Psychonomic Bulletin & Review
2007, 14 (2), 194-199

The memorial consequences of multiple-choice testing

Elizabeth J. Marsh
Duke University, Durham, North Carolina

Henry L. Roediger III
Washington University, St. Louis, Missouri

and

Robert A. Bjork and Elizabeth L. Bjork
University of California, Los Angeles, California

The present article addresses whether multiple-choice tests may change knowledge even as they attempt to measure it. Overall, taking a multiple-choice test boosts performance on later tests, as compared with non-tested control conditions. This benefit is not limited to simple definitional questions, but holds true for SAT II questions and for items designed to tap concepts at a higher level in Bloom's (1956) taxonomy of educational objectives. Students, however, can also learn false facts from multiple-choice tests; testing leads to persistence of some multiple-choice lures on later general knowledge tests. Such persistence appears due to faulty reasoning rather than to an increase in the familiarity of lures. Even though students may learn false facts from multiple-choice tests, the positive effects of testing outweigh this cost.

Testing is ubiquitous in education. Tests are used to hold schools accountable for their students' progress and to monitor the advancement (or lack thereof) of individual students. Recent policy shifts such as *No Child Left Behind* have added to a zeitgeist of testing, albeit one that is controversial. In policy and in educational research, tests are viewed as tools for measuring students' mastery of skills and knowledge. Research questions and practical concerns revolve around the fairness of the tests and whether the tests are measuring the qualities they are supposed to assess. Although issues of test content, scoring, and bias are important, the emphasis on testing as assessment can lead to the presumption that tests measure a student's knowledge without affecting that knowledge. In fact, however, research conducted in both experimental and educational settings demonstrates that tests not only measure what is learned, but also alter the nature and accessibility of students' knowledge (e.g., Roediger & Karpicke, 2006b; Sternberg & Grigorenko, 2001).

Our focus in the present review is on multiple-choice tests, the most common format of objective tests. Multiple-choice tests are popular with educators because of the ease and perceived objectivity of grading them. The question here is how taking a multiple-choice test may change students' knowledge.

The Testing Effect

Testing generally improves students' performance on a later test relative to conditions in which students are not tested after learning (Bjork, 1975; Carrier & Pashler, 1992; Foos & Fisher, 1988; Gates, 1917; Glover, 1989; McDaniel & Masson, 1985; Roediger & Karpicke, 2006a; Spitzer, 1939). A review of 35 studies on frequent testing in classrooms supports the experimentalist's belief that the testing effect generalizes to the classroom; 29 studies found positive effects of testing and only 6 found negative effects (Bangert-Drowns, Kulik, & Kulik, 1991).

Although many of the classic studies on the testing effect examined the memorial consequences of initial free recall tests, recognition or multiple-choice tests also yield benefits (Hogan & Kintsch, 1971). One of the earliest studies on testing, conducted by Spitzer (1939) and involving 3,605 sixth graders from Iowa, examined the effects of eight different testing schedules on the students' memory for two passages, one on peanuts and the other on bamboo. Each of Spitzer's 25 multiple-choice items paired the correct answer with four incorrect choices. At various points in time after reading the passages, Spitzer compared the performance of students who were taking the test for the first time with that of those who were taking it for the second time. As long as the initial test occurred within three weeks of learning, students who had previously taken a test outperformed students who had not. A week after learning, for example, previously tested sixth graders answered an average of 11.8 items correctly on the final test, whereas students taking a test for the first time answered only 7.9 items correctly.

E. J. Marsh, emarsh@psych.duke.edu

The benefits of taking an initial multiple-choice test are also obtained when students later have to produce a fact, rather than recognize it. One example of this benefit, using educationally relevant materials, comes from our own work (Roediger & Marsh, 2005). Our subjects read a subset of passages on nonfiction topics (e.g., the sun). Across subjects, passage facts were either tested on an initial multiple-choice test or not tested at that time; they received no feedback on their multiple-choice answers. Five minutes later, all subjects took a final open-ended (cued recall) test. Students correctly answered more questions such as "How many planets separate Jupiter from the sun?" if they had previously answered a parallel multiple-choice question ($M = 63\%$) than if they had not been tested on the fact ($M = 40\%$). When the final test was delayed one week, the testing effect was reduced but, importantly, still significant (Fazio, Marsh, & Roediger, 2006).

In the domain of semantic memory, which is perhaps of most interest to educators, testing effects occur even when students have not recently studied the material. That is, simply taking a multiple-choice general knowledge test boosts performance on a later cued recall test, even if there was no study phase in the experiment (Roediger & Marsh, 2005). In this case, the test links to and activates preexperimental knowledge. The test may cue knowledge that might not have been retrieved otherwise, such as marginal knowledge (Berger, Hall, & Bahrick, 1999), which is defined as knowledge that is available but not readily accessible (Tulving & Pearlstone, 1966). One example might be the name of the author of the fable "The Fox and the Sour Grapes." Berger and colleagues demonstrated that such marginal knowledge can be made accessible and is subsequently very slowly forgotten if the answer (Aesop, in this case) is presented. In principle, and almost certainly in practice, presenting marginal knowledge as one alternative on a multiple-choice test is likely to have similar positive effects. In addition, multiple-choice tests may teach students new facts, because students use reasoning to eliminate lures, select the correct answer, and thereby learn it.

In short, multiple-choice tests yield large memorial benefits for numerous reasons (Roediger & Guynn, 1996; Roediger & Karpicke, 2006b). Tests serve as an additional study opportunity, offer retrieval practice, and provide retrieval cues in the form of answer options. Compared with restudying material, testing is also a better processing match to later tests. Given the myriad ways in which multiple-choice tests can aid memory, more testing might seem the obvious recommendation to improve learning and performance. Recent research, however, suggests that testing may also change memory in ways that are not always for the better. The question we now address is whether multiple-choice tests also provide opportunities for the learning and retrieval practice of incorrect answers.

Negative Side Effects of Testing

One concern about multiple-choice tests is that they routinely expose students to wrong answers. In four-alternative multiple-choice tests, for example, three alternatives are wrong and only one is correct. If subjects read all choices carefully, they read three (usually) plausible wrong answers and only one correct answer. Even if subjects pick the correct answer, reading the wrong statements may make those answers seem true later. That is, simply repeating statements increases the probability that those statements will be judged true later (Hasher, Goldstein, & Toppino, 1977). Consistent with this analysis, testing increases later ratings of the truth of multiple-choice lures, although they are still rated as less true than known facts (Toppino & Brochin, 1989; Toppino & Luipersbeck, 1993). Similarly, testing increases the production of multiple-choice lures as answers to later cued recall questions, even when students are strictly warned against guessing (Roediger & Marsh, 2005). Specifically, multiple-choice lures were used to answer 5% of questions when subjects had not been previously tested; testing increased the use of these specific wrong answers to 12% on the later cued recall test.

What mechanism drives the persistence of multiple-choice lures on later tests? By persistence, we mean that an error made on a previous multiple-choice test is also produced—persists—on the final cued recall test. In our experiments involving multiple-choice tests, many errors were made that were never intruded again. What mechanism allows for some, but not all, errors to persist?

To answer this question, we examined the relationship between students' answers on the two tests. That is, if mere familiarity were driving the effect (akin to an illusory truth effect), then persistence of errors should not have been limited to previously selected lures—because, presumably, even nonselected lures would have been read, and thus would have accrued familiarity. When, however, students answered a final cued recall question with a lure they had read on the previous multiple-choice test, their error was almost always the same as the answer they had previously chosen. Rarely did students select the correct answer on the initial test and then produce a lure on the final test. Nor were students likely to select Lure A on the first test and then produce Lure B on the final test (Butler, Marsh, Goode, & Roediger, 2006; Roediger & Marsh, 2005). Errors that persisted were those that had been endorsed on the first test.

The importance of selecting a multiple-choice lure sheds light on a puzzle in the literature. Specifically, manipulations of the number of multiple-choice lures have had inconsistent effects on later tests (Brown, Schilling, & Hockensmith, 1999; Roediger & Marsh, 2005; Whitten & Leonard, 1980). For example, Whitten and Leonard found a benefit in later free recall for words that had been grouped with more lures on a prior recognition test, whereas Roediger and Marsh showed a cost. Although there were many differences between Whitten and Leonard (1980) and Roediger and Marsh (2005), follow-up studies suggested that the critical difference was the level of performance on the initial multiple-choice test (Butler et al., 2006). Subjects in Whitten and Leonard's list learning experiment were successful in selecting the studied item even when it was grouped with additional nonstudied words, whereas subjects in our prose learning experiments were less able to select the correct answer when it was

grouped with more lures. The memorial consequences of adding multiple-choice lures depends on whether subjects are able to select the correct answer or if they are increasingly likely to select a lure.

Given the importance of selecting a lure on an initial multiple-choice test, we conducted a study in which we examined subjects' reasons for their answer choices (Huelser & Marsh, 2006). All subjects took an initial multiple-choice test, and then a final cued recall test after a short delay. The key manipulation was that half the subjects had to explain why they had chosen their answers for the multiple-choice questions. After answering each question, they responded to the prompt *Please type why you selected that answer for the previous question*. The control subjects simply answered each question to the best of their ability.

Writing explanations for answer choices did not change the pattern of the data, nor did it interact with other variables. We felt comfortable, therefore, using subjects' overtly reported strategies as a measure of what they typically did when not instructed to verbalize their thought processes. A coding scheme was created that classified most of the explanations into one of eight categories: guessing, process of elimination, reasoning based on supporting knowledge, just knowing, selection based on past personal experience, familiarity, a combination of strategies, or other.

One result of that analysis—namely, the differing consequences of guessing a wrong answer as opposed to selecting a wrong answer through a reasoning process—is particularly instructive. Guesses were unlikely to persist to the final test (15% were reproduced as answers on the final general knowledge test), whereas errors resulting from faulty reasoning were much more likely to persist. That is, the persistence of errors on the final test was higher (36%) when subjects reasoned about their choices using supporting information (i.e., *most animals sleep on their bellies*, after incorrectly selecting *on their bellies* as the position in which sea otters sleep), or using their own personal experience (persistence of 67%; e.g., *I went there on vacation with my Aunt*). Persistence of errors tends not to arise from a simple boost in familiarity, but rather from a reasoning process in which the error is linked to other world knowledge and integrated into a knowledge base.

Testing Effects with More Complex Materials

To the dismay of some cognitive psychologists, educators tend not to be interested in experiments involving word lists or paired associates, even if such experiments are beautifully controlled. To have an impact on the educational community, studies must use more realistic materials, which is why we have highlighted studies involving prose and general knowledge questions. And yet, these are still relatively simple materials. Educators want much more from students than the ability to list capitals of countries and to provide the definitions of vocabulary words; they want students to move beyond the *who*, *what*, and *where* questions to answer *why*. In Bloom's (1956) taxonomy of educational objectives, knowing basic facts is only the first of his six goals for the student: knowledge, comprehension, application, analysis, synthesis, and evaluation. Thus, a critical issue is, What are the effects of multiple-choice tests when the questions tap higher levels of learning, as defined by Bloom's taxonomy?

To find an answer, we selected questions routinely used in education to tap higher level knowledge. In an experiment similar to those already described, we used published SAT II test materials (Marsh & Roediger, 2006). SAT II tests measure domain-specific knowledge. Many colleges and universities use them as part of the admissions process, and to determine which college courses are appropriate for students, given their prior work.

We selected SAT II questions from the following subjects: biology, chemistry, U.S. history, and world history. The questions included some definitional questions (Level 1 in Bloom's taxonomy), but also questions that tapped higher levels of understanding, such as those that required students to apply their knowledge (Level 3). A research assistant coded each question for its level in Bloom's taxonomy; the average question level was 2.5.

All subjects took four "mini" SAT II tests, with standard SAT II instructions (the SAT II differs from many standardized tests in that it penalizes students for wrong answers and pairs the four alternatives with a "don't know" option). On average, Duke undergraduates answered 55% of the multiple-choice questions correctly, skipped 23%, and selected a lure for 22% of the questions. The tests, therefore, were not easy (Duke undergraduates are expert test-takers).

Of interest were the consequences of taking the SAT II on a later cued recall test. Importantly, a large positive testing effect was obtained: Students correctly answered more questions if they had been tested on the prior multiple-choice test ($M = 48\%$) than if they had not ($M = 22\%$). Taking the SAT II also boosted production of the multiple-choice lures on the final test from a baseline of 7% of responses (when students had not taken the SAT prior to the cued recall test) to 16% (when they had). The overall error rate, however, was not higher following testing. That is, although students learned incorrect answers from the test and their use of specific multiple-choice lure answers was boosted on the final test, they also produced fewer other wrong answers after testing.

These data provide strong evidence that complex multiple-choice questions yield testing effects. Further support comes from a study in which we manipulated how concepts were tested (Marsh, Bjork, & Bjork, 2006). Each concept was tested at Level 1 (definitional) or Level 3 (application) in Bloom's taxonomy. For example, consider the parallel questions created to test the concept of acclimation. The Level 1 (definitional) version of the question read *What biological term describes an organism's slow adjustment to new conditions?* whereas the Level 3 (application) version read *What biological term describes fish slowly adjusting to water temperature in a new tank?* Critically, the answer choices were the same for the two conditions: in this example, *acclimation*, *gravitation*, *maturation*, and *migration*.

Supporting our manipulation of level in Bloom's taxonomy, subjects answered more Level 1 (definitional) questions correctly than Level 3 (application) items. Ques-

tions at both levels led to positive testing effects, with performance rising from 30% correct in the nontested condition to 47% and 48% on final cued recall following testing with Level 1 and Level 3 questions, respectively. Similarly, testing with either a Level 1 (definitional) or a Level 3 (application) multiple-choice question increased lure answers on the final test, compared with the nontested condition. Only 2% of new questions were answered with lures, but 9% and 11% of questions previously tested in Level 1 and Level 3 multiple-choice questions, respectively, were later answered with multiple-choice lures. In short, changing the multiple-choice question to tap a higher level in Bloom's taxonomy did not change the memorial consequences of testing.

We also asked whether subjects would be willing to apply and reason with the incorrect information learned from the multiple-choice test. The answer proved to be yes. For example, selecting a lure (e.g., *gravitation*) as the answer to a multiple-choice question such as "Allowing new fish to adjust slowly to tank water temperature is an example of what biological phenomenon?" increased students' likelihood of later using that lure to answer a transfer question such as "Animals that thicken their fur during winter are exhibiting what biological phenomenon?" Note that the superficial similarity between the questions is minimal, but the questions are conceptually similar: They both test an application of the concept *acclimation*. Again, the data suggest that effects of multiple-choice tests go beyond simple priming of errors; multiple-choice lures may become integrated into subjects' more general knowledge and lead to erroneous reasoning about concepts.

To Test or Not to Test?

Knowing that tests can teach students wrong information, what should an educator do, given the genuine need to assess a student's knowledge? Before we create a false alarm, we need to emphasize again that the overall positive effect of testing (see Figure 1) outweighs any negative consequences (see Figure 2). In addition, in several of our studies the learning of lure answers was balanced by a decrease in other wrong answers on the final test.

To the educator who shudders at the very idea of students acquiring false information from a test, however, we offer the following advice: First, give immediate feedback. This reduces multiple-choice lure production on a later test (Butler & Roediger, 2006). This idea has already been captured in a commercial application, the IFAT (Immediate Feedback Test), which permits teachers to order custom made "Scantrons" that allow students to keep scratching off response options until they reach the correct answer, marked by a star. When a student selects the wrong answer, the lack of a star provides immediate feedback that the answer is wrong (Epstein, Epstein, & Brosvic, 2001).

A second recommendation is to follow the SAT II's example of offering a "don't know" option, with a penalty for selecting a wrong answer. Free responding yielded a small but significant reduction in lure production on a later cued recall test. A final recommendation is to change the ways in which concepts are tested across exams. Switching from a definitional multiple-choice question to an application cued recall question reduced but did not eliminate negative testing effects.

Concluding Comments

The research reviewed here demonstrates that the conventional view of tests as a means of measuring knowledge is overly simple. As has also been demonstrated in research on metacognitive judgments (Spellman & Bjork, 1992), tests modify the knowledge they are designed to assess. In the present research, using a variety of multiple-choice formats, from simple definitional questions to the SAT II, testing helped students to answer questions on later tests. However, we also found that tests can teach

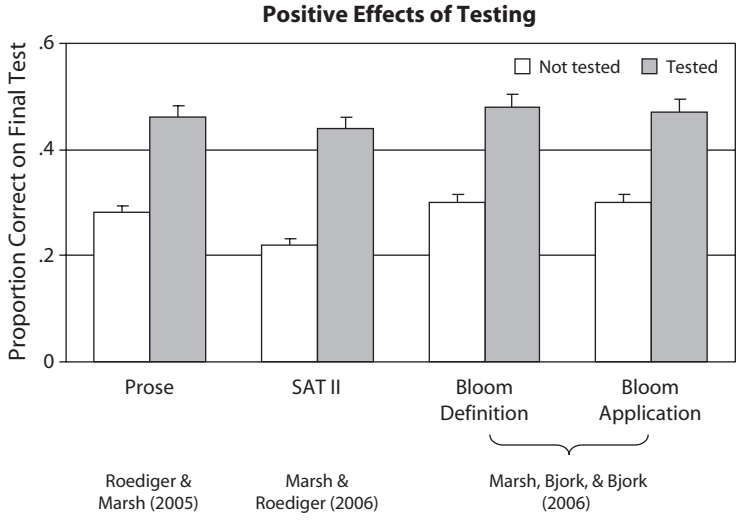

Figure 1. Proportion correct on the final general knowledge test as a function of whether or not concepts had been tested previously, for different types of materials.

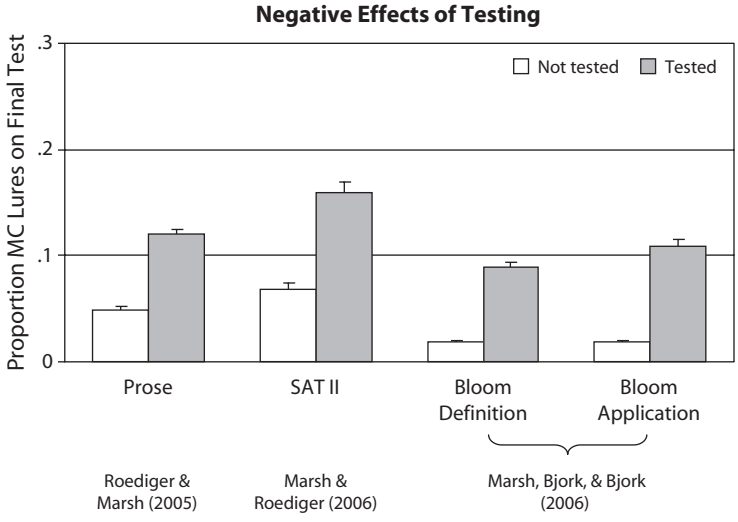

Figure 2. Intrusions of multiple-choice (MC) lures on the final general knowledge test as a function of whether or not concepts had been tested previously, for different types of materials.

students incorrect facts and that such negative effects of testing are not driven simply by rote reproduction of erroneous responses. Rather, the errors reflect meaningful shifts in the ways in which students reason with their knowledge.

More generally, the prevailing societal emphasis on testing as assessment is unfortunate, because it obscures the critical pedagogical aspects of testing. Tests, optimally constructed, can enhance later performance, provide feedback to the learner on what has and has not been learned, and potentiate the efficiency of subsequent study opportunities (see McDaniel, Roediger, & McDermott, 2007). It is not the case, though, that just any test will have all those virtues and at the same time avoid the negative consequences. What is required to construct optimal tests is an understanding of the processing dynamics triggered by testing. We see our research as a step in that direction.

AUTHOR NOTE

This work was supported by a collaborative activity award from the James S. McDonnell Foundation. We thank Barbie Huelser and Keith Payne for their comments on the manuscript. Correspondence concerning this article should be addressed to E. J. Marsh, Department of Psychology & Neuroscience, Duke University, 9 Flowers Drive, Durham, NC 27708-0086 (e-mail: emarsh@psych.duke.edu).

REFERENCES

BANGERT-DROWNS, R. L., KULIK, J. A., & KULIK, C. C. (1991). Effects of frequent classroom testing. *Journal of Educational Research*, **85**, 89-99.

BERGER, S. A., HALL, L. K., & BAHRICK, H. P. (1999). Stabilizing access to marginal and submarginal knowledge. *Journal of Experimental Psychology: Applied*, **5**, 438-447.

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

BLOOM, B. S. (ED.) (1956). *Taxonomy of educational objectives: The classification of educational goals* (Vol. 1). New York: Longman.

BROWN, A. S., SCHILLING, H. E. H., & HOCKENSMITH, M. L. (1999). The negative suggestion effect: Pondering incorrect alternatives may be hazardous to your knowledge. *Journal of Educational Psychology*, **91**, 756-764.

BUTLER, A. C., MARSH, E. J., GOODE, M. K., & ROEDIGER, H. L., III (2006). When additional multiple-choice lures aid versus hinder later memory. *Applied Cognitive Psychology*, **20,** 941-956.

BUTLER, A. C., & ROEDIGER, H. L., III (2006, May). *Feedback neutralizes the detrimental effects of multiple-choice testing*. Poster presented at the annual meeting of the Association for Psychological Science, New York.

CARRIER, M. L., & PASHLER, H. (1992). The influence of retrieval on retention. *Memory & Cognition*, **20**, 633-642.

EPSTEIN, M. L., EPSTEIN, B. B., & BROSVIC, G. M. (2001). Immediate feedback during academic testing. *Psychological Reports*, **88**, 889-894.

FAZIO, L. K., MARSH, E. J., & ROEDIGER, H. L., III (2006, November). *Consequences of multiple-choice testing persist over one week*. Poster presented at the Annual Meeting of the Psychonomic Society, Houston, TX.

FOOS, P. W., & FISHER, R. P. (1988). Using tests as learning opportunities. *Journal of Educational Psychology*, **80**, 179-183.

GATES, A. I. (1917). Recitation as a factor in memorizing. *Archives of Psychology*, **40**, 104.

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

HASHER, L., GOLDSTEIN, D., & TOPPINO, T. C. (1977). Frequency and the conference of referential validity. *Journal of Verbal Learning & Verbal Behavior*, **16**, 107-112.

HOGAN, R. M., & KINTSCH, W. (1971). Differential effects of study and test trials on long-term recognition and recall. *Journal of Verbal Learning & Verbal Behavior*, **10**, 562-567.

HUELSER, B. J., & MARSH, E. J. (2006, November). *Does guessing on a multiple-choice test affect later cued recall?* Poster presented at the Annual Meeting of the Psychonomic Society, Houston, TX.

MARSH, E. J., BJORK, R. A., & BJORK, E. L. (2006). *Testing effects occur with questions that tap higher levels in Bloom's taxonomy*. Manuscript in preparation.

MARSH, E. J., & ROEDIGER, H. L., III (2006). *Positive effects of taking the SAT II on general knowledge*. Unpublished manuscript.

MCDANIEL, M. A., & MASSON, M. E. J. (1985). Altering memory representations through retrieval. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **11**, 371-385.

MCDANIEL, M. A., ROEDIGER, H. L., III, & MCDERMOTT, K. B. (2007).

Generalizing test-enhanced learning from the laboratory to the classroom. *Psychonomic Bulletin & Review*, **14**, 200-206.

ROEDIGER, H. L., III, & GUYNN, M. J. (1996). Retrieval processes. In E. L. Bjork & R. A. Bjork (Eds.), *Memory: Handbook of perception and cognition* (pp. 197-236). San Diego: Academic Press.

ROEDIGER, H. L., III, & KARPICKE, J. D. (2006a). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, **1**, 181-210.

ROEDIGER, H. L., III, & KARPICKE, J. D. (2006b). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, **17**, 249-255.

ROEDIGER, H. L., III, & MARSH, E. J. (2005). The positive and negative consequences of multiple-choice testing. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **31**, 1155-1159.

SPELLMAN, B. A., & BJORK, R. A. (1992). When predictions create reality: Judgments of learning may alter what they are intended to assess. *Psychological Science*, **3**, 315-316.

SPITZER, H. F. (1939). Studies in retention. *Journal of Educational Psychology*, **30**, 641-656.

STERNBERG, R. J., & GRIGORENKO, E. L. (2001). All testing is dynamic testing. *Issues in Education*, **7**, 137-170.

TOPPINO, T. C., & BROCHIN, H. A. (1989). Learning from tests: The case of true–false examinations. *Journal of Educational Research*, **83**, 119-124.

TOPPINO, T. C., & LUIPERSBECK, S. M. (1993). Generality of the negative suggestion effect in objective tests. *Journal of Educational Research*, **86,** 357-362.

TULVING, E., & PEARLSTONE, Z. (1966). Availability versus accessibility of information in memory for words. *Journal of Verbal Learning & Verbal Behavior*, **5**, 381-391.

WHITTEN, W. B., & LEONARD, J. M. (1980). Learning from tests: Facilitation of delayed recall by initial recognition alternatives. *Journal of Experimental Psychology: Human Learning & Memory*, **6**, 127-134.