = 2 %) was enhanced relative to their performance on the appropriate control items (M = 29 %, SE = 1 %), $t(95)$ = 11.46, $p < .001$, d = 1.25, and $t(95) = 9.58$, $p < .001$, d = 1.04, respectively. The interaction between feedback and item type (previously tested with competitive vs. noncompetitive alternatives) was not reliable, $F < 1$, but the presence of feedback improved performance, overall, for both types of items, $F(1,$

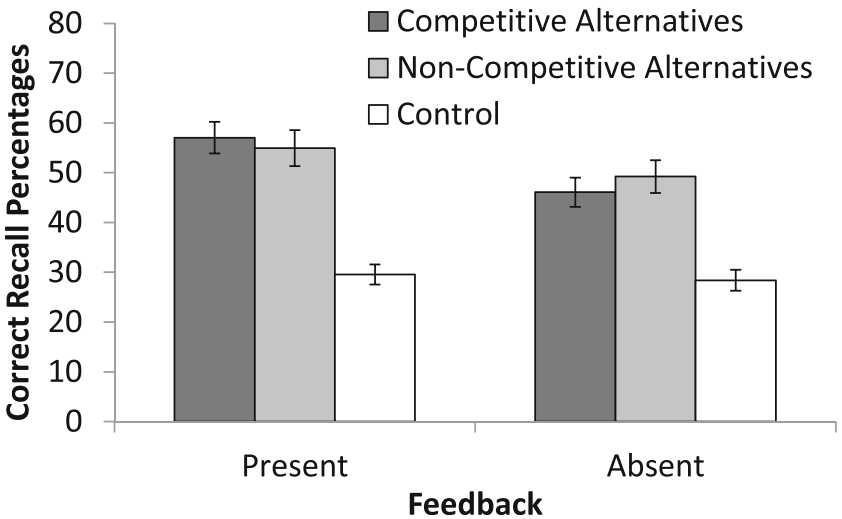

Fig. 3 Correct performance percentages on the final cued-recall test for previously tested questions as a function of competitiveness of the incorrect alternatives on the initial multiple-choice test and whether feedback was present or absent on the initial test in Experiment 2. Error bars represent ±1 *SEM*

94) = 4.93, MSE = 4.33, η_p^2 = 0.05, p = .03. This improvement in performance was only reliable for competitive items, however, $t(94) = 2.53$, $p = .01$. Looking at only those questions that had been correctly answered on the initial multiple-choice test, performance on the later cued-recall test was reliably greater for those that were originally presented as multiple-choice items with competitive alternatives (M = 65 %, SE = 2 %) than for those originally presented as multiple-choice items with noncompetitive alternatives (M = 57 %, SE = 2 %), $t(95) = 2.89$, $p < .01$.

Feedback did not interact with the competitiveness of the alternatives. One expectation pertaining to the influence of feedback on previously tested information was that feedback might be more beneficial for questions with competitive alternatives than for those with noncompetitive alternatives. We did not find clear evidence for this assumption in the present experiment, however. The data are most supportive of the notion that feedback helped both types of items to about the same extent. Although we did find conditional performance to be better for questions previously tested with competitive alternatives than for those previously tested with noncompetitive alternatives, we cannot rule out specific item effects as the reason.

As we suggested earlier, one reason for why competitive alternatives might not have improved retention of the previously tested information, as compared to noncompetitive alternatives, in the present situation makes sense in light of the present evidence for the retrieval hypothesis: To the extent that competitive alternatives induce learners to process them and recall information pertaining to them, less attention might be allocated to forming a relationship between the question and its correct answer. Although Whitten and Leonard (1980) found that competitive distractors were remembered better, it is possible that their distractors did not induce the level or type of processing that would take attention away from the target.

General discussion

In the present experiments, we assessed two possible theoretical explanations (referred to earlier as the *retrieval hypothesis* vs. *general deep processing*) as to why the taking of an initial multiple-choice test with competitive incorrect alternatives can lead to enhanced performance in answering questions based on related information pertaining to the incorrect alternatives. We did so by manipulating the level of competitiveness of the alternatives in the prior multiple-choice questions and demonstrating that competitive alternatives are necessary for this benefit to occur. Specifically, in both experiments, we found that when an incorrect alternative had been a competitive choice in a previous multiple-choice question, then a question for which it was the correct answer on a delayed cued-recall test was more likely to be answered correctly than a corresponding control question, whereas such enhancement did not occur for answering a question about the same alternative when it had appeared as a noncompetitive choice in a previous multiple-choice question. Additionally, we found that other alternatives (which were never correct) were not more likely to be intruded as incorrect responses when they had served as competitive alternatives than when they had served as noncompetitive alternatives. Furthermore, these findings were replicated when feedback was provided during the initial multiple-choice tests. Across two experiments, then, we observed results consistent with the proposed retrieval explanation and inconsistent with a general deep-processing explanation as to why the taking of competitive multiple-choice tests can lead to enhanced performance on related questions.

The role of noncompetitive alternatives

The results of the present research point toward the notion that when test-takers answer multiple-choice questions with competitive alternatives, they are led to think about or retrieve specific information pertaining to those alternatives in the process of selecting their answer, but when they answer multiple-choice questions with noncompetitive alternatives, they are less likely to think about or retrieve specific information pertaining to those alternatives in the process of selecting their answer. This is not to say, however, that they do not think about any information pertaining to noncompetitive incorrect alternatives.

In the present materials, noncompetitive alternatives were not so noncompetitive as to be outlandish choices—in fact, participants chose a noncompetitive alternative 14 %–17 % of the time on the initial test. Why, then, was facilitation not seen for questions related to the earlier tested noncompetitive multiple-choice questions? We suggest the following possibility. Although participants may have needed to recall some information to reject these noncompetitive alternatives, this information was likely not specific enough to have been useful in the answering of related questions on the cued-recall test, which were based upon specific information pertaining to the previous incorrect alternatives. Consider, for example, the first question pair shown in Table 1. All of the incorrect alternatives are planets, but these planets can be generally classified into inner/terrestrial planets (*Mars*, *Mercury*, *Venus*) and outer/gaseous planets (*Uranus*, *Neptune*, *Saturn*). Thus, when answering the question "Which outer planet was discovered by mathematical prediction rather than by direct observation," given *Mars* and *Mercury* as the incorrect alternatives, rejecting both of them could be done on the basis of the general information that they are not outer planets, without the need for recalling specific information about each of them from the passage. As a consequence, although some

information might be brought to mind from the passage about Mars and Mercury in attempting to answer this question, that information would likely be more general than would be needed to answer the type of related questions on the final test. In contrast, when *Uranus* and *Saturn* were the incorrect alternatives, such general information about them (i.e., that they are both outer planets) would not discriminate them from *Neptune*, making it more likely that the test-taker might try to retrieve some more specific information about them from the passage, with such information possibly forming the basis for a related question appearing on the final test.

Relation to previous work

The present research contributes to and expands upon earlier findings regarding the effects of testing as it pertains to different types of relationships between the tested and nontested information. Some past work, for example, has shown that recall of similar or related nontested information can be enhanced if, for example, it has appeared in close temporal proximity to the tested information in a studied passage (e.g., Frase, 1971; and see also Chan et al., 2006, who defined relatedness in a similar manner). In such cases, the proposed mechanism is that when trying to recall information from the passage to answer a specific question, one might spontaneously also recall information that had been presented in close proximity to it. In one study demonstrating such benefits, for example, McGaw and Grotelueschen (1972) drew the tested and nontested related information from the same sentence. In another study, Watts and Anderson (1971) used initial test questions that had learners define concepts or statements of general principles, and then the researchers examined performance for related questions that would require learners to identify examples of these concepts. That these types of relationships between tested and nontested information might lead to enhanced recall for the initially nontested information on a later test seems reasonable, but the occurrence of such facilitation seems less reasonable for other types of relationships—specifically, relationships between tested and nontested pieces of information that are competitive.

Some evidence outside of the standard testing-and-adjunct-question literature suggests that recall of nontested competitive information can be impaired as a consequence of testing related information (e.g., M. C. Anderson, Bjork, & Bjork, 1994). The explanation for such impairment remains controversial, with some researchers suggesting that in trying to recall the answer to the tested question, competitive information or alternatives have to be selected against, leading to their suppression or diminished accessibility on a later test (e.g., Anderson et al., 1994; M. C. Anderson & Spellman, 1995; Johansson, Aslan, Bäuml, Gäbel, & Mecklinger, 2007). Others have rejected this selection-plus-suppression account in favor of more general interference accounts (e.g., Camp, Pecher, & Schmidt, 2007; Dodd, Castel, & Roberts, 2006; MacLeod, Dodd, Sheard, Wilson, & Bibi, 2003; Perfect et al., 2004; Williams & Zacks, 2001)—the idea being that the information that is strengthened as a consequence of testing would then interfere with the recall of competitively related nontested information on a later test. Regardless of why the testing of some information leads to impaired retention of competitive nontested information, such impairment has been documented many times, even with educational materials—both when the initial test utilizes cued-recall questions (Carroll, Campbell-Ratcliffe, Murnane, & Perfect, 2007; Chan, 2009; Little et al., 2012, Exp. 1; Macrae & MacLeod, 1999) and free recall (Little, Storm, & Bjork, 2011)—and particularly when short delays from the initial to the final test are utilized (although retrieval-induced forgetting has been observed with longer delays, as well; e.g., Storm, Bjork, & Bjork, 2012). In short, given the literature, it seems fair to say that testing with open-ended types of tests (i.e., cued- or free-recall) does not tend to lead to facilitated recall for competitive related information. The present findings, however, demonstrate that the use of multiple-choice testing with competitive alternatives offers a way to reliably strengthen the accessibility of such competitive nontested information—a finding that we would thus contend to be an important contribution to this general body of previous work.

Educational implications

The present research has demonstrated that a critical factor in the proper construction of multiple-choice questions, in terms of increasing their ability to invoke the type of retrieval processes known to support retention—particularly of related information—is that they have alternatives that are competitive. Thus, the implications of the present findings for the learning of educational materials seem both clear and important. Competitive information occurs naturally in a variety of the materials learned in educational contexts. Students, for example, are often required to learn competitive information about various regions of the world (e.g., the geography, climate, and people of different countries or regions) in a geography class. In anatomy, students must learn massive amounts of competitive information pertaining to the parts of the body. For the learning of such materials, the present research indicates that the use of properly constructed multiple-choice tests might be particularly valuable. Providing students with practice tests consisting of multiple-choice questions constructed with competitive alternatives, for example, could aid the students' learning of both the tested information and the competitive information associated with the incorrect alternatives.

Two possible limitations to the application of the present findings to the classroom are that the delays following the taking of the initial multiple-choice tests in the present experiments were considerable shorter than those used in the typical

educational setting, and that we did not compare multiple-choice testing to a time-on-task control. Pertaining to delay, there is reason to believe that such effects for multiple-choice questions with competitive alternatives would last over considerably longer delays. In recent research aimed at this issue, for example, we (Little & Bjork, 2012) found that the benefits for the retention of related information (as well as for tested information) following the giving of a multiple-choice test composed of questions with competitive alternatives persisted over a 48-h delay. Pertaining specifically to multiple-choice testing in educational contexts, E. L. Bjork, Little, and Storm (2014) found that multiple-choice quizzing consisting of questions constructed with competitive alternatives improved performance for questions pertaining to previously incorrect alternatives on the final exam in a large undergraduate course. Pertaining to the control conditions that we used as baselines in the two experiments, we showed that answering multiple-choice questions with competitive alternatives improved performance for related questions as compared to engaging in a distractor task or doing nothing. But we also showed that such testing improved performance relative to answering multiple-choice questions with noncompetitive alternatives, which was a time-on-task control. Thus, it is important to consider that simply spending extra time engaging with the material is not what improves performance, and not even testing in general; rather, it is a specific type of testing, with multiple-choice questions containing competitive alternatives, that leads to the improved retention of related information.

Concluding comment

Multiple-choice tests are typically considered *necessary evils* in educational contexts: a form of testing that should only be used when absolutely necessary. As we have clearly demonstrated in the present research, however, that reputation is unwarranted—at least with respect to their use as tools to promote learning. As tools of learning, properly constructed multiple-choice tests—that is, those with questions that include competitive alternatives—far from being an evil, appear to be particularly effective, especially for the learning of nontested competitive information.

Author Note J.L.L. is now at the Department of Psychology, Hillsdale College. A Collaborative Activity Award from the James S. McDonnell Foundation funded this research. We thank Ashley Kees for creating the materials and helping with data collection. We thank Robert Bjork, Barbara Knowlton, and the members of CogFog for helpful insights. Aspects of this research were reported in a poster presented at the 32nd Annual Conference of the Cognitive Science Society in Portland, Oregon, and appear as part of the dissertation of J.L.L.

References

Anderson, R. C., & Biddle, W. B. (1975). On asking people questions about what they are reading. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 9, pp. 89–132). New York, NY: Academic Press.

Anderson, M. C., Bjork, R. A., & Bjork, E. L. (1994). Remembering can cause forgetting: Retrieval dynamics in long-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 1063–1087. doi:10.1037/0278-7393.20.5.1063

Anderson, M. C., & Spellman, B. A. (1995). On the status of inhibitory mechanisms in cognition: Memory retrieval as a model case. *Psychological Review, 102,* 68–100. doi:10.1037/0033-295X.102.1.68

Bjork, R. A. (1975). Retrieval as a memory modifier: An interpretation of negative recency and related phenomena. In R. L. Solso (Ed.), *Information processing and cognition: The Loyola symposium* (pp. 123–144). Hillsdale, NJ: Erlbaum.

Bjork, E. L., Little, J. L., & Storm, B. C. (2014). Multiple-choice testing as a desirable difficulty in the classroom. *Journal of Applied Research in Memory and Cognition*.

Boker, J. R. (1974). Immediate and delayed retention effects of interspersing questions in written instructional passages. *Journal of Educational Psychology, 66,* 96–98.

Camp, G., Pecher, D., & Schmidt, H. G. (2007). No retrieval-induced forgetting using item-specific independent cues: Evidence against a general inhibitory account. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33,* 950–958. doi:10.1037/0278-7393.33.5.950

Carpenter, S. K., & DeLosh, E. L. (2006). Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. *Memory & Cognition, 34,* 268–276. doi:10.3758/BF03193405

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition, 20,* 633–642. doi:10.3758/BF03202713

Carroll, M., Campbell-Ratcliffe, J., Murnane, H., & Perfect, T. (2007). Retrieval-induced forgetting in educational contexts: Monitoring, expertise, text integration, and test format. *European Journal of Cognitive Psychology, 19,* 580–606. doi:10.1080/09541440701326071

Chan, J. C. K. (2009). When does retrieval induce forgetting and when does it induce facilitation? Implications for retrieval inhibition, testing effect, and text processing. *Journal of Memory and Language, 61,* 153–170. doi:10.1016/j.jml.2009.04.004

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

Dodd, M. D., Castel, A. D., & Roberts, K. E. (2006). A strategy disruption component to retrieval-induced forgetting. *Memory & Cognition, 34,* 102–111. doi:10.3758/BF03193390

Duchastel, P. C. (1981). Retention of prose following testing with different types of test. *Contemporary Educational Psychology, 6,* 217–226.

Foos, P. W., & Fisher, R. P. (1988). Using tests as learning opportunities. *Journal of Educational Psychology, 80,* 179–183.

Frase, L. T. (1967). Learning from prose material: Length of passage, knowledge of results and position of questions. *Journal of Educational Psychology, 58,* 266–272.

Frase, L. T. (1968). Effect of question location, pacing, and mode of retention of prose material. *Journal of Educational Psychology, 59,* 244–249.

Frase, L. T. (1971). Effect of incentive variables and type of adjunct questions upon text learning. *Journal of Educational Psychology, 62,* 371–375.

Glover, J. A. (1989). The "testing" phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology, 81,* 392–399.

Hamaker, C. (1986). The effects of adjunct question on prose learning. *Review of Educational Research, 56,* 212–242.

Jacoby, L. L., Shimizu, Y., Daniels, K. A., & Rhodes, M. G. (2005). Modes of cognitive control in recognition and source memory: Depth of retrieval. *Psychonomic Bulletin & Review, 12,* 852–857. doi:10.3758/BF03196776

Johansson, M., Aslan, A., Bäuml, K., Gäbel, A., & Mecklinger, A. (2007). When remembering causes forgetting: Electrophysiological correlates of retrieval-induced forgetting. *Cerebral Cortex, 17,* 1335–1341. doi:10.1093/cercor/bhl044

Little, J. L., & Bjork, E. L. (2012). The persisting benefits of using multiple-choice tests as learning events. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (pp. 683–688). Austin, TX: Cognitive Science Society.

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,* 1337–1344.

Little, J. L., Storm, B. C., & Bjork, E. L. (2011). The costs and benefits of testing text materials. *Memory, 19,* 346–359. doi:10.1080/09658211.2011.569725

MacLeod, C. M., Dodd, M. D., Sheard, E. D., Wilson, D. E., & Bibi, U. (2003). In opposition to inhibition. In B. H. Ross (Ed.), *The psychology of learning and motivation* (Vol. 43, pp. 163–214). San Diego, CA: Academic Press.

Macrae, C. N., & MacLeod, M. D. (1999). On recollections lost: When practice makes imperfect. *Journal of Personality and Social Psychology, 77,* 463–473. doi:10.1037/0022-3514.77.3.463

McGaw, B., & Grotelueschen, A. (1972). Direction of the effect of questions in prose material. *Journal of Educational Psychology, 63,* 586–588.

Perfect, T. J., Stark, L.-J., Tree, J. J., Moulin, C. J. A., Ahmed, L., & Hunter, R. (2004). Transfer appropriate forgetting: The cued-dependent nature of retrieval-induced forgetting. *Journal of Memory and Language, 51,* 399–417. doi:10.1016/j.jml.2004.06.003

Rickards, J. P. (1976). Interaction of position and conceptual level of adjunct questions on immediate and delayed retention of text. *Journal of Educational Psychology, 68,* 210–217.

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

Rothkopf, E. Z. (1966). Learning from written instructive materials: An exploration of the control of inspection behavior by test-like events. *American Educational Research Journal, 3,* 241–249.

Rothkopf, E. Z., & Bisbicos, E. E. (1967). Selective facilitative effects of interspersed questions on learning from written materials. *Journal of Educational Psychology, 58,* 56–61.

Rothkopf, E. Z., & Bloom, R. D. (1970). Effects of interpersonal interaction on the instructional value of adjunct questions in learning from written material. *Journal of Educational Psychology, 61,* 417–422.

Storm, B. C., Bjork, E. L., & Bjork, R. A. (2012). On the durability of retrieval-induced forgetting. *Journal of Cognitive Psychology, 24,* 617–629.

Watts, G. H., & Anderson, R. C. (1971). Effects of three types of inserted questions on learning from prose. *Journal of Educational Psychology, 62,* 387–394.

Whitten, W. B., & Leonard, J. M. (1980). Learning from tests: Facilitation of delayed recall by initial recognition alternatives. *Journal of Experimental Psychology: Human Learning and Memory, 6,* 127–134. doi:10.1037/0278-7393.6.2.127

Williams, C. C., & Zacks, R. T. (2001). Is retrieval-induced forgetting an inhibitory process? *American Journal of Psychology, 114,* 329–354.