Research Article

aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Psychological Science
23(11) 1337–1344

DOI: 10.1177/0956797612443370
http://pss.sagepub.com
SAGE

Multiple-Choice Tests Exonerated, at Least of Some Charges: Fostering Test-Induced Learning and Avoiding Test-Induced Forgetting

Jeri L. Little[1], Elizabeth Ligon Bjork[1], Robert A. Bjork[1], and Genna Angello[2]
[1]University of California, Los Angeles, and [2]Texas A&M University

Abstract

Among the criticisms of multiple-choice tests is that—by exposing the correct answer as one of the alternatives—such tests engage recognition processes rather than the productive retrieval processes known to enhance later recall. We tested whether multiple-choice tests could trigger productive retrieval processes—provided the alternatives were made plausible enough to enable test takers to retrieve both why the correct alternatives were correct and why the incorrect alternatives were incorrect. In two experiments, we found not only that properly constructed multiple-choice tests can indeed trigger productive retrieval processes, but also that they had one potentially important advantage over cued-recall tests. Both testing formats fostered retention of previously tested information, but multiple-choice tests also facilitated recall of information pertaining to incorrect alternatives, whereas cued-recall tests did not. Thus, multiple-choice tests can be constructed so that they exercise the very retrieval processes they have been accused of bypassing.

Keywords

memory, forgetting, educational psychology, cognitive processes, learning

Received 10/26/11; Revision accepted 2/20/12

Multiple-choice testing has a bad reputation and is frequently the target of disparaging comments, not only in the everyday conversations of students and teachers, but also in the academic literature. Nonetheless, multiple-choice tests are used in diverse settings and often for high-stakes purposes—ranging from obtaining a license to drive or practice medicine, to gaining acceptance to competitive academic programs. The use of multiple-choice tests tends to be justified, if at all, by pragmatic arguments (e.g., the argument that in large undergraduate courses, the grading of short-answer or essay tests becomes prohibitively burdensome or unreliable). Multiple-choice testing is, in short, often regarded as a necessary evil.

Among the long-standing criticisms of multiple-choice tests is that they do not measure complex learning well (e.g., Frederiksen, 1984). Additionally, they are accused of failing to engage the kind of retrieval processes that support long-term retention—that is, retrieval-induced learning (e.g., Chan, McDermott, & Roediger, 2006; Foos & Fisher, 1988). The validity of that second criticism is the focus of the present research.

Test-Induced Learning

It has long been known that retrieving information from memory is a powerful learning event—making the retrieved information more recallable in the future than it would have been otherwise (e.g., Bjork, 1975)—and that testing can trigger such retrieval processes. Recently, however, interest in and research on the practical implications of test-induced learning has intensified, and articles on the implications of such research for educational practices have appeared not only in discipline-based journals, but also in broader publications of scientific research (e.g., Karpicke & Blunt, 2011; Karpicke & Roediger, 2008; Pyc & Rawson, 2010; Roediger & Finn, 2009; Smith et al., 2009) and in public media such as *The New York Times* (e.g., Belluck, 2011; Carey, 2010).

Exactly why retrieval is more powerful than, for example, a restudy opportunity is a matter of current research, but there

Corresponding Author:
Jeri L. Little, Psychology Department, Washington University in St. Louis, Box 1125, One Brookings Dr., St. Louis, MO 63130
E-mail: jerilittle@gmail.com

is general agreement that retrieving information from memory alters its representation in a way that improves its future accessibility, and the empirical evidence of such retrieval-based learning is dramatic and widespread (e.g., Carrier & Pashler, 1992; for an excellent review, see Roediger & Karpicke, 2006). In contrast, recognition tests tend not to improve subsequent retention to the same degree (Carpenter & DeLosh, 2006; Glover, 1989). Thus, critics have argued that multiple-choice tests are less effective as learning events than cued-recall or short-answer tests are because they present the correct answer among the alternatives, thereby bypassing the need for retrieval and allowing test takers to rely on recognizing the correct answer. In accord with that view, studies have shown multiple-choice tests to be less effective than cued-recall tests in enhancing the later recall of tested information (e.g., R. C. Anderson & Biddle, 1975; Duchastel, 1981; Foos & Fisher, 1988; Hamaker, 1986; McDaniel, Anderson, Derbish, & Morrisette, 2007).

Is it actually the case, however, that multiple-choice questions fail to trigger productive retrieval processes? Although it is possible for them to be written so as to rely primarily on recognition processes, it is certainly not necessary that they do so. It is difficult to see, for instance, how recognition processes—in comparison with actual computation or recall processes—can play much of a role in answering a multiple-choice mathematics question for which finding the correct answer requires a calculation. Moreover, it seems that plausible incorrect alternatives in a multiple-choice question would encourage the recall of why they are incorrect. One goal of the present research was to see whether multiple-choice questions, when constructed with competitive incorrect alternatives, can trigger retrieval processes that enhance more than just the later recall of correct answers and their associated information: Can such questions also enhance the recall of initially incorrect answers (and their associated information) when, on a later cued-recall test, such alternatives become correct answers to related questions?

From a standpoint of applications to educational contexts, the possibility that properly constructed multiple-choice practice questions might facilitate the recall of related, but not explicitly tested, information is particularly important. Practice tests, after all, are not typically constructed to be identical to a later examination, but, instead, are constructed to consist of questions of the type that will be asked later. Therefore, currently incorrect alternatives could become correct answers to related questions on later examinations, and this possibility presents an important consideration given the evidence, summarized in the next section, that initial cued-recall testing can, under some circumstances, impair rather than assist the later recall of competitive alternatives.

Test-Induced Forgetting

The possibility that cued-recall practice tests could have negative effects is suggested by evidence that repeated recall of a target item can result in *retrieval-induced forgetting* of other items associated with the same or similar cues (M. C. Anderson, Bjork, & Bjork, 1994). M. C. Anderson et al. used the retrieval-practice paradigm for category-exemplar pairs from a number of categories (e.g., *Fruit: Banana; Fruit: Orange; Tree: Fir; Tree: Redwood*); after the initial study of all of the pairs, participants practiced the retrieval of half of the items from half of the categories (e.g., *Fruit: Ba____*). The results for a later test showed that recall of the unpracticed items in a practiced category (e.g., *Fruit: Orange*) was lower than recall of items from initially studied categories from which no members had appeared in retrieval practice (e.g., *Trees*; control condition). Since this initial observation, retrieval-induced forgetting has proved to be a robust phenomenon, occurring in a variety of similar paradigms and for different types of material (e.g., Levy, McVeigh, Marful, & Anderson, 2007; Macrae & MacLeod, 1999; Radvansky, 1999; Saunders, Fernandes, & Kosnes, 2009; Shaw, Bjork, & Handal, 1995; Storm, Bjork, & Bjork, 2005).

The educational implications of these results are significant. Whether it is the case, however, that giving practice quizzes actually impairs students' future recall of related, but initially untested, material on a later more comprehensive test is unclear. Work investigating retrieval-induced forgetting with educational materials has produced mixed results, with some researchers finding reduced recall for related materials (e.g., Carroll, Campbell-Ratcliffe, Murnane, & Perfect, 2007; Little, Storm, & Bjork, 2011; Macrae & MacLeod, 1999), and others finding facilitated recall for related materials (e.g., Chan et al., 2006; Frase, 1967; Rothkopf & Billington, 1974). It is possible that positive effects occur when initial-test questions induce the retrieval of related information in order to access the target information. That is, in some cases, nontarget information may serve as a kind of mediator in the search for the target information and, thus, be retrieved and considered in that process (e.g., Chan et al., 2006), rather than needing to be selected against and suppressed. Similarly, in the case of a multiple-choice question, the presence of plausible incorrect alternatives might lead a learner to access information pertaining to those alternatives while trying to decide on the correct answer. Consequently, to the extent that such information is helpful in answering a related question on a later test, performance on that related question could be facilitated.

Experiment 1 was designed to compare the effects of an initial cued-recall test and an initial multiple-choice test on participants' performance on a later test containing both questions that were identical to those asked on the initial test and questions that were related to them.

Experiment 1

Method

Thirty-two undergraduates at the University of California, Los Angeles, participated for course credit. Participants studied two 1,100-word passages (about Yellowstone Park and Saturn)

in succession for 9 min each, with presentation order counterbalanced across participants. They then received either a cued-recall or a multiple-choice test on information from one of the passages, with the specific passage tested being counterbalanced across participants. Following a 5-min distractor task (playing Tetris), a final cued-recall test on information from both passages was administered.

Ten pairs of multiple-choice questions were constructed for each passage. Questions in each pair were randomly assigned to one of two sets (A and B) and related by virtue of addressing the same topic (e.g., geysers in Yellowstone Park) and having the same alternative answers (e.g., "Old Faithful," "Steamboat Geyser," "Castle Geyser," and "Daisy Geyser"), but different correct answers (e.g., "Question: "What is the tallest geyser in Yellowstone National Park?" Answer: "Steamboat Geyser"; vs. Question: "What is the oldest geyser in Yellowstone National Park?" Answer: "Castle Geyser"). For passages initially tested, participants receiving a multiple-choice test were asked the 10 questions from just one set (counterbalanced across participants), with the alternative answers presented. The cued-recall test contained the same questions, but without the alternatives presented. On the final cued-recall test, all 40 questions were asked (i.e., all questions from Sets A and B for both passages), but without the alternatives presented.

On the initial test, participants had 24 s to answer each question and were encouraged to spend the full time in thinking about the question and their answer. Our goal was to provide students with ample time to consider possible answers to the questions. After the 10 questions were presented once, they were asked again in a new random order, with no feedback provided during either cycle.

On the final test, participants received up to 20 s to answer each question. Each participant received the initial-test questions about only one of the two texts and from Set A or Set B only. Therefore, if one question from a given question pair had been asked twice earlier (during the initial test), the other question had not been asked earlier. Accordingly, we designated these items from the tested passage as previously tested and related items, respectively, and designated items from the nontested passage as control items. Counterbalancing procedures ensured that, across participants, all questions served as previously tested, related, and control items.

To control for output interference on the final test, and because the comparison of performance on related and control questions was the most crucial comparison, we ensured that the first 20 questions were the 10 related questions and 10 of the 20 control questions, and that the last 20 questions were the 10 previously tested questions and the remaining 10 control questions. In the analyses reported in the next section, final-test performance on previously tested items and on related items was compared with performance on corresponding control items, that is, control items presented in the same half of the test.

Results and discussion

Initial-test performance. As might be expected, participants given a multiple-choice test following their reading of the passages answered more of the questions correctly (M = 70%, SD = 17%) than did participants given a cued-recall test (M = 43%, SD = 16%).

Final-test performance. The critical comparison—with respect to whether multiple-choice tests might result in as much test-induced learning as cued-recall tests—was between final-test performance on previously tested items and final-test performance on their corresponding control items. As is apparent from the percentages of correct responses, illustrated in Figure 1, taking either type of initial test enhanced cued-recall performance on the final test. Planned paired-samples t tests indicated that both multiple-choice and cued-recall tests enhanced final-test performance for previously tested items compared with that for control items, $t(15) = 5.093$, $d = 1.30$, $p < .001$, and $t(15) = 3.578$, $d = 0.67$, $p < .01$, respectively. A 2 × 2 mixed-model analysis of variance (ANOVA) revealed that the apparent interaction between item type (previously tested vs. control) and initial-test type (multiple choice vs. cued recall; see Fig. 1) was indeed significant, $F(1, 30) = 5.499$, $MSE = 1.503$, $\eta_p^2 = .15$, $p < .05$. Thus, the initial multiple-choice test improved later recall of the tested information more than did the initial cued-recall test.

Figure 2 illustrates the appropriate comparisons regarding the other critical question: What happened to the retention of related items (compared with control items) following an initial multiple-choice test versus an initial cued-recall test? The results are striking: Although the initial cued-recall test led to lower final-test performance on related items relative to

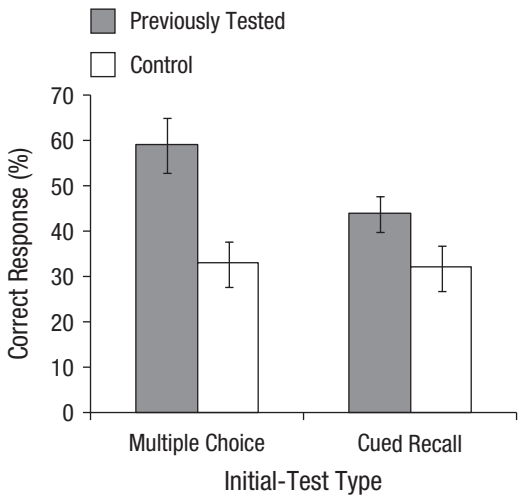

Fig. 1. Results of Experiment 1: percentage of correct responses on the final test as a function of initial-test type (multiple choice vs. cued recall) and whether the final-test items had been previously tested or were control items. Error bars represent ±1 *SEM*.

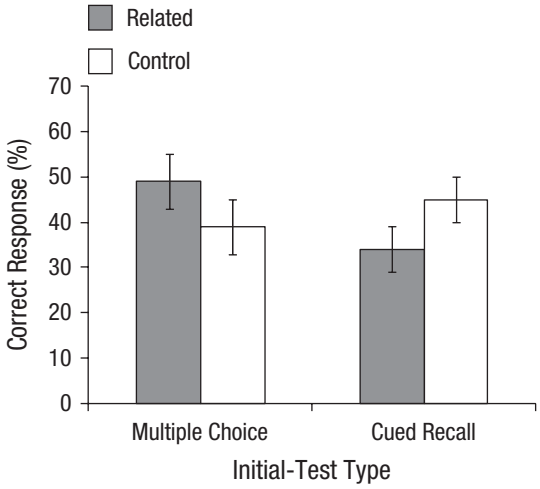

Fig. 2. Results of Experiment 1: percentage of correct responses on the final test as a function of initial-test type (multiple choice vs. cued recall) and whether the final-test items were related to the questions on the initial test or were control items. Error bars represent ±1 *SEM*.

control items, the initial multiple-choice test produced, if anything, enhanced final-test recall of related items. A 2 × 2 mixed-model ANOVA revealed that the interaction between final-test item type (related vs. control) and initial-test type (multiple choice vs. cued recall) was indeed significant, $F(1, 30) = 7.39$, $MSE = 2.17$, $\eta_p^2 = .19$, $p < .05$. Planned paired-samples t tests confirmed that when participants had taken an initial cued-recall test, final-test performance on related items was worse than that on their corresponding control items, $t(15) = 2.18$, $d = 0.55$, $p < .05$, but when participants had taken an initial multiple-choice test, final-test performance on related items was numerically, though not significantly, higher than that on their corresponding control items, $t(15) = 1.70$, $d = 0.54$, $p > .05$.

The pattern of results obtained in Experiment 1 is consistent with our conjecture that multiple-choice items with competitive alternatives can trigger productive retrieval processes, but the results go beyond that general expectation and suggest that multiple-choice practice tests may have an important advantage over cued-recall tests. An initial multiple-choice test can not only improve performance for items that are repeated on a final test, but also enhance retrieval of information associated with incorrect alternatives on the initial practice test.

One might argue, however, that performance on previously tested information during the final test in Experiment 1 reflects initial-test performance more than the relative potency of each type of test as a learning tool. In fact, for questions answered correctly on the initial test, taking an initial cued-recall test led to better retention than did taking an initial multiple-choice test. That result could reflect nothing more than item selection, but it nonetheless reinforces concerns about how to interpret final-test performance given differences in initial-test performance between the two testing formats.

To address that concern, we replicated Experiment 1, with this difference: In Experiment 2, we provided participants with feedback, so that they would have the opportunity to see the correct answer to every question on the initial test, regardless of the initial test's type. In fact, given prior findings (see, e.g., Kang, McDermott, & Roediger, 2007), providing feedback after each initially tested item might be expected even to reverse the differences found in Experiment 1. That is, after receiving feedback, participants might recall answers to questions initially tested with a multiple-choice test less well than answers to questions initially tested with a cued-recall test.

More critical to the focus of the present research, however, is how the provision of feedback on an initial multiple-choice test might affect the retention of related information. On the one hand, receiving feedback might improve one's ability to answer a later related question—because the feedback might provide an opportunity to correct an erroneous belief. If a student incorrectly chooses Steamboat Geyser as the oldest geyser in Yellowstone National Park, for example, he or she might also be thinking, incorrectly, that Castle Geyser (another alternative) is the tallest geyser, when actually Steamboat Geyser is the tallest one, and Castle Geyser is the oldest one. If the student is then given feedback indicating that Castle Geyser is the oldest geyser, then—because there can be only one oldest geyser and only one tallest geyser—he or she might reconsider the significance of Steamboat Geyser, perhaps remembering that it is actually the tallest geyser. From this perspective, a test with feedback might help one recall both previously tested and nontested, but related, information more than a test without feedback would. On the other hand, the answers provided as corrective feedback might sometimes interfere later with the recall of initially incorrect answers that are the correct answers to related questions on the final test, and any such interference might lead to lower—rather than enhanced—recall of that related information.

Experiment 2

Method

Ninety-six undergraduates at Washington University in St. Louis participated in Experiment 2 for partial course credit. The materials used in Experiment 2 were the same as those employed in Experiment 1. The procedure, too, was the same—with two exceptions. First, feedback was provided for half of the participants. That is, although for half of the participants (no-feedback condition), the initial-test procedure was the same as in Experiment 1 (except for the time difference noted in the next sentence), the remaining participants (feedback condition) received corrective feedback after answering each question during the initial test. Second, the time allocated to answer each question was modified: Participants in the no-feedback condition were given 25 s to answer each question; participants in the feedback condition were given 22 s to answer each question, and then immediately received feedback (i.e., the correct

answer) for 3 s. Feedback was the same for both initial-test types (multiple choice and cued recall).

Results and discussion

Initial-test performance. Participants completed an initial test comprising two cycles of 10 questions in immediate succession. When given an initial multiple-choice test without feedback, participants answered 78% (*SD* = 19%) and 80% (*SD* = 19%) of the questions correctly on the first and second cycles, respectively. When given an initial multiple-choice test with feedback, participants answered 75% (*SD* = 14%) and 99% (*SD* = 4%) of the questions correctly on the first and second cycles, respectively. Participants given an initial cued-recall test without feedback answered 47% (*SD* = 20%) and 50% (*SD* = 20%) of the questions correctly on the first and second cycles, respectively, and those given an initial cued-recall test with feedback answered 43% (*SD* = 17%) and 87% (*SD* = 19%) of the questions correctly on the first and second cycles, respectively.

Final-test performance. Final-test performance for previously tested items and the corresponding control items is shown in Figure 3. As is readily apparent, the provision of feedback—compared with the lack of feedback—during the initial test improved performance for information tested by both initial cued-recall and multiple-choice tests. A 2 × 2 ANOVA testing the apparent interactive effect of type of initial test (multiple choice vs. cued recall) and provision of feedback (no feedback vs. feedback) on performance for previously tested information revealed, however, that although performance was generally better with feedback than without, this improvement in performance was greater for the initial cued-recall test than for the initial multiple-choice test, $F(1, 92) = 16.63$, $MSE = 2.90$, $\eta_p^2 = .15$, $p < .01$. In both cases, the provision of feedback led to very high levels of final-test performance, and although performance was better for items initially tested with cued recall, this difference was not significant, $t(46) = 1.55$, $p > .05$.

Figure 4 shows the appropriate comparisons regarding the other critical question: How did performance on the related items differ from performance on the corresponding control items following the two initial-test types and the presence or absence of feedback? As the figure illustrates, regarding performance for related information, we observed no interaction between initial-test type and whether feedback was provided, $F < 1$. Furthermore, an independent-samples *t* test confirmed that regardless of whether feedback was provided on the initial test, initial multiple-choice testing resulted in better final-test performance for related information than did initial cued-recall testing, $t(94) = 2.88$, $d = 0.59$, $p < .01$, and that finding is consistent with the effect demonstrated in Experiment 1. Paired-samples *t* tests confirmed that recall of related, but nontested, information was enhanced, compared with recall of control information, as a consequence of taking an initial multiple-choice test, $t(47) = 3.01$, $d = 0.43$, $p < .01$, whereas taking an initial cued-recall test led to no benefits for recall of related information; in fact, recall of related information was generally worse than recall of control information following an initial cued-recall test, although this difference was nonsignificant, $t(47) = 0.95$, $p > .05$.

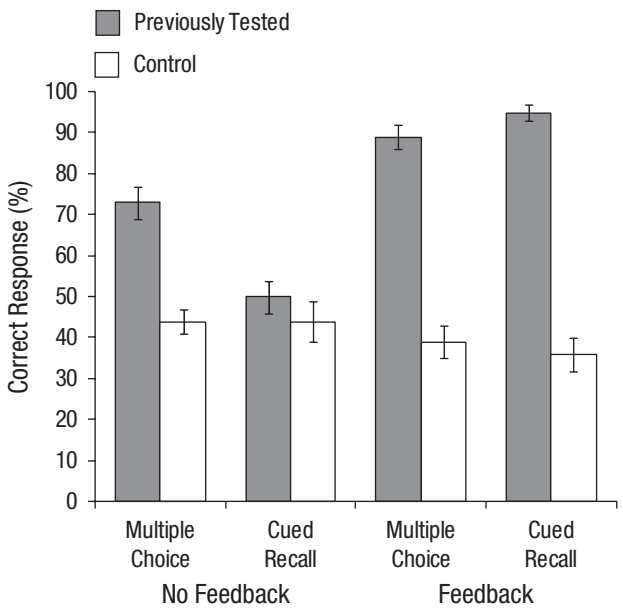

Fig. 3. Results of Experiment 2: percentage of correct responses on the final test as a function of initial-test type (multiple choice vs. cued recall), whether or not feedback had been provided on the initial test, and whether the final-test items had been previously tested or were control items. Error bars represent ±1 *SEM*.

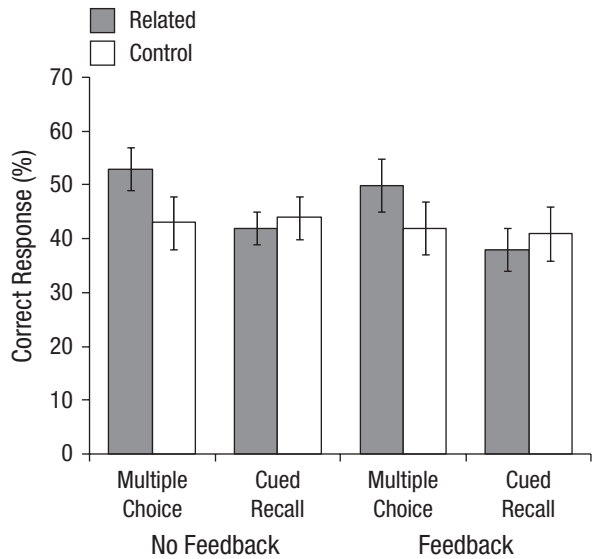

Fig. 4. Results of Experiment 2: percentage of correct responses on the final test as a function of initial-test type (multiple choice vs. cued recall), whether or not feedback had been provided on the initial test, and whether the final-test items were related to the questions on the initial test or were control items. Error bars represent ±1 *SEM*.

Thus, regardless of whether feedback was provided, an initial cued-recall test did not result in reliable retrieval-induced forgetting in Experiment 2. Nevertheless, the lower performance for related than for control items after an initial cued-recall test remains striking given the facilitated performance for related relative to control items following an initial multiple-choice test. In both experiments, an initial multiple-choice test produced retrieval-induced learning (numerically so in Experiment 1; significantly so in Experiment 2), as measured by performance on related questions (vs. control questions) on the final test, and that was true regardless of whether feedback was provided.

The results of Experiment 2 suggest that, with respect to performance on final-test questions that are the same as initial-test questions, providing feedback increases the potency of initial cued-recall testing more than it increases the potency of initial multiple-choice testing. This finding is not surprising and, indeed, is consistent with prior research (Kang et al., 2007). Perhaps because performance was much lower on the initial cued-recall test than on the initial multiple-choice test, there was more opportunity for an increase in final-test performance in the cued-recall condition when feedback was given.

These results are interesting from the standpoint of transfer-appropriate processing. Morris, Bransford, and Franks (1977) demonstrated that retention is enhanced when the processes engaged at the time of study overlap those engaged at the time of a later test—that is, when the processing at the time of study is "transfer appropriate." We argue that the initial test in our experiments functioned as an additional opportunity to encode the information. Thus, from the standpoint of transfer-appropriate processing, it is surprising that for items retested on the final test, an initial cued-recall test did not improve final-test performance more than did an initial multiple-choice test. The final, or criterion, test was a cued-recall test, not a multiple-choice test, and transfer from an initial test to a final test might be expected to be greater when the tests are in the same format than when they mismatch, especially when feedback is provided during the initial test. What we found, however, was when feedback was provided, final-test recall of previously tested information was approximately the same regardless of whether the initial test was a multiple-choice test or a cued-recall test. Moreover, we found that the advantage of multiple-choice testing for the recall of information related to initially incorrect alternatives was sustained even when participants were given feedback during the initial test.

General Discussion

The present findings vindicate multiple-choice tests, at least of charges regarding their use as practice tests. In fact, our findings suggest that when multiple-choice tests are used as practice tests, they can provide a win-win situation: Specifically, they can foster test-induced learning not only of previously tested information, but also of information pertaining to the initially incorrect alternatives. This latter advantage is especially important because, typically, few if any practice-test items are repeated verbatim on the subsequent real test. From that standpoint, the advantage of initial multiple-choice testing over initial cued-recall testing is a truly significant one.

The importance of alternatives' being competitive

A major proviso with respect to the benefits of multiple-choice testing, however, is that such tests must be properly constructed; that is, they must include plausible (i.e., competitive) incorrect alternatives of the kind that can trigger the retrieval processes that foster test-induced learning and deter test-induced forgetting. In fact, in earlier work, we were able to demonstrate that when the incorrect alternatives on an initial test are not competitive, the later recall of information pertaining to those alternatives is not enhanced (Little & Bjork, 2010). For example, participants answered an initial-test question asking which outer planet was discovered by mathematical predictions rather than by direct observation (the correct answer is Neptune). Later, their recall of information pertaining to one of the incorrect alternatives provided for this question was enhanced only if those alternatives had been competitive (e.g.,"Uranus" and "Saturn," which are generally known to be outer planets)—not if they had been noncompetitive (e.g., "Mercury" and "Mars," which are generally known to be inner planets). An implication of this result is that simple exposure to incorrect alternatives does not by itself convey the benefits of initial multiple-choice testing, because when an item was used as a noncompetitive alternative on the initial multiple-choice test, its recall on a later test was not enhanced compared with control items (Little & Bjork, 2010). Thus, for multiple-choice tests to function as effective practice tests, it appears that they must present incorrect alternatives in a way that will induce students to recall why those alternatives are incorrect; such recall, in turn, can lead to subsequent recall of the information if it is tested later, say, during a final examination.

Remaining issues

Although the present findings support the possibility that properly constructed multiple-choice practice tests can be important learning events, some issues remain. One has to do with how long the benefits of multiple-choice testing last. Existing findings (see Roediger & Karpicke, 2006, for a review) suggest that the benefits for retested items are likely to be long-lived, but it is especially important, from a practical standpoint, to know how long the benefits for untested or related information persist. In a recent study aimed at this issue, we found that the benefits for untested or related information, as well as for tested information, persist over a 48-hr period (Little & Bjork, 2012).

Another important question is whether taking a properly constructed multiple-choice test can trigger retrieval strategies

more broadly, rather than simply triggering the retrieval of why a given alternative is the right or wrong answer. If such a test triggers retrieval of studied content more broadly, it is possible that the later recall of items that could have been presented as incorrect alternatives, but were not, will be enhanced, and such enhancement would increase the benefits of using multiple-choice tests in educational settings.

Conclusion

The present work demonstrates that properly constructed multiple-choice practice tests can be important learning events for students. Achieving "proper construction" of such tests—which requires that incorrect alternatives be plausible, but not so plausible that they are unfair—is, however, a challenge. As any teacher who has used multiple-choice tests can testify, writing good multiple-choice items is very hard work, whereas writing poor ones is relatively easy. Thus, when people accuse multiple-choice tests of being bad tests, that accusation, statistically, has some truth to it. We argue, however, that the statistical accuracy of such accusations has more to do with human nature than with the multiple-choice format per se.

Declaration of Conflicting Interests

The authors declared that they had no conflicts of interest with respect to their authorship or the publication of this article.

Funding

A Collaborative Activity Award from the James S. McDonnell Foundation funded this research.

References

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.

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.

Belluck, P. (2011, January 20). To really learn, quit studying and take a test. *The New York Times*. Available from http://www.nytimes.com

Bjork, R. A. (1975). Retrieval as a memory modifier. In R. Solso (Ed.), *Information processing and cognition: The Loyola Symposium* (pp. 123–144). Hillsdale, NJ: Erlbaum.

Carey, B. (2010, September 6). Forget what you know about good study habits. *The New York Times*. Available from http://www.nytimes.com

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.

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

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.

Chan, J. C., 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.

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.

Frederiksen, N. (1984). The real test bias: Influences of testing on teaching and learning. *American Psychologist*, *39*, 193–202.

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.

Kang, S. H. K., McDermott, K. B., & Roediger, H. L., III. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology*, *19*, 528–558.

Karpicke, J. D., & Blunt, J. R. (2011). Retrieval practice promotes more learning than elaborative studying with concept mapping. *Science*, *331*, 772–775.

Karpicke, J. D., & Roediger, H. L., III. (2008). The critical importance of retrieval for learning. *Science*, *319*, 966–968.

Levy, B. J., McVeigh, N. D., Marful, A., & Anderson, M. C. (2007). Inhibiting your native language: The role of retrieval-induced forgetting during second-language acquisition. *Psychological Science*, *18*, 29–34.

Little, J. L., & Bjork, E. L. (2010). Multiple-choice testing can improve the retention of non-tested related information. In S. Ohisson & R. Catrabone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 1535–1540). Austin, TX: Cognitive Science Society.

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., Storm, B. C., & Bjork, E. L. (2011). The costs and benefits of testing text materials. *Memory*, *19*, 346–359.

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

McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology*, *19*, 494–513.

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior*, *16*, 519–533.

Pyc, M. A., & Rawson, K. A. (2010). Why testing improves memory: Mediator effectiveness hypothesis. *Science*, *330*, 335.

Radvansky, G. A. (1999). Memory retrieval and suppression: The inhibition of situation models. *Journal of Experimental Psychology: General*, *128*, 563–579.

Roediger, H. L., & Finn, B. (2009, October 20). Getting it wrong: Surprising tips on how to learn. *Scientific American*. Retrieved from http://www.scientificamerican.com/article.cfm?id=getting-it-wrong

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.

Rothkopf, E. Z., & Billington, M. J. (1974). Indirect review and priming through questions. *Journal of Educational Psychology*, *66*, 669–679.

Saunders, J., Fernandes, M., & Kosnes, L. (2009). Retrieval-induced forgetting and mental imagery. *Memory & Cognition*, *37*, 819–828.

Shaw, J. S., Bjork, R. A., & Handal, A. (1995). Retrieval-induced forgetting in an eyewitness memory paradigm. *Psychonomic Bulletin & Review*, *2*, 249–253.

Smith, M. K., Wood, W. B., Adams, W. K., Weiman, C., Knight, J. K., Guild, N., & Su, T. (2009). Why peer discussion improves student performance on in-class concept questions. *Science*, *323*, 122–124.

Storm, B. C., Bjork, E. L., & Bjork, R. A. (2005). Social metacognitive judgments: The role of retrieval-induced forgetting in person memory and impressions. *Journal of Memory and Language*, *52*, 535–550.